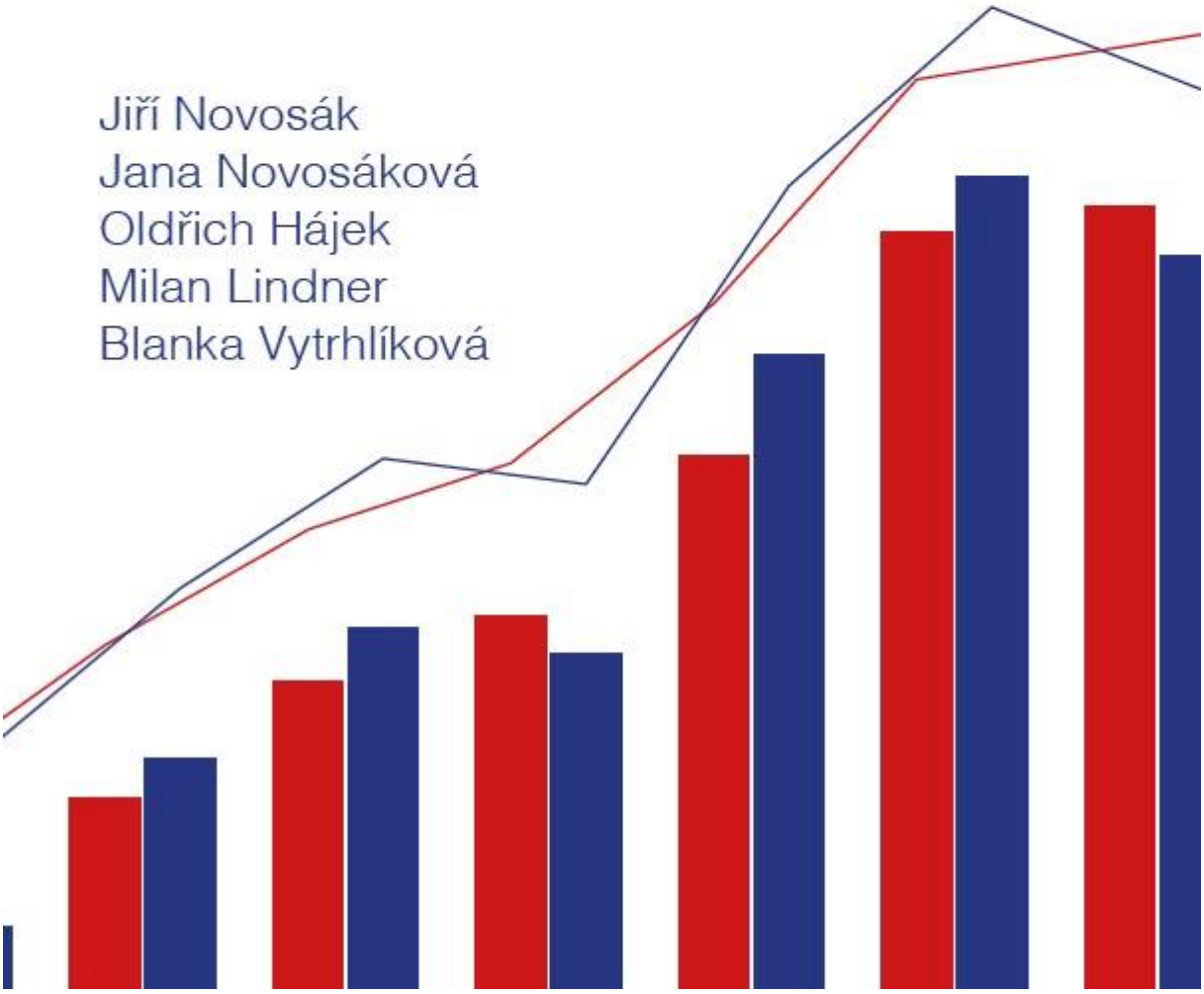
O MOŽNOSTECH VYHODNOCENÍ VÝSLEDKŮ OVĚŘOVACÍHO TESTOVÁNÍ V POČÁTEČNÍM VZDĚLÁVÁNÍ

Jiří Novosák
Jana Novosáková
Oldřich Hájek
Milan Lindner
Blanka Vytrhlíková



O možnostech vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání

Jiří Novosák

Jana Novosáková

Oldřich Hájek

Milan Lindner

Blanka Vytrhlíková

2021

O možnostech vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání

Mgr. Jiří Novosák, Ph.D.

Ing. Jana Novosáková, PhD., MBA

Doc. RNDr. PhDr. Oldřich Hájek, Ph.D., MBA

Ing. Milan Lindner, Ph.D.

Ing. Blanka Vytrhlíková, MBA

Recenzenti:

prof. PhDr. Jaroslav Veteška, Ph.D., MBA

doc. RNDr. Aleš Ruda, Ph.D.

PhDr. Eva Jurášková, Ph.D., MBA

V roce 2021 vydala:

NEWTON Academy; 5. května 1640/65; 140 21 Praha

ISBN: 978-80-87325-39-1

Tato kniha byla zpracována jako výstup řešení projektu Technologické agentury České republiky číslo TL01000385 s názvem „Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání a její aplikace v modelových případových studiích“, a to v rámci programu TL – Program na podporu aplikovaného společenskovědního a humanitního výzkumu, experimentálního vývoje a inovací ÉTA. Řešitelé projektu děkují Technologické agentuře České republiky za finanční podporu při řešení projektu.

Obsah

1. Úvod

Koncept lidského kapitálu patří mezi hlavní východiska vzdělávacích politik současného světa, ať již na národní, nebo nadnárodní úrovni (např. Gillies, 2017). Vzdělávání je v tomto ohledu označováno jako klíčový mechanismus, který zvyšuje kvalitu lidského kapitálu, čímž umožňuje dosahovat budoucích zisků jak nositeli lidského kapitálu v podobě vyššího finančního ohodnocení jeho práce, tak společnosti v podobě ekonomického růstu. Přes řadu kritických připomínek jsou právě tyto přínosy politicky vlivnou motivací k „investicím do vzdělávání" na úrovni jedince i na úrovni vzdělávacího systému (např. Gillies, 2017; Becker, 1992).[1]

Vedle ekonomických přínosů je v politické rovině téma vzdělávání diskutováno v kontextu rozvoje kompetencí pro 21. století (např. Erstad a Voogt, 2018).[2] Voogt a Roblin (2012), Dede (2010), Binkley et al. (2013) označují za motivaci této diskuse odlišnosti kompetencí (znalostí a dovedností) potřebných pro práci, občanský život a sebeuplatnění ve 20. a 21. století. Tyto odlišnosti mimo jiné spočívají ve vyšších nárocích kladených na: (i) zpracování (porozumění, vyhodnocení a interpretaci) množství snadno dostupných informací různé kvality; (ii) hledání strategií řešení problémů mimo standardizované postupy (řešení komplexních problémů a neočekávaných situací); a (iii) spolupráci s dalšími aktéry vlastnícími dílčí části potřebných znalostí a dovedností (multidimenzionální charakter znalostí a dovedností). Důležitou roli také hraje stále častější využívání sofistikovaných informačních a komunikačních technologií a důležitost kompetencí pro uplatnění člověka v současné znalostní společnosti (např. Erstad a Voogt, 2018; Voogt a Roblin, 2012; Dede, 2010; Binkley et al., 2013).

Koncept kompetencí pro 21. století a jeho důležitost pro vzdělávání (např. začlenění do obsahu kurikula) se promítají ve vytvoření několika koncepčních rámců, které tyto kompetence vymezují[3] a mezi něž především patří (např. Erstad a Voogt, 2018; Voogt a Roblin, 2012; Dede, 2010; Binkley et al., 2013):[4]

- kompetence vztahující se k dovednostem učit se a myslet – kritické myšlení; analýza a řešení komplexních problémů v prostředí nejistoty; myšlení vyšší úrovně, simulace a modelování; kreativita a inovace; kontextové učení; plánování, stanovení priorit a orientace na výsledky;
- kompetence vztahující se k dovednostem spolupráce a komunikace – týmová práce a řešení konfliktů; komunikační a argumentační dovednosti; komunikace v interaktivním prostředí a využití podpůrných nástrojů; informační, mediální a vizuální gramotnost;

[1] Gillies (2017) upozorňuje, že koncept lidského kapitálu se v kontextu vzdělávání stává v čase širším, když zahrnuje vedle znalostí a dovedností také postoje či individuální odpovědnost člověka. V praktické rovině se uváděný vztah například promítá v častém důrazu na posun kurikula ve směru prakticky uplatnitelných dovedností (např. Gillies, 2017).

[2] Kompetence pro 21. století jsou označovány také jako celoživotní kompetence, případně klíčové kompetence (např. Voogt a Roblin, 2012).

[3] Koncepční rámce kompetencí pro 21. století se odlišují především ve způsobu jejich kategorizace, respektive v důrazu kladeném na jednotlivé znalosti a dovednosti (např. Erstad a Voogt, 2018).

[4] Voogt a Roblin (2012) poukazují na chybějící shodu na tom, zda kompetence pro 21. století je potřebné vnímat jako kompetence nové, nebo zda se jedná o existující kompetence, které jsou vysoce relevantní pro znalostní společnost.

- vzdělávací obsah relevantní pro 21. století (např. globální a kulturní uvědomění; finanční, ekonomická a podnikatelská gramotnost; občanské vzdělávání; zdraví a životní styl);
- kompetence vztahující se k dovednostem využívat informační a komunikační technologie, včetně dovedností pro sdílení informací (např. sociální sítě), myšlení (např. blog, podcast, online diskusní fóra) a společné tvoření;
- kompetence vztahující se k dovednostem pro život – leadership; odpovědnost; etika; osobní a společenská odpovědnost; zvědavost, sebeřízení a adaptabilita; samostatnost a akceptování rizika; život a kariéra.

Dede (2010) dále zdůrazňuje potřebu vložit kompetence pro 21. století do výuky hlavních předmětů[5], tj. nalézt soulad mezi změnou a kontinuitou vzdělávání (např. Kereluik et al., 2013), přičemž Dede (2010) poukazuje na významnost psychologických, politických a kulturních překážek pro dosažení tohoto záměru (např. také Voogt a Roblin, 2012). Další související otázky pak zahrnují: (i) „zeštíhlení" kurikula[6] a posílení vhodných metod vzdělávání (např. problémově orientované učení, kooperativní učení, učení prožitkem a zkušeností, formativní hodnocení); (ii) profesní rozvoj učitelů k výuce kompetencí pro 21. století (např. dovednosti využívat vhodné vzdělávací metody, stejně jako ICT ve výuce)[7]; a (iii) zajištění vybavenosti škol potřebnými ICT (např. Dede, 2010; Voogt a Roblin, 2012).

Významnost proměn ve společnosti pro další směřování vzdělávání ve 21. století přiznává také hlavní koncepční dokument České republiky v této oblasti – *Strategie vzdělávací politiky České republiky do roku 2030+* (dále jen „Strategie vzdělávací politiky 2030+"). Ta za klíčové společenské proměny současnosti označuje hospodářské a jiné transformace měnící soubory dovedností potřebných pro výkon tradičních i nových povolání, důležitost digitálních technologií pro komunikaci a socializaci člověka a prakticky neomezený přístup člověka k informacím, které je však nezbytné kriticky vyhodnocovat a dále s nimi pracovat (viz Fryč et al., 2020). Soulad zaměření Strategie vzdělávací politiky 2030+ s uváděnými úvahami o směřování vzdělávání v 21. století potvrzuje také obsahová analýza záměrů, které jsou v tomto dokumentu uváděny (viz tabulka č. 1).

Strategie vzdělávací politiky 2030+ ve svém obsahu vyzvedává význam hodnocení ve vzdělávání, a to včetně ověřování dosažení očekávaných vzdělávacích výstupů v uzlových bodech na úrovni školy i vzdělávacího systému a s důrazem na úroveň kompetencí a gramotností žáků (viz Fryč et al., 2020). Důležitost hodnocení v úvahách o rozvoji kompetencí pro 21. století zmiňují také Erstad a Voogt (2018), Voogt a Roblin (2012), a rovněž Dede (2010) poukazuje na potřebu společného hodnocení kompetencí pro 21. století a učiva hlavních předmětů. Voogt a Roblin (2012), Gillies (2017) zároveň poukazují na příležitosti ke zkvalitňování systémů hodnocení rozvoje kompetencí pro 21. století, které by měly splňovat následující znaky (např. Binkley et al., 2013; Voogt a Roblin, 2012):

[5] Erstad a Voogt (2018) poukazují na nalezení souladu mezi otázkami „vědět co" a „vědět jak".
[6] Voogt a Roblin (2012) uvádějí tři možnosti, jak do kurikula začlenit znalosti a dovednosti pro 21. století: (i) nový předmět nebo nový obsah v tradičních předmětech; (ii) průřezová témata napříč kurikulem; a (iii) součást nového kurikula založeného na transformaci tradiční struktury předmětů.
[7] Voogt a Roblin (2012) zdůrazňují tři vhodné formy profesního rozvoje učitelů: (i) pozorování modelových příkladů; (ii) účast v průběžných s výukou spojených iniciativách profesního rozvoje učitelů; a (iii) účast na aktivitách komunit učitelů.

Tabulka č. 1: Záměry uváděné ve Strategii vzdělávací politiky 2030+

Strategická linie	Záměry
1. Proměna obsahu, způsobů a hodnocení vzdělávání	- Zlepšování vztahu žáků ke škole a učení; zvyšování vnitřní motivace žáků; uchopení vzdělávacího potenciálu všech žáků (individualizace vzdělávání); snižování intenzity problémů spojených s rizikovým chováním žáků - Rozvíjení schopností žáků učit se, kriticky myslet a řešit náročnější úkoly vyžadující hlubší porozumění a praktickou aplikaci; posun od memorování znalostí k rozvoji kompetencí a gramotností žáků; odklon od učení se na zkoušky - Rozvíjení schopností žáků spolupracovat a hledat společná řešení - Rozvíjení schopností žáků pochopit, provázat a využít znalosti; realizace vzdělávání v souladu s potřebami trhu práce a praxe - Jasně vymezený obsah kurikula s formulací očekávaných výstupů různých úrovní pro výběr vhodných vzdělávacích strategií - Redukce učiva pro vytvoření potřebného časového prostoru ve výuce - Rozmanitost nabídky metod a forem výuky pro individualizaci vzdělávání; šíření inovací a příkladů dobré praxe - Propojení obsahu vzdělávání a ověřování výsledků vzdělávání, posilování praxe hodnocení a ověřování výsledků žáků v práci učitele - Uchopení potenciálu využití digitálních technologií pro vzdělávání, rozvíjení digitální gramotnosti a informatického myšlení žáků - Posilování občanského vzdělávání – demokratické hodnoty, zájem o věci veřejné a život kolem sebe, udržitelný rozvoj, kritické myšlení, mediální gramotnost - Přijetí proměny obsahu, metod a forem vzdělávání zainteresovanými aktéry
2. Rovný přístup ke kvalitnímu vzdělávání	- Zajištění srovnatelné a kvalitní výuky ve všech školách - Snižování rozdílů ve vzdělávání na úrovni škol, žáků i území
3. Podpora pedagogických pracovníků	- Kvalitní pedagogické vedení školy, dostatečná a kvalitní podpora pedagogické práce škol - Dostatečný počet motivovaných, dobře připravených a spokojených učitelů
4. Zvýšení odborných kapacit, důvěry a vzájemné spolupráce	- Posilování komunikace a spolupráce - Zlepšení využití dat a výzkumů pro oblast vzdělávání - Snížení administrativní zátěže škol
5. Zvýšení financování a zajištění jeho stability	- Zajištění potřebného financování pro naplnění záměrů v oblasti vzdělávání

Zdroj: vlastní zpracování podle Fryč et al. (2020)

- využívání koncepčního rámce hodnocení, který vychází z formulovaných cílů vzdělávání (např. rozvíjení kompetencí pro 21. století);[8]
- využívání hodnocení pro poskytování zpětné vazby a identifikaci vzdělávacích potřeb (formativní charakter hodnocení);
- využívání potenciálu ICT pro hodnocení;
- naplňování kritérií kvalitního hodnocení vzhledem k interpretaci výsledků (např. validita, reliabilita, spravedlnost, psychometrické vlastnosti testových položek a testů, diferenciace mezi různými odpověďmi žáků, vazba na konstrukty/dimenze hodnocení).

A právě poslední z uvedených znaků hodnocení ve vzdělávání je předmětem zájmu této knihy. Specificky se kniha zaměřuje na hodnocení výsledků ověřovacích testů v počátečním vzdělávání (dále jen „test"), jejichž záměrem bývá poskytovat informace o: (a) vzdělávacích výsledcích žáků a jejich determinantech (Zhang, 2008); (b) pokroku žáků ve vzdělávání (De Champlain, 2010; Ryan a Brockmann, 2009; Udofia a Uko, 2016); a (c) vzdělávacích potřebách žáků (Lambert et al., 2018). Získané informace následně slouží jako informační vstupy pro rozhodovací proces aktérů v různých situacích, jako jsou například přijímací zkoušky, udělení profesního certifikátu, schválení podoby kurikula, financování škol či zdůvodnění akontability veřejných financí alokovaných na oblast vzdělávání (např. Ryan a Brockmann, 2009; Cook a Eignor, 1991; Lambert et al., 2018; Betebenner a Linn, 2010; Thompson, 2015; Udofia a Uko, 2016; Betebenner, 2009). V tomto kontextu je utvářena poptávka po kvalitních ověřovacích testech, přičemž Thompson (2016) upozorňuje, že ačkoliv jsou testy využívány každodenně, jejich kvalita není vždy v souladu s existujícími standardy.

Kniha je strukturována následujícím způsobem. Ve druhé kapitole jsou představeny cíle knihy a k nim doplňující východiska jejího zpracování. Ve třetí kapitole jsou představeny teoreticko-metodické přístupy k problematice vyhodnocování ověřovacích testů, které jsou ve čtvrté kapitole ilustrovány v podobě modelových případových studií. Pátá kapitola syntetizuje poznatky předchozích dvou kapitol v metodice vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání a konečně šestá kapitola je závěrem celé knihy.

[8] Gillies (2017), Voogt a Roblin (2012), Dede (2010) v tomto ohledu poukazují na některé problémy současného nastavení testů, které jsou například spojené s chybějícími možnostmi testovat některé z kompetencí pro 21. století (např. týmová spolupráce, strategie řešení problémů, využití ICT aplikací, komplexní úlohy vyžadující aktivaci vyššího počtu kompetencí).

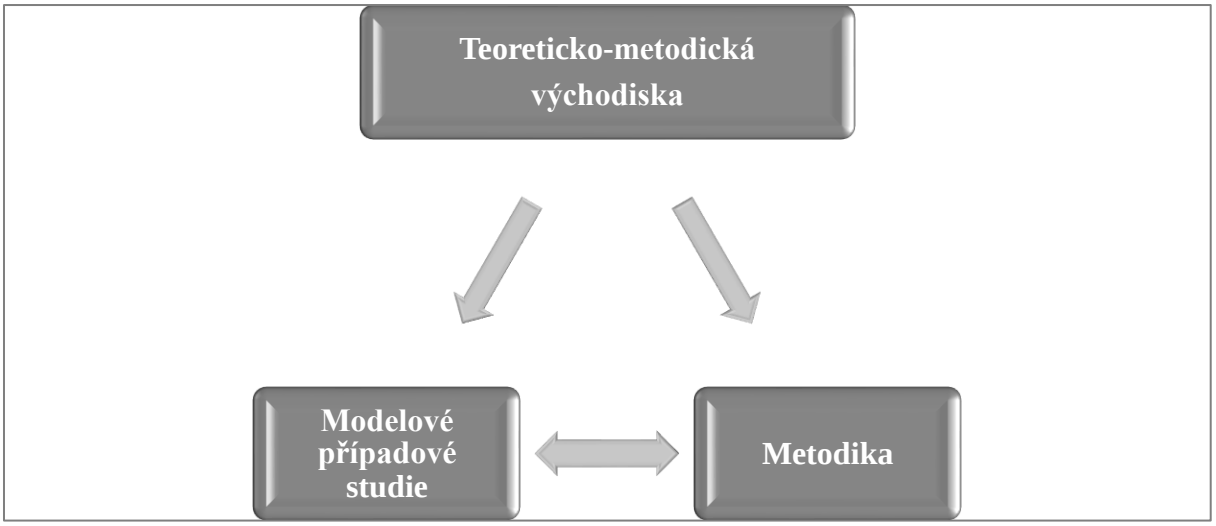
2. Cíle a doplňující východiska

Tato kniha je jedním z výstupů projektu „Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání a její aplikace v modelových případových studiích“, který byl řešen projektovým týmem Newton College s podporou Technologické agentury České republiky v období let 2018-2021. Cíle knihy, které jsou plně v souladu s cíli projektovými, byly formulovány následujícím způsobem:

- shrnout podstatu teoreticko-metodických východisek problematiky vyhodnocení ověřovacího testování v počátečním vzdělávání;
- ilustrovat průmět teoreticko-metodických východisek do řešení vybraných modelových případových studií vyhodnocení ověřovacího testování v počátečním vzdělávání;
- představit podstatu metodiky vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání (hlavní výstup řešeného projektu) jako syntézy teoreticko-metodických východisek problematiky vyhodnocení ověřovacího testování v počátečním vzdělávání a řešení modelových případových studií.

Obrázek č. 1 zachycuje podstatu vazeb mezi dílčími tematickými částmi této knihy, které logicky vycházejí z formulovaných cílů.

Obrázek č. 1: Dílčí tematické části knihy a vztahy mezi nimi



Formulace uvedených cílů knihy je potřebné doplnit o související východiska. Primárně jsou cíle knihy naplňovány pro řešení testů tvořených dichotomickými testovými položkami, které mohou být zodpovězeny buď správně, nebo nesprávně. Problematika polytomických položek, kde jsou možnosti jejich vyhodnocení širší, není v této knize diskutována.

Druhé východisko souvisí se skutečností, že naplňování třetího cíle knihy je spojeno jen se vstupním představením Metodiky vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání. V tomto ohledu je vhodné čtení této knihy spojit se čtením i samotné metodiky,

neboť informace uváděné v těchto dvou výstupech jsou provázané, a to včetně odkazu na využívané datové zdroje a metodické postupy zpracování.

3. Teoreticko-metodická východiska

Tvorba testů a jejich vyhodnocení se setkává s řadou teoreticko-metodických i praktických otázek (např. Ziegler a Hagemann, 2015).[9] Dva hlavní přístupy k hledání odpovědí na ně zahrnují jednak klasickou teorii testů, jednak teorii odpovědi na položku (např. Wang, Ma a Chen, 2010; Ryan a Brockmann, 2009). Právě tyto dva přístupy jsou blíže představeny v této kapitole.

3.1 Klasická teorie testů

Klastická teorie testů (dále jen „CTT“) představuje tradiční, dlouhodobě rozvíjený přístup k práci s testy (např. Traub, 1997; Wang, Ma a Chen, 2010). Typickým znakem tohoto přístupu je využití testů k měření hodnoty, která vyjadřuje úroveň zvládnutí hodnoceného konstruktu, není však pozorovatelná přímo (např. dovednosti žáků, povahové rysy žáků).[10] Význam testu spočívá v odhadu úrovně nepozorované (latentní) proměnné (např. DeVellis, 2006; De Champlain, 2010). Z uvedené motivace následně vychází základní vztah CTT (např. Hambleton a Jones, 1993; DeVellis, 2006; Revelle, 2012; Krishnan, 2013; Brennan, 2011; De Champlain, 2010; Graham, 2006):

$$X = T + E,$$

kde: (a) X odpovídá pozorovanému skóre testované osoby v testu; (b) T odpovídá skutečnému (pravdivému) skóre testované osoby v testu[11]; a (c) E odpovídá chybovému skóre[12]. Takto je pozorované skóre testované osoby mixem informací plynoucích z její úrovně zvládnutí hodnoceného konstruktu (T) a chybového skóre (E), přičemž právě proměnné T a E uvedeného vztahu jsou latentního (nepozorovatelného) charakteru (např. Graham, 2006; DeVellis, 2006; Krishnan, 2013; Hambleton a Jones, 1993; Brennan, 2011; Traub, 1997).

V ideálním případě by platilo, že hodnota pozorovaného skóre testované osoby odpovídá úrovni zvládnutí hodnoceného konstruktu, tj. platí vztah $X = T$ (např. Revelle, 2012; Krishnan, 2013). Je zjevné, že v praxi taková situace nenastává, a to z řady důvodů, mezi které mimo jiné patří (např. Krishnan, 2013; Brennan, 2011):

- nekontrolované podmínky testování (např. odlišné kódování hodnotitelů, prostorové a časové podmínky testování);

[9] Takto například Zhang (2008) uvádí otázku volby podoby testových položek, kdy výběrové (*multi-choice*) testové položky spojuje s možností širokého obsahového zaměření testu a otevřené testové položky se zaměřením na hlubší učení.

[10] Z tohoto důvodu doporučuje Toland (2014), aby se na tvorbě škály pro měření úrovně latentní schopnosti osob podíleli jak experti pro obsah testových položek, tak experti pro hodnocení testových položek.

[11] Skutečné (pravdivé) skóre testované osoby, tj. úroveň jejího zvládnutí hodnoceného konstruktu, teoreticky odpovídá střední hodnotě nekonečného počtu měření skóre testované osoby v tzv. paralelních testech. Paralelní testy jsou charakteristické tím, že: (a) měří stejný konstrukt (latentní proměnná); (b) mají stejné hodnoty průměrů, variancí a kovariancí pozorovaného skóre X; (c) kovariance mezi chybovými skóre a skutečnými skóre je rovna nule; a (d) kovariance mezi chybovými skóre testů je rovna nule. Charakteristiky (c) a (d) jsou dány tím, že očekávaná hodnota chybového skóre je rovna nule (např. Brennan, 2011).

[12] Brennan (2011) upozorňuje, že chybové skóre není ve skutečnosti chybou, ale nedostatkem souladu hodnot s modelem CTT. Zároveň základní vztah CTT nelze zaměňovat s regresním modelem.

- náhodná fluktuace výsledků testované osoby daná jejími osobnostními charakteristikami (např. motivace, únava);
- chyba vznikající z odlišnosti testů;
- chyba spojená s kvalitou slovního vyjádření testové položky (např. matoucí vyjádření pro určité skupiny testovaných osob).

Přítomnost chyby je motivací k hodnocení spolehlivosti testu, tj. odhadu toho, do jaké míry pozorované skóre testované osoby v testu odpovídá skóre skutečnému, respektive do jaké míry jsou výsledky testovaných osob v testu konzistentní (např. Thompson, 2016; DeVellis, 2006). Tabulka č. 2 představuje různé postupy odhadu hodnoty spolehlivosti testu, přičemž předpokladem správnosti takového odhadu je unidimenzionalita testu, kdy všechny testové položky měří stejný konstrukt, tj. latentní proměnnou (např. Graham, 2006; DeVellis, 2006). Pokud není předpoklad unidimenzionality testu naplněn, snižuje se důvěryhodnost odhadu jeho spolehlivosti. Uveďme, že pro tzv. rozhodné (*high-stake*) testy je potřebná vyšší spolehlivost s ohledem na významnost jejich praktických dopadů (např. De Champlain, 2010).[13]

Tabulka č. 2: Postupy odhadu hodnoty spolehlivosti testu a jejich stručná charakteristika

Postup	Stručná charakteristika postupu
Znalost pozorovaného i skutečného skóre v testu	Spolehlivost testu obecně odpovídá podílu pravdivé variance uchopené v celkové varianci testu (např. Graham, 2006; DeMars, 2010), respektive druhé mocnině korelace mezi skutečným skóre a pozorovaným skóre (např. Brennan, 2011). Skutečné skóre v testu, tj. úroveň zvládnutí hodnoceného konstruktu, ovšem nebývá známo, proto jsou hledány alternativní postupy stanovení spolehlivosti testu.
Využití paralelních testů	Odhad hodnoty spolehlivosti testu s využitím paralelních testů vychází z existence dvou testů následujících charakteristik: (a) každá testovaná osoba má stejné skutečné skóre pro každý test; (b) variance skutečného skóre každého testu je stejná; a (c) variance chybového skóre každého testu je stejná (např. Revelle, 2012). Následně platí, že korelace skóre dvou paralelních testů odpovídá druhé mocnině korelace skóre každého z těchto testů a jejich pravdivého skóre, tj. pro stanovení spolehlivosti testu stačí vypočítat korelaci mezi dosaženými skóre v daném testu a v libovolném testu paralelním. I zde však zůstává problém nalezení paralelního testu, který by splňoval uvedené podmínky. Typicky pak zadáváme pouze jeden test. Z těchto důvodů jsou hledány postupy stanovení hodnoty spolehlivosti testu prostřednictvím vnitřní struktury jediného testu (např. Revelle, 2012).
Rozdělení testu	Odhad hodnoty spolehlivosti testu jeho rozdělením (tzv. *split-half* metoda) vychází z rozdělení testu na dvě poloviny náhodně vybraných testových položek s tím, že následně je vypočtena korelace dosažených skóre testovaných osob v nich. S ohledem na zkrácení testů je potřebná korekce vypočtené hodnoty korelace tzv. Spearman-Brownovým „věšteckým“ vzorcem (např. Revelle, 2012). Kritika tohoto přístupu vychází z velkého počtu možných způsobů rozdělení testu a otázkou, zda bude hodnota spolehlivosti testu vycházet vždy stejně (např. Traub, 1997).

[13] V případě Cronbachova alfa označují Krishnan (2013), Thompson (2016) za minimální hodnotu spolehlivosti testu hodnotu 0,7, pro rozhodné testy je typicky vyžadována vyšší hodnota.

Postup	Stručná charakteristika postupu
Vnitřní struktura testu	Odhad hodnoty spolehlivosti testu na základě jeho vnitřní struktury vychází z motivace mít k dispozici jednoduchý nástroj, který stanoví spolehlivost testu s využitím vlastností testových položek (variance a kovariance) bez potřeby hledat k němu paralelní test (např. Revelle, 2012). Tradičně používané ukazatele spolehlivosti testu, jako je například Cronbachovo alfa, vycházejí z tzv. esenciálně tau-ekvivalentních modelů, pro které platí: (a) testové položky měří stejný konstrukt; (b) na stejné škále (stejná variance testových položek); (c) s možnou odlišnou hodnotou skutečného skóre testované osoby[14]; a (d) s odlišným množstvím chyby (např. Graham, 2006). Podstata odhadu hodnoty spolehlivosti testu pak vychází ze dvou základních myšlenek (Graham, 2006): • Celková variance testu je součtem kovariancí a variancí testových položek. • Kovariance testových položek je součástí pravdivé variance testu, variance testových položek je součástí pravdivé i chybové variance.[15] Spolehlivost testu je následně odhadována v závislosti na tom, jaký je podíl chybové variance na celkové varianci testu, přičemž různé ukazatele zohledňují podíl pravdivé a chybové variance ve varianci testových položek odlišně. Definovány tak jsou různé ukazatele, které zahrnují: (a) Guttmanových šest odhadů dolní spolehlivosti testu λ_1 až λ_6, kdy ukazatel λ_3 odpovídá hodnotě Cronbachova alfa; a (b) ukazatele ω_h a ω_t, které zohledňují potenciální přítomnost skupinových faktorů v testu utvářených pouze některými testovými položkami. Takto ukazatel ω_h odhaduje spolehlivost testu vztahující se pouze k obecnému faktoru, ukazatel ω_t zohledňuje také vliv skupinových faktorů dílčích testových položek, tj. dalších subdimenzí testu (např. Revelle, 2012). Revelle (2012) následně uvádí, že pokud dochází k situaci, kdy se hodnota ukazatele ω_h blíží nule, zatímco hodnota ukazatele ω_t se blíží jedničce, pak je význam obecného faktoru relativně slabý a vzniká otázka, zda raději nerozdělit test na více skupinových škál. Rovněž hodnota Cronbachova alfa není v tomto případě důvěryhodná s ohledem na nenaplnění předpokladu unidimenzionality testu. Graham (2006), Brennan (2011) pak upozorňují na další problém spojený s využitím Guttmanových odhadů dolní spolehlivosti testu, tj. také Cronbachova alfa, který spočívá v nenaplnění předpokladů esenciálně tau-ekvivalentních modelů, na nichž je Cronbachovo alfa založeno (např. významně odlišná úroveň variance i jediné testové položky; využití různých formátů testových položek). V takovém případě bude hodnota Cronbachova alfa významně podhodnocena (např. Graham, 2006).

[14] Graham (2006) uvádí příklad dvou testových položek, které měří stejný konstrukt, ale liší se v intenzitě projevu: (a) Někdy se cítím smutný. (b) Téměř vždy se cítím smutný. Odpovědi respondentů by měly zůstat podobné, nicméně posunuté na definované škále.

[15] Tento předpoklad je založen na myšlence, že pokud testové položky měří stejný konstrukt, měly by být mezi sebou korelovány, tj. měly by mít mezi sebou dostatečně silné korelace, respektive kovariance (např. Krishnan, 2013). DeVellis (2006) charakterizuje tuto úvahu rozlišením sdílené a jedinečné variance testových položek s tím, že vyšší podíl sdílené variance znamená, že testové položky mají více společného a tím má testová škála vyšší úroveň spolehlivosti.

3.1.1 Analýza testových položek a hodnocení vzoru odpovědí testovaných osob

Pro vyhodnocení testu (např. identifikace silných a slabých stránek testovaných osob), stejně jako pro tvorbu testu (např. zvyšování kvality testu zamítnutím využití málo kvalitních testových položek), hraje významnou roli analýza testových položek (např. Krishnan, 2013; De Champlain, 2010; Thompson, 2016). Na tomto místě je proto žádoucí uvést hlavní statistiky, které jsou nedílnou součástí analýzy testových položek.

- *Základní ukazatele deskriptivní statistiky úspěšnosti odpovědí na testové položky*

První okruh statistik, které jsou součástí analýzy testových položek, je založen na výpočtu základních ukazatelů deskriptivní statistiky vztahujících se k úspěšnosti testovaných osob při odpovědích na ně. Jedná se tedy o ukazatele střední hodnoty a směrodatné odchylky úspěšnosti odpovědí testovaných osob na testové položky (např. Krishnan, 2013).

- *Obtížnost testové položky*

CTT definuje obtížnost dichotomických testových položek jako podíl testovaných osob, kteří zodpověděli danou testovou položku správně. Platí tedy, že čím je hodnota obtížnosti testové položky vyšší, tím je tato testová položka jednodušší (např. DeVellis, 2006). Obtížnost testové položky je významným aspektem její kvality, kdy Krishnan (2013) doporučuje sledovat následující pravidla:

- Čím se hodnota obtížnosti testové položky více blíží hodnotě 0,5, tím vyšší má taková testová položka schopnost diferenciace mezi testovanými osobami. Testové položky s obtížností 0, respektive s obtížností 1, jsou extrémními testovými položkami, které nejsou schopny mezi testovanými osobami diferencovat.
- V úvahách o zařazení testových položek do testu je potřeba zvážit přínosnost těch položek, jejichž obtížnost je vyšší než 0,90 a nižší než 0,10 (či 0,20). Takové testové položky je nutné považovat za kandidáty k vyřazení z testu, pokud je takový postup odůvodněný.
- Žádoucí obtížnost testové položky odpovídá hodnotě mírně vyšší, než je průměr maximální možné hodnoty (tj. hodnoty 1) a šance testované osoby správně zodpovědět testovou položku prostřednictvím hádání.[16]

Uveďme, že Krishnan (2013) charakterizuje testové položky vzhledem k obtížnosti následujícím způsobem: (a) velmi jednoduché testové položky (obtížnost vyšší než hodnota 0,90); (b) jednoduché testové položky (obtížnost v intervalu hodnot 0,76 až 0,90); (c) optimální testové položky (obtížnost v intervalu hodnot 0,26 až 0,75)[17]; (d) obtížné testové položky (obtížnost v intervalu hodnot 0,10 až 0,25)[18]; a (e) velmi obtížné testové položky (obtížnost

[16] V případě výběru ze čtyř odpovědí je 25% šance správně zodpovědět testovou položku hádáním. Žádoucí obtížnost testové položky pak odpovídá hodnotě mírně vyšší, než je hodnota 0,625.
[17] Ryan a Brockmann (2009) uvádějí ideální hodnoty obtížnosti testových položek v intervalu 0,40 až 0,65.
[18] Thompson (2016) označuje za obtížné testové položky s obtížností nižší než 0,30, a to rovněž v kontextu faktoru uhodnutí správné odpovědi.

nižší než hodnota 0,10). Právě velmi jednoduché a velmi obtížné testové položky jsou problematické, neboť jen omezeně diferencují mezi testovanými osobami.

- *Diskriminace testové položky*

Koncept diskriminace vyjadřuje, do jaké míry je testová položka schopna rozlišit mezi testovanými osobami s různou úrovní zvládnutí hodnoceného konstruktu (např. DeVellis, 2006; DeMars, 2010; De Champlain, 2010; Thompson, 2016). Takto jsou testové položky s vyšší úrovní diskriminace schopny lépe rozlišit mezi testovanými osobami, které lépe zvládají hodnocený konstrukt a testovanými osobami, které hodnocený konstrukt zvládají hůře (např. DeMars, 2010). Krishnan (2013), Thompson (2016) uvádějí několik postupů hodnocení diskriminace testové položky:

- Index diskriminace testové položky vychází z vytvoření dvou skupin testovaných osob podle jejich celkového výsledku v testu. První skupina testovaných osob zahrnuje 27 % testovaných osob s nejlepším dosaženým výsledkem a druhá skupina 27 % testovaných osob s nejhorším dosaženým výsledkem. Pro tyto dvě skupiny testovaných osob je vypočten podíl správně odpovídajících testovaných osob na testové položky s tím, že korespondující hodnoty jsou odečteny a tento rozdíl utváří samotný index diskriminace. Platí, že vyšší hodnoty indexu diskriminace jsou spojeny s lepší schopností testové položky rozlišit mezi testovanými osobami podle toho, jak dobře zvládají hodnocený konstrukt.
- Diskriminaci testové položky lze dále měřit úrovní souladu odpovědí testovaných osob na danou testovou položku s jejich výsledky v celém testu. Platí, že vysoké kladné hodnoty korelace jsou spojeny s testovými položkami, které lépe diskriminují mezi testovanými osobami podle toho, jak dobře zvládají hodnocený konstrukt (např. DeVellis, 2006). Upravená bodově biseriální korelace je analogií k předchozí korelaci s tím, že celkové skóre v testu je počítáno bez hodnocené testové položky, aby tato nepřispívala ke zvyšování hodnoty korelace (např. De Champlain, 2010).

Z uváděných přístupů k hodnocení diskriminace testové položky je nejčastěji doporučovaným ukazatel upravené bodově biseriální korelace (např. DeMars, 2010; Krishnan, 2013). I v případě tohoto ukazatele lze v odborné literatuře najít referenční hodnoty. Ryan a Brockmann (2009), Krishnan (2013) zmiňují hodnotu 0,30 jako minimální hodnotu upravené bodově biseriální korelace dobrých testových položek, Thompson (2016) pak hodnotu 0,20 s tím, že silně diskriminující položky dosahují hodnot v intervalu 0,50 až 0,60. Naopak za nejvíce problematické označuje Krishnan (2013) testové položky s nízkou hodnotou upravené bodově biseriální korelace a zároveň s velmi vysokou či velmi nízkou obtížností. Pokud je hodnota upravené bodově biseriální korelace velmi nízká či dokonce záporná, pak Thompson (2016) zmiňuje tři možné příčiny: (a) existence chyby v určení správné odpovědi; (b) existence velmi atraktivního distraktoru; a (c) příliš obtížná či naopak příliš jednoduchá testová položka i pro testované osoby s nejvyšší či nejnižší úrovní zvládnutí hodnoceného konstruktu. Konečně v případě testů s nízkým počtem testových položek doporučuje Krishnan (2013) výpočet také průměrné korelace mezi testovými položkami s doporučenou minimální hodnotou 0,70.

Vztah k hodnocení diskriminace testové položky mají rovněž ukazatele spolehlivosti testu. Podstata tohoto přístupu je založena na výpočtu dvou hodnot zvoleného ukazatele spolehlivosti testu, kdy se: (a) první hodnota váže k celému testu, tj. ke všem testovým položkám; a (b) druhá hodnota k testu, v němž byla vynechána hodnocená testová položka. Rozdíl obou hodnot pak naznačuje, zda vynechání testové položky vede ke zvýšení spolehlivosti testu či nikoliv. Kladná odpověď na tuto otázku indikuje nižší kvalitu testové položky vzhledem k hodnocenému konstruktu, přičemž tato skutečnost se projeví v hodnotě dalších ukazatelů diskriminace testové položky.

- *Kvalita distraktorů testové položky*

V případě testových položek, které testovaným osobám nabízejí výběr z nabídky odpovědí (*multi-choice*), je předmětem analytického zájmu také hodnocení kvality distraktorů těchto testových položek, tj. nabízených nesprávných odpovědí. Postup je v tomto ohledu analogický s výpočtem diskriminace testové položky, kdy je opětovně počítána upravená bodově biseriální korelace, a to mezi odpověďmi testovaných osob spojenými s daným distraktorem[19] a celkovým dosaženým skóre testovaných osob v celém testu (např. Krishnan, 2013; Thompson, 2016). Žádoucí jsou v tomto případě záporné hodnoty upravené bodově biseriální korelace, kladné hodnoty je naopak nutné považovat za problematické, a to například z následujících důvodů: (a) správná odpověď na testovou položku je matoucí či pochybná, a proto testované osoby vybírají odpověď jinou; (b) více než jedna odpověď přitahuje pozornost testovaných osob s vysokou úrovní zvládnutí hodnoceného konstruktu; a (c) klíč řešení testové položky je chybný (např. FDE, 2017). Uveďme rovněž, že distraktor má pro testovou položku omezený význam, a jeho kvalita je proto nízká, pokud si jej vybral jen velmi omezený počet testovaných osob (např. Krishnan, 2013; Thompson, 2016; Ryan a Brockmann, 2009).

- *DIF analýza testových položek*

Záměrem DIF[20] analýzy testových položek je posoudit jejich spravedlnost vzhledem k odpovědím různých skupin testovaných osob (např. dívky oproti chlapcům, kulturně-etnické skupiny). DIF analýza tedy hledá odpověď na otázku, zda testová položka určitým způsobem neznevýhodňuje některou skupinu testovaných osob, například kvůli své formulaci (např. Krishnan, 2013; Özdemir, 2015).[21] Karami (2012) hovoří o nespravedlnosti testové položky v případě, kdy dvě různé skupiny testovaných osob se stejnou úrovní zvládnutí hodnoceného konstruktu neodpovídají na testovou položku stejně, přičemž faktor, který stojí v pozadí těchto odlišností, není spojený s hodnoceným konstruktem, nýbrž má vazbu k charakteristice těchto skupin testovaných osob. Takto má DIF analýza zásadní význam pro validitu testů, neboť rozdíly dané charakteristikami skupin testovaných osob utváří nežádoucí šum v kvalitě odhadovaných statistik (např. Wiberg, 2007; Yavuz et al., 2018).

[19] Hodnocený distraktor zde tedy de facto plní roli „správné odpovědi".
[20] *Differentiated Item Functioning*, tj. analýza diferenciálního fungování položek
[21] Karami (2012) přitom uvádí, že zájem o DIF analýzu se významně zvýšil kvůli rostoucímu zájmu o otázku rovných příležitostí ve vzdělávání (viz rovněž Wiberg, 2007).

Z terminologického hlediska DIF analýza rozlišuje: (a) ohniskovou skupinu testovaných osob, u které je předpokládáno znevýhodnění; a (b) referenční skupinu testovaných osob, vůči které je ohnisková skupina posuzována (např. Wiberg, 2007; Karami, 2012). DIF analýza dále definuje tzv. uniformní DIF, který nastává tehdy, pokud rozdíly v odpovědích testovaných osob ohniskové a referenční skupiny jsou napříč různými úrovněmi zvládnutí hodnoceného konstruktu stejné. Naopak v případě neuniformního DIF jsou rozdíly v odpovědích testovaných osob ohniskové a referenční skupiny odlišné pro různé úrovně zvládnutí hodnoceného konstruktu (např. Yavuz et al., 2018; Özdemir, 2015). Úroveň zvládnutí hodnoceného konstruktu je typicky měřena dosaženým skóre v testu (např. Karami, 2012). Pro DIF analýzu byla navržena řada metodických přístupů, tabulka č. 3 charakterizuje některé z nich.

Tabulka č. 3: Charakteristika vybraných metod DIF analýzy

Metoda	Charakteristika metody
Logistická regrese	DIF analýza založená na logistické regresi vychází z modelování pravděpodobnosti správné odpovědi na testovou položku (*log-odds*) v závislosti jednak na zvládnutí hodnoceného konstruktu testovanou osobou (typicky skóre v testu θ) a jednak na příslušnosti testované osoby k ohniskové či referenční skupině testovaných osob S (např. Wiberg, 2007; Karami, 2012). Následně jsou vytvořeny tři modely logistické regrese pro pravděpodobnost správné odpovědi testované osoby i na testovou položku m (např. Karami, 2012; Wiberg, 2007): Model (1): $\ln\left(\frac{P_{mi}}{1-P_{mi}}\right) = \beta_0 + \beta_1\theta_i,$ Model (2): $\ln\left(\frac{P_{mi}}{1-P_{mi}}\right) = \beta_0 + \beta_1\theta_i + \beta_2 S_i,$ Model (3): $\ln\left(\frac{P_{mi}}{1-P_{mi}}\right) = \beta_0 + \beta_1\theta_i + \beta_2 S_i + \beta_3\theta_i S_i,$ přičemž β_1 vyjadřuje vliv zvládnutí hodnoceného konstruktu testovanou osobou na pravděpodobnost její správné odpovědi na testovou položku m, β_2 vyjadřuje vliv uniformního DIF a β_3 vyjadřuje vliv neuniformního DIF. DIF analýza vychází z odhadu modelů (1) až (3), načež jsou primárně porovnány modely (2) a (3). Statistická významnost rozdílu ukazatele dobré shody (např. ukazatel -2LL), indikuje neuniformní DIF testové položky. Pokud není neuniformní DIF zaznamenán, jsou ve druhém kroku porovnány modely (1) a (2) a analogicky je posuzována přítomnost uniformního DIF testové položky. DIF analýza dále využívá hodnocení úrovně DIF, které je založeno na srovnání hodnot pseudo-R^2 modelů. Wiberg (2007), Lambert et al. (2018) uvádějí, že rozdíly menší než 0,035 indikují nízkou úroveň DIF, rozdíly v intervalu 0,035 až 0,070 středně silnou úroveň DIF a rozdíly vyšší než 0,070 vysokou úroveň DIF.

Metoda	Charakteristika metody
Metoda standardizovaných p-rozdílů	DIF analýza využívající metodu standardizovaných p-rozdílů je založena na výpočtu rozdílů v podílech správně odpovídajících testovaných osob ohniskové (P_{FK}) a referenční skupiny (P_{RK}), a to pro různé úrovně zvládnutí hodnoceného konstruktu (K), tj. typicky pro různé hodnoty dosaženého skóre v testu. Současně je zohledněn podíl testovaných osob pro jednotlivé úrovně zvládnutí hodnoceného konstruktu (W_K). Na tomto základě je definován vztah pro výpočet standardizovaných p-rozdílů (např. Wiberg, 2007; Karami, 2012): $$STD\ P = \sum_K W_K(P_{FK} - P_{RK})$$ Pro posouzení DIF testové položky je využita referenční hodnota ±0,10, kdy přítomnost DIF v testové položce je indikována při hodnotách vyšších než 0,10 či nižších než -0,10 (např. Wiberg, 2007; Karami, 2012). Uveďme, že za nevýhody této metody DIF analýzy je považována absence testů statistické významnosti, a dále pak požadavek na vysoký počet testovaných osob pro aplikaci metody (např. Karami, 2012).
Mantel-Haenszelův (MH) přístup	Mantel-Haenszelův (MH) přístup k DIF analýze je nejčastěji používaným neparametrickým přístupem k posouzení spravedlnosti testové položky (např. Wiberg, 2007). Podstata MH přístupu je založena na následujících krocích (např. Narayanan a Swaminathan, 1994; Wiberg, 2007; Michaelides, 2008): (a) V prvním kroku jsou testované osoby rozděleny do k kategorií podle úrovně zvládnutí hodnoceného konstruktu, typicky podle dosaženého skóre v testu. (b) Pro definované kategorie podle bodu (a) je vytvořeno k četnostních tabulek (2 krát 2), které uvádějí podíly testovaných osob ohniskové a referenční skupiny, kteří zodpověděli správně, respektive chybně, danou testovou položku. (c) Ve třetím kroku jsou odhadovány hodnoty dvou statistik označených jako MH_{χ}^2 a ΔMH, a to na bázi četností správných a nesprávných odpovědí testovaných osob ohniskové, respektive referenční skupiny. Výsledné hodnocení spravedlnosti testových položek je založeno na jejich kategorizaci do tří skupin (např. Yavuz et al., 2018; FDE, 2017; Wiberg, 2007; Karami, 2012; Dorans a Holland, 1992; Michaelides, 2008): (A) Testové položky s nízkou úrovní DIF jsou charakteristické statisticky nevýznamnou hodnotou MH_{χ}^2 nebo hodnoty |ΔMH| menší než 1. (B) Testové položky se střední úrovní DIF jsou charakteristické statisticky významnou hodnotou MH_{χ}^2 statistiky a hodnotou |ΔMH| v intervalu 1 až 1,5. (C) Testové položky s vysokou úrovní DIF jsou charakteristické statisticky významnou hodnotou MH_{χ}^2 statistiky a hodnotou |ΔMH| vyšší než 1,5. Mezi nevýhody MH přístupu k DIF analýze patří: (a) neschopnost uchopit neuniformní DIF (např. Wiberg, 2007); a (b) závislost χ^2 rozdělení četností na velikosti vzorku testovaných osob s hrozbou častého zamítnutí nulové hypotézy o spravedlnosti testové položky (nepřítomnost DIF). Na druhé straně Narayanan a Swaminathan (1994) označují MH přístup k DIF analýze za efektivní metodu neparametrického charakteru, která nevyžaduje ověřování striktních požadavků na parametry modelu (např. také Wiberg, 2007).

- *Hodnocení neobvyklého vzoru odpovědí testovaných osob*

Záměrem hodnocení vzoru odpovědí testovaných osob na testové položky je identifikovat neobvyklé vzory odpovědí, které mohou vznikat z různých důvodů, jako jsou (např. Tendeiro, Meijer a Niessen, 2016; DeMars, 2010):

- neetické chování testovaných osob, včetně znalosti testových položek dopředu;
- nízká motivace testovaných osob řešit test vedoucí k hádání odpovědí;
- nedostatečné znalosti testovaných osob v některé z testovaných oblastí.

Hodnocení neobvyklosti vzoru odpovědí testované osoby na testové položky typicky posuzuje, zda se její vzor odpovědí významně odlišuje od ideálního vzoru odpovědí, přičemž je předpokládána vyšší pravděpodobnost správné odpovědi testované osoby na méně obtížné testové položky a naopak vyšší pravděpodobnost nesprávné odpovědi testované osoby na více obtížné testové položky (např. Tendeiro, Meijer a Niessen, 2016). Tabulka č. 4 představuje statistiky využívané pro tento účel.

Tabulka č. 4: Statistiky pro hodnocení neobvyklého vzoru odpovědí testovaných osob

Statistika	Charakteristika statistiky
r.pbis	Statistika *r.pbis* vyjadřuje hodnotu bodově biseriální korelace mezi odpověďmi testované osoby na testové položky a obtížností testových položek (podíly správných odpovědí na testové položky). Nízké hodnoty *r.pbis* naznačují vyšší neobvyklost vzoru odpovědí testované osoby.
C, C a U3*	Statistiky *C, C** a *U3* jsou založeny na myšlence srovnání vzoru odpovědí testovaných osob na testové položky a ideálního vzoru odpovědí testovaných osob na testové položky (tzv. Guttmanův vzor odpovědí). Guttmanův vzor odpovědí vychází ze seřazení testových položek od nejjednodušší k nejvíce obtížné s tím, že ideální vzor odpovědí testované osoby s *n* správnými odpověďmi, tj. Guttmanův vzor odpovědí, je spojen se správnými odpověďmi testované osoby na prvních *n* nejjednodušších testových položek a s nesprávnými odpověďmi na testové položky ostatní (obtížnější). Platí, že čím více se vzor odpovědí testované osoby odlišuje od Guttmanova vzoru odpovědí, tím více neobvyklý tento vzor odpovědí je. Statistika *C* je přitom počítána vztahem: $$C = 1 - \frac{cov(x_{ni}, p_i)}{cov(x_{ni}^*, p_i)},$$ kde x_{ni} je vzor odpovědí testované osoby *n* na testové položky *i*; x_{ni}^* je Guttmanův vzor odpovědí testované osoby *n* na testové položky *i*; p_i je podíl správných odpovědí na testovou položku *i*. Skutečný vzor odpovědí odpovídá ideálnímu vzoru odpovědí při hodnotě nula. Statistika C^* převádí hodnoty statistiky *C* do intervalu od nuly (nejvíce obvyklý vzor odpovědí) do jedné (nejméně obvyklý vzor odpovědí) vztahem: $$C^* = \frac{cov(x_{ni}^*, p_i) - cov(x_{ni}, p_i)}{cov(x_{ni}^*, p_i) - cov(\widehat{x_{ni}}, p_i)},$$ kde x_{ni} se stříškou je opačný vektor Guttmanova vzoru odpovědí. Statistika *U3* se pak od statistiky C^* odlišuje svým logaritmickým tvarem.

3.1.2 Společná škála testů, equating

V praktických úlohách se lze často setkat se situacemi, kdy je potřeba reportovat dosažené výsledky testovaných osob na stejné škále (např. Ryan a Brockmann, 2009). Typickým příkladem takové situace je využití vyššího počtu odlišných testů v případě rozhodného (*high-stake*) testování (např. Lamprianou, 2007; Livingston, 2014).[22] Propojení testů na společnou škálu, tzv. *equating*[23], vede k odstranění znevýhodnění, které je dáno řešením obtížnějšího testu (např. Dorans, Moses a Eignor, 2010; Lamprianou, 2007; Livingston, 2014).

Equating je tedy metodický postup využívaný pro přizpůsobení škál různých testů měřících stejný konstrukt tak, aby tyto škály byly vzájemně zaměnitelné (např. Ryan a Brockmann, 2009; Lamprianou, 2007). Takto dochází k situaci, kdy úroveň zvládnutí hodnoceného konstruktu testovanou osobou na škále jednoho testu má odpovídající úroveň zvládnutí hodnoceného konstruktu na škále druhého testu, přičemž za tímto účelem lze využít různé typy škál (např. Livingston, 2014). *Equating* se může uplatnit v řadě situací, ke kterým také patří (např. Ryan a Brockmann, 2009):

- propojení škál dvou různých testů stejného programu hodnocení;
- propojení škál starší a novější verze testů;
- utváření vertikální škály měřící úroveň pokroku testované osoby v čase,
- propojení škál národních a mezinárodních šetření.

Metodický postup pro propojování škál testů měřících stejný konstrukt primárně vychází z plánu sběru dat (tzv. *equating design*). Livingston (2014), Cook a Eignor (1991), Dorans, Moses a Eignor (2010), Lamprianou (2007), Ryan a Brockmann (2009) v tomto ohledu uvádějí tři základní přístupy:

- První přístup je založený na jedné skupině testovaných osob (SG), které řeší nový i původní (referenční) test, přičemž poznatky jsou zobecněny na celou populaci testovaných osob.
- Druhý přístup je založený na výběru dvou ekvivalentních skupin testovaných osob (EG), tj. skupin testovaných osob ze stejné populace a s odpovídajícím si rozdělením úrovně zvládnutí hodnoceného konstruktu, přičemž každá z těchto skupin testovaných osob řeší jiný test.
- Třetí přístup využívá pro dvě neekvivalentní skupiny testovaných osob (vzhledem k populaci i vzhledem k úrovni zvládnutí hodnoceného konstruktu) dva různé testy, které

[22] Cook a Eignor (1991) uvádějí, že jedním z motivů pro využívání vyššího počtu různých testů je snižování rizika zveřejnění zadání testů ještě před jejich řešením (např. také Sansivieri, Wiberg a Matteucci, 2018). Sansivieri, Wiberg a Matteucci (2018) pak přidávají další dva motivy: (a) stále častější požadavek na zveřejňování příkladů testových položek; a (b) měnící se obsahové zaměření testů při změně standardů/obsahu kurikula.

[23] Dorans, Moses a Eignor (2010) upozorňují na skutečnost, že zastřešujícím pojmem pro propojování škál testů je pojem *linking* s tím, že *equating* je jeho specifický typ s nejsilnějšími požadavky. Mezi další typy propojování škál testů (*linking*) řadí Dorans, Moses a Eignor (2010): (a) predikci skóre testované osoby z informací testu jiného, případně z dalších informací; a (b) tzv. *scale aligning* chápaný ve smyslu propojení dvou škál různých testů, které hodnotí odlišné konstrukty (např. rovněž Ryan a Brockmann, 2009). Zároveň Ryan a Brockmann (2009) upozorňují na skutečnost, že oba pojmy *linking* a *equating* jsou v praxi často zaměňovány, neboť využívají analogické metody výpočtů vedoucí k vytvoření dvojic odpovídajících si skóre dvou testů.

však obsahují určitý počet stejných, tzv. kotvících, testových položek (NEAT). Kotvící položky, které společně utvářejí tzv. kotvící test, pomáhají odstranit vliv odlišné obtížnosti testu, přičemž jejich potřebnost je dána neekvivalentním charakterem výběrových souborů testovaných osob.[24]

Tabulka č. 5: Výhody a nevýhody různých přístupů k plánu sběru dat pro *equating* testů

Plán sběru dat	Výhody	Nevýhody
SG	(a) Přístup kontroluje vliv charakteristik testovaných osob (např. Dorans, Moses a Eignor, 2010). (b) Přístup vyžaduje spíše nižší počet zapojených testovaných osob (např. Dorans, Moses a Eignor, 2010).	(a) Testovaná osoba musí absolvovat dva testy (např. Livingston, 2014; Ryan a Brockmann, 2009). (b) Na výsledek může mít vliv pořadí testů působením faktorů učení a klesající motivace testovaných osob (např. Ryan a Brockmann, 2009).
EG	(a) Testovaná osoba řeší pouze jeden test, čímž odpadá i nežádoucí vliv pořadí testů (např. Dorans, Moses a Eignor, 2010).	(a) Přístup vyžaduje spíše vysoký počet zapojených testovaných osob (např. Ryan a Brockmann, 2009; Cook a Eignor, 1991). (b) Přístup klade náročné požadavky na ekvivalentnost dvou výběrových souborů testovaných osob (např. Ryan a Brockmann, 2009; Cook a Eignor, 1991).
NEAT	(a) Testovaná osoba řeší pouze jeden test, čímž odpadá také nežádoucí vliv pořadí testů (např. Ryan a Brockmann, 2009; Livingston, 2014). (b) Přístup kontroluje vliv charakteristik testovaných osob kotvícím testem, a proto nevyžaduje ekvivalentnost výběrových souborů testovaných osob (např. Ryan a Brockmann, 2009; Livingston, 2014).	(a) Přístup klade vyšší nároky na přípravu testů (např. Livingston, 2014). (b) Přístup klade vysoké nároky na kvalitu kotvícího testu, a proto je potřeba vzít do úvahy hrozby ovlivňující kvalitu kotvícího testu, jako je například obecná změna obtížnosti kotvících testových položek v čase s ohledem na změnu jejich důležitosti v kurikulu a výuce (např. Ryan a Brockmann, 2009; Livingston, 2014).

[24] Pro tvorbu kotvících testů lze identifikovat řadu doporučení, které vycházejí z teze, že kotvící test by měl být ideálně mini-verzí obou propojovaných testů, a to vzhledem k obsahu testu (měřený konstrukt), k typům testových položek či k vlastnostem testových položek (např. Lamprianou, 2007; Ryan a Brockmann, 2009; Livingston, 2014; Dorans, Moses a Eignor, 2010; Cook a Eignor, 1991). Livingston (2014) dále doporučuje, aby: (a) testové položky kotvícího testu nebyly řazeny na konec testů pro předcházení negativního vlivu časového tlaku; (b) testové položky kotvícího testu byly v obou řešených testech umístěny na přibližně stejném místě pro předcházení negativního vlivu klesající motivace (např. také Lamprianou, 2007; Ryan a Brockmann, 2009; Dorans, Moses a Eignor, 2010); (c) testové položky kotvícího testu pokrývaly široké spektrum obtížností pro možnost rozlišit dobré i slabší testované osoby (např. také Ryan a Brockmann, 2009); a (d) pro delší testy (např. kolem 100 testových položek) bylo k dispozici 15-20 testových položek kotvícího testu (např. také Ryan a Brockmann, 2009) s tím, že Cook a Eignor (1991) doporučují, aby nejméně 20 % testových položek bylo součástí kotvícího testu. Konečně také platí, že testové položky kotvícího testu mají vyšší kvalitu v případě vysoké korelace se skóre obou hodnocených testů (např. Livingston, 2014; Dorans, Moses a Eignor, 2010).

Tabulka č. 5 shrnuje hlavní výhody a nevýhody tří přístupů pro propojování škál testů měřících stejný konstrukt. S ohledem na nepraktičnost přístupu založeného na jedné skupině testovaných osob (SG přístup) a na nereálnost požadavků kladených na ekvivalentnost dvou výběrových souborů testovaných osob (EG přístup) označují Sansivieri, Wiberg a Matteucci (2018) za nejvhodnější třetí přístup založený na dvou neekvivalentních skupinách testovaných osob a na využití kotvících testových položek (NEAT přístup).

Pro vlastní propojení škál testů lze využít různé přístupy, přičemž při jejich volbě hrají roli především tři dílčí aspekty (např. Dorans, Moses a Eignor, 2010):

- zvolený plán sběru dat s rozlišením SG a EG přístupu na jedné straně a NEAT přístupu na straně druhé;
- volba přístupu založeného na skutečných skóre testů, nebo na pozorovaných skóre testů, přičemž však skutečná skóre testů jsou neznámá, a jedná se proto spíše o teoretický koncept;
- zvolený způsob transformace skóre testů s volbou lineární, nebo nelineární transformace.

V případě SG a EG přístupu ke sběru dat je aplikován postup propojení škál testů, který je založen na společné populaci testovaných osob. Klíčovým metodickým prvkem postupu je kumulativní distribuční funkce, která vyjadřuje podíl testovaných osob dosahujících v testu dané skóre či skóre nižší a která je rostoucí funkcí nabývající hodnot v intervalu od nuly do jedné (např. Dorans, Moses a Eignor, 2010). Tabulka č. 6 charakterizuje další kroky propojení škál testů, a to s využitím buď ekvipercentilního, nebo lineárního přístupu.

Tabulka č. 6: Ekvipercentilní a lineární *equating* pro společné populace testovaných osob

Equating	Charakteristika
Ekvipercentilní	Ekvipercentilní přístup k propojování škál testů vychází z předpokladu, že hodnoty skóre nového (x) a referenčního (y) testu, které odpovídají stejnému percentilu jejich kumulativních distribučních funkcí $F_T(x)$ a $G_T(y)$ pro společnou populaci testovaných osob T, jsou ekvivalentní (např. Livingston, 2014; Sansivieri, Wiberg a Matteucci, 2018; Ryan a Brockmann, 2009). Pro nalezení odpovídajících si dvojic skóre je následně využita funkce (např. Sansivieri, Wiberg a Matteucci, 2018): $$y = {G_T}^{-1}F_T(x),$$ kde G_T^{-1} je inverzní funkce ke kumulativní distribuční funkci $G_T(y)$. Odpovídající si hodnoty skóre obou testů jsou tedy odvozeny z kumulativních funkcí.
Lineární	Lineární přístup k propojování škál testů předpokládá, že kumulativní distribuční funkce $F_T(x)$ a $G_T(y)$ mají stejný tvar a liší se pouze v průměru a směrodatné odchylce (např. Ryan a Brockmann, 2009; Livingston, 2014; Sansivieri, Wiberg a Matteucci, 2018). Pro nalezení odpovídajících si dvojic skóre je následně využita funkce (např. Sansivieri, Wiberg a Matteucci, 2018; Ryan a Brockmann, 2009): $$y = \bar{y}_T + \frac{\sigma_{YT}}{\sigma_{XT}}(x - \bar{x}_T).$$ V případě shodnosti průměrů i směrodatných odchylek je propojení škál nového a referenčního testu identitou (např. Dorans, Moses a Eignor, 2010).

Postup propojení škál testů na bázi NEAT přístupu ke sběru dat lze ve srovnání se SG a EG přístupem charakterizovat jako více komplexní, neboť nepracuje s předpokladem stejné úrovně zvládnutí hodnoceného konstruktu dvou výběrových souborů testovaných osob, nýbrž využívá kotvící test (např. Livingston, 2014).[25] Propojení škál testů je v tomto případě typicky spojeno s vytvořením tzv. syntetické populace T ze dvou různých populací výběrových souborů testovaných osob X a Y, a to prostřednictvím vztahu (např. Dorans, Moses a Eignor, 2010; Livingston, 2014):

$$T = w_1X + w_2Y, kde\ w_1 + w_2 = 1,$$

kde w_i je váha každé populace testovaných osob, která je často stanovena proporčně k velikostem obou populací. Zdůrazněme, že propojení škál testů pro NEAT přístup vychází z předpokladu, že vlastnosti, které propojují skóre dvou testů, platí pro každou syntetickou populaci T (např. Dorans, Moses a Eignor, 2010; Livingston, 2014).

Dorans, Moses a Eignor (2010), Livingston (2014) rozlišují dva základní metodické přístupy pro propojení škál testů na bázi NEAT přístupu ke sběru dat. První metodický přístup je tzv. řetězový *equating*, který je založen na myšlence, že skóre nového testu jsou propojena se skóre kotvícího testu a následně jsou skóre kotvícího testu propojena se skóre referenčního testu. Přes kotvící test je rovněž zajištěno propojení škál nového a referenčního testu, přičemž platí, že propojení škál nového a kotvícího testu, stejně jako škál referenčního a kotvícího testu, je stejné pro každou syntetickou populaci testovaných osob T (např. Livingston, 2014; Dorans, Moses a Eignor, 2010). Při utváření řetězce skóre nového, kotvícího a referenčního testu lze využít jak ekvipercentilní přístup, tak lineární přístup (např. Livingston, 2014; Sansivieri, Wiberg a Matteucci, 2018; Lamprianou, 2007).

Druhý metodický přístup je tzv. post-stratifikační *equating*, který je založený na předpokladu, že rozdělení skóre nového i referenčního testu podmíněné dosaženým skóre kotvícího testu je stejné pro každou syntetickou populaci testovaných osob T (např. Livingston, 2014; Dorans, Moses a Eignor, 2010). Pro post-stratifikační *equating* je proto potřeba zajistit informaci o rozdělení skóre testovaných osob syntetické populace T v novém a referenčním testu, a to s využitím skóre v kotvícím, tj. ve společném, testu. Následně je možné opětovně využít jak ekvipercentilní přístup, tak lineární přístup pro propojení škál testů (např. Livingston, 2014).

Postup propojení škál testů se může setkat s řadou metodických otázek spojených mimo jiné: (a) s diskrétním charakterem dosažených skóre v testu; (b) s výraznými nepravidelnostmi v rozdělení dosažených skóre v testu, případně s vyšší četností výskytu určitých skóre v testu; a (c) s problematickým nalezením odpovídajících si skóre v testu pro málo početné extrémní hodnoty či hodnoty vůbec se nevyskytující (např. Livingston, 2014).[26] Pro snižování negativních vlivů uvedených skutečností doporučují Livingston (2014), Dorans, Moses

[25] Livingston (2014) vysvětluje tuto úvahu s využitím následujícího příkladu. Předpokládejme dva výběrové soubory testovaných osob, z nichž jeden zaznamenává skóre: (a) v novém testu; a (b) v kotvícím testu, zatímco druhý soubor testovaných osob zaznamenává skóre: (a) v referenčním testu; a (b) v kotvícím testu. Následně platí, že testované osoby s vysokým skóre v kotvícím testu a nízkým skóre v (celém) novém testu řeší obtížný test, zatímco testované osoby s nízkým skóre v kotvícím testu a vysokým skóre v (celém) referenčním testu řeší jednoduchý test. *Equating* slouží ke kompenzaci rozdílů v obtížnosti testů.

[26] Livingston (2014) v tomto ohledu označuje za nejvíce problematické testy řešené vysokým počtem testovaných osob s nízkým počtem možných skóre k dosažení.

a Eignor (2010) aplikaci metod jádrového vyhlazení, a to včetně vyhlazení nepravidelností v rozdělení dat a doplnění dat tam, kde žádná testovaná osoba nedosáhla daného skóre (např. extrémně vysoká či nízká skóre). Zároveň však je potřeba vzít do úvahy nevýhody takového postupu spojené se ztrátou části informace (např. Dorans, Moses a Eignor, 2010).

Konečně uveďme, že pro reporting výsledků testovaných osob lze definovat ještě další škálu (např. bodová škála v intervalu hodnot 200 až 800), což je typicky motivováno záměrem vyhnout se reportingu založenému na dosaženém skóre (např. Livingston, 2014). Při sledování přístupu „třetí" škály je potřeba rozhodnout o následujících otázkách (např. Livingston, 2014; Dorans, Moses a Eignor, 2010):

- rozhodnutí o jednotce škály (např. 1 bod, 10 bodů);[27]
- rozhodnutí o horní a dolní hranici škály,[28]
- rozhodnutí o podobě škály ve vazbě na dosažená skóre,[29]
- rozhodnutí o reportingu výsledků rovněž s využitím širokých kategorií úrovní.

3.1.3 Hodnocení pokroku ve vzdělávání

V teorii i praxi tvorby a hodnocení testů roste zájem o časový aspekt problematiky, který je spojený s měřením vzdělávacího pokroku testovaných osob (např. Betebenner, 2009; O'Malley et al., 2011). Hlavní otázka se v tomto ohledu ptá: „Jaký pokrok testovaná osoba ve svých vzdělávacích výsledcích zaznamenala a je na správné cestě k dosažení očekávaných vzdělávacích výsledků?" V kontextu zvyšující se pozornosti věnované uvedeným otázkám hovoří Thompson (2015), Betebenner (2009) o posunu paradigmatu od hodnocení statusu, tj. stavu v daném čase, k hodnocení pokroku, tj. změn v čase. Využívány jsou za tímto účelem různé metodické přístupy, pro něž jsou však některé aspekty typické (Betebenner a Linn, 2010):

- Pro možnost hodnocení pokroku ve vzdělávání je potřebné mít k dispozici informace o vzdělávacích výsledcích testované osoby ve více časových okamžicích (longitudinální data), a to například ve více ročnících studia (např. také Briggs a Domingue, 2013; Udofia a Uko, 2016). V tomto ohledu je rovněž klíčová kvalita využívaného nástroje hodnocení (např. Schafer, 2006).
- Vzdělávací pokrok testované osoby je možné hodnotit s využitím různých typů škál (např. percentily, skóre, kategorie úspěšnosti), jejichž kvalita je zásadním předpokladem kvality hodnocení vzdělávacího pokroku testované osoby (např. O'Malley et al., 2011; Udofia a Uko, 2016).

[27] Livingston (2014) upozorňuje na hrozbu nesprávného vnímání vzdáleností na škále, kdy rozdíly mohou být považovány za velmi velké, přičemž skutečnost je odlišná.

[28] Livingston (2014) uvádí dva motivy těchto úvah. První motiv je spojený s otázkou, zda testovaná osoba dosahující skóre 100 % v lehčím testu má dosáhnout stejné hodnoty skóre jako testovaná osoba dosahující skóre 100 % v těžším testu. Druhý motiv je spojený se snahou předejít vnímání výsledku testované osoby, který byl dosažen hádáním, jako dobrého výsledku výrazně nad hodnotou nejnižšího možného, tj. nulového, skóre.

[29] Livingston (2014) uvádí možnosti: (a) využití dosaženého průměru (střední hodnota škály) a směrodatné odchylky (rozpětí škály) dosažených skóre testovaných osob; a (b) definice nejvyšší a nejnižší hodnoty škály s lineárním rozdělením dalších skóre v rámci takto vymezené škály.

- Podstata hodnocení vzdělávacího pokroku testované osoby je založena na výpočtu rozdílů vzdělávacích výsledků, kterých tato osoba dosáhla v různých časových okamžicích na zvolené škále, případně na hodnocení posunu testované osoby mezi definovanými kategoriemi úspěšnosti (např. také Briggs, 2013; Schafer, 2006; O'Malley et al., 2011).

Pokrok ve vzdělávání může být hodnocen s využitím různých metodických přístupů, mezi které především patří: (a) měření rozdílů vzdělávacích výsledků na vertikální škále; (b) model přidané hodnoty; a (c) žákovský percentil růstu. Tabulka č. 7 blíže představuje podstatu těchto metodických přístupů.

Tabulka č. 7: Metodické přístupy k hodnocení vzdělávacího pokroku testované osoby

Metodický přístup	Charakteristika metodického přístupu
Měření rozdílů ve vzdělávacích výsledcích na vertikální škále	Podstata metodického přístupu k hodnocení vzdělávacího pokroku testované osoby, který je založený na měření rozdílů jí dosažených vzdělávacích výsledků na vertikální škále, vychází z propojení škál testů, které jsou za tímto účelem využity. Testy jsou přitom konstruovány s předpokladem zvyšující se úrovně zvládnutí hodnoceného konstruktu v čase, tj. s rostoucí obtížností testu (např. Udofia a Uko, 2016), přičemž pro propojení škál testů je využíván kotvící test a NEAT přístup ke sběru dat (např. Briggs, 2013; Schafer, 2006). Udofia a Uko (2016) uvádějí, že vertikální škála funguje lépe v případě, že rozdíly uvnitř skupiny testovaných osob (např. ročník žáků) jsou relativně vysoké, zatímco rozdíly mezi skupinami testovaných osob relativně nízké. Využití vertikální škály tak je méně vhodné pro propojení skóre testů zadávaných v dlouhém časovém odstupu. Vzdělávací pokrok testované osoby pak je logicky měřen a reportován jejím posunem na vertikální škále (např. Udofia a Uko, 2016; Betebenner a Linn, 2010). Popsaný způsob využití vertikální škály pro hodnocení vzdělávacího pokroku testované osoby se ovšem setkává se silnou kritikou, a to především ze dvou důvodů: (a) První důvod je spojený s otázkou, zda vertikální škála je intervalového typu (např. Briggs, 2013; Ballou, 2009; Schafer, 2006). V případě nenaplnění tohoto předpokladu platí, že stejné rozdíly dvou různých dvojic hodnot na vertikální škále nejsou shodné i ve skutečnosti, což následně zpochybňuje informaci o metrice posunu testovaných osob na této škále. (b) Druhý důvod je spojený s narušením předpokladu unidimenzionality testů, tj. předpokladu, že testy hodnotí (měří) stejný konstrukt. Hrozba tohoto problému se zdá být opodstatněná tím, že obsah výuky v různých ročnících se odlišuje, přičemž tato skutečnost má svůj vliv rovněž na kvalitu fungování testových položek kotvícího testu.

Metodický přístup	Charakteristika metodického přístupu
Model přidané hodnoty	Koedel, Mihaly a Rockoff (2015) označují model přidané hodnoty za klíčový metodický nástroj pro hodnocení pokroku ve vzdělávání. Podstata modelu spočívá v modelování přidané hodnoty vzdělávání mezi dvěma časovými okamžiky typicky ve vazbě na další vstupy do vzdělávání, ke kterým především patří práce učitele (např. Betebenner a Linn, 2010; Thompson, 2015; Braun, 2005; Briggs a Domingue, 2013; O'Malley et al., 2011). Koedel, Mihaly a Rockoff (2015), Briggs a Domingue (2013) uvádějí jednu z používaných podob modelu přidané hodnoty vztahem: $$Y_{isjt} = \beta_0 + Y_{isjt-1}\beta_1 + X_{isjt}\beta_2 + S_{isjt}\beta_3 + T_{isjt}\theta + \varepsilon_{isjt},$$ kde Y_{isjt} je výsledek testované osoby *i* ze skupiny *s* učitele *j* v testu v čase *t*; Y_{isjt-1} je výsledek testované osoby *i* ze skupiny *s* učitele *j* v testu v čase *t-1*; kde X_{isjt} jsou další charakteristiky testované osoby (např. pohlaví, etnický původ); kde S_{isjt} jsou další charakteristiky skupiny (např. průměrné skóre všech testovaných osob skupiny v testu); a kde T_{isjt} jsou další charakteristiky učitele. Za pozornost stojí, že v uvedeném vztahu je využíván časově opožděný výsledek testovaných osob v testu, který kontroluje výchozí úroveň zvládnutí hodnoceného konstruktu testovanou osobou, přičemž díky koeficientu β_1 není pro odhad modelu přidané hodnoty vyžadována vertikální škála, tj. propojení využívaných testů na společnou škálu (např. Betebenner a Linn, 2010; Briggs a Domingue, 2013). Koedel, Mihaly a Rockoff (2015) pak přidávají některé další metodické poznámky k modelu přidané hodnoty: • Zahrnutí dalších charakteristik testovaných osob a skupin do modelu je motivováno skutečností, že tito nejsou učitelům přiřazování náhodným výběrem, a proto je potřeba kontrolovat jejich vliv na pokrok ve vzdělávání (např. také Betebenner a Linn, 2010; Braun, 2005; O'Malley et al., 2011). Koedel, Mihaly a Rockoff (2015) nicméně uvádějí, že oproti výsledkům testovaných osob v dřívějších testech je vliv dalších charakteristik omezený. • Kvalitě odhadů modelu přidané hodnoty napomáhá, pokud je do modelu zahrnuto více proměnných vztahujících se k výsledkům testovaných osob v dřívějších testech, a to dokonce i v testech, které primárně neměří hodnocený konstrukt. • Vedle představeného jednoúrovňového modelu přidané hodnoty lze tyto modely odhadovat také jako víceúrovňové, kdy například efekt učitele či skupiny je odhadován teprve na druhé úrovni z reziduí modelu, který nezahrnuje tento efekt. Koedel, Mihaly a Rockoff (2015) nicméně poukazují na podobnost závěrů, které oba typy modelů přinášejí. • Model přidané hodnoty stále častěji využívá Bayesovy přístupy k redukci vlivu extrémních hodnot vyhlazením. Přínos tohoto kroku se zdá být vyšší, než je velikost chyby vznikající z vynechání části informace. Alternativní modely přidané hodnoty (např. Tennessee model) představují například Thompson (2015), Braun (2005), Briggs a Domingue (2013).

Metodický přístup	Charakteristika metodického přístupu
Studentský percentil růstu	Podstata metody studentského percentilu růstu (také Colorado model) spočívá ve srovnání výsledků testované osoby v testu s výsledky jiných testovaných osob, a to těch, které v dřívějších testech dosáhly podobného výsledku jako testovaná osoba, o kterou má hodnocení zájem (např. Thompson, 2015). Z takto provedeného srovnání je stanoven percentil, jakého testovaná osoba dosáhla mezi testovanými osobami podobných výsledků (podmíněné hodnocení). Vysoké hodnoty percentilu znamenají relativně vysoký pokrok testované osoby, zatímco nízké hodnoty percentilu relativně nízký pokrok testované osoby (např. Betebenner a Linn, 2010). Vlastní výpočet percentilu pak je založen na odhadu 100 regresních modelů (kvantilová regrese), kde kvantily odpovídají hodnotám příslušných skóre testované osoby v aktuálním testu a výsledky testovaných osob v dřívějších testech jsou využity jako nezávislé proměnné. Výsledky testované osoby v dřívějších testech slouží pro výpočet příslušného percentilu (např. Betebenner, 2009).

3.2 Teorie odpovědi na položku

Thorpe a Favia (2012), Toland (2014), van der Linden (2010) označují teorii odpovědi na položku (dále jen „IRT“) za moderní a dynamicky se rozvíjející přístup k hodnocení a tvorbě testů, Thompson (2009) pak za současné teoreticko-metodické paradigma v této oblasti (např. také Zhang, 2013 pro označení IRT za široce aplikovaný teoreticko-metodický přístup k práci s testy). Hambleton a Jones (1993) doplňují, že zájem o IRT byl posílen skutečností, že při splnění stanovených podmínek nezávisí statistiky vycházející z IRT na výběrovém souboru testovaných osob, což není případ základních charakteristik CTT.

Podstata IRT spočívá v modelování vztahu mezi úrovní zvládnutí hodnoceného konstruktu testovanou osobou, která je tradičně označena písmenem řecké abecedy Θ, a vzorem odpovědí této osoby na testové položky (např. DeMars, 2010; Hambleton a Jones, 1993; Toland, 2014; Orlando a Thissen, 2000). De Champlain (2010) uvádí, že IRT je spojena s odhadem pravděpodobnosti správné odpovědi testované osoby na testovou položku v závislosti na: (a) charakteristikách testové položky (obtížnost testové položky, diskriminace testové položky); a (b) úrovni zvládnutí hodnoceného konstruktu Θ testovanou osobou. V tomto kontextu jsou dobře srozumitelné také rozdíly IRT oproti CTT, které především spočívají (např. Thompson, 2009):

- v měření zvládnutí hodnoceného konstruktu testovanou osobou nikoliv hodnotou dosaženého skóre v testu, ale latentní (nepozorovanou) spojitou proměnnou Θ, která je modelována ze vzoru odpovědí testované osoby na jednotlivé testové položky;[30]
- ve vyšším důrazu kladeném na hodnocení jednotlivých testových položek.

Vlastní metrika zvládnutí hodnoceného konstruktu Θ může být definována různě, základem jejího utváření je fixace středu (např. nulová hodnota) a fixace jednotky vzdálenosti (např.

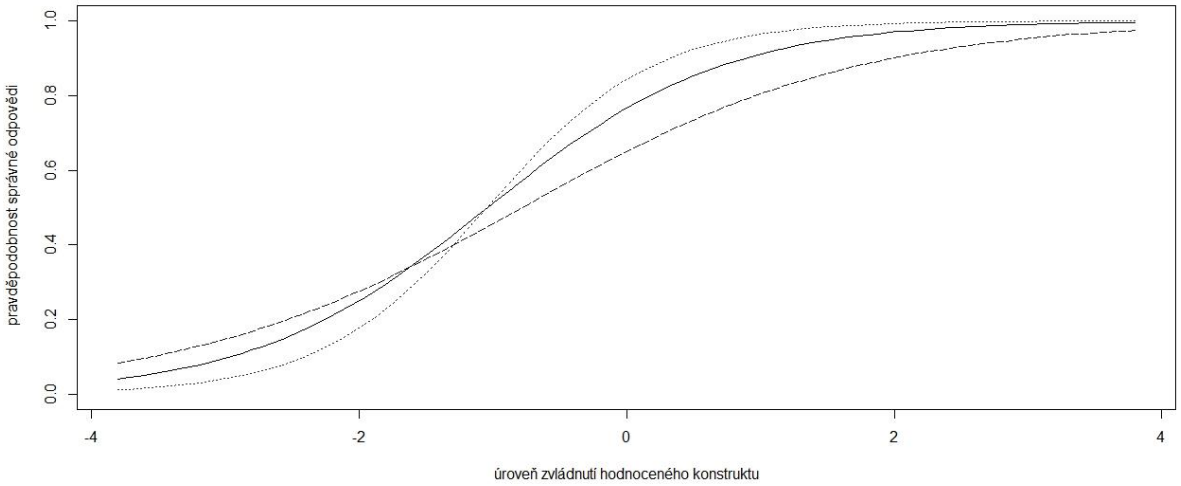
[30] I v tomto případě se jedná o konstrukt, který není pozorován přímo, ale je měřen prostřednictvím testu.

hodnota směrodatné odchylky).[31] Hodnoty Θ se teoreticky mohou pohybovat v intervalu od -∞ do +∞, nicméně typicky se pracuje s hodnotami ve výrazně kratším intervalu (viz také pojednání o obtížnosti a diskriminaci testové položky v následující podkapitole).

3.2.1 Analýza testových položek

IRT analyzuje testové položky s využitím charakteristik analogických k CTT, nicméně zásadním způsobem se odlišuje způsob jejich konstrukce. Charakteristická křivka testové položky (ICC), která zachycuje pravděpodobnost jejího správného zodpovězení testovanou osobou v závislosti na její úrovni zvládnutí hodnoceného konstruktu Θ (viz obrázek č. 2), je základním nástrojem grafické prezentace výstupů analýzy testových položek (např. Edelen a Reeve, 2007; Thorpe a Favia, 2012; Toland, 2014).[32] V případě dichotomických testových položek má ICC rostoucí tvar, kdy při vyšší úrovni zvládnutí hodnoceného konstruktu Θ je také vyšší pravděpodobnost úspěšného zodpovězení dané testové položky (např. Toland, 2014; Edelen a Reeve, 2007; DeMars, 2010). Technicky je ICC vyhlazením pozorovaných podílů správně odpovídajících testovaných osob na danou testovou položku v závislosti na intervalech úrovně zvládnutí hodnoceného konstruktu Θ (např. DeMars, 2010). Za tímto účelem jsou typicky odhadovány parametry obtížnosti, diskriminace, případně jiných charakteristik testových položek, a to s využitím vhodného IRT modelu.

Obrázek č. 2: Charakteristická křivka testové položky (ICC)



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

[31] Obecně však lze využít libovolnou škálu, kdy například střed odpovídá hodnotě 500 a vzdálenost hodnotě 100 (např. Toland, 2014).

[32] Hambleton a Jones (1993) poukazují na skutečnost, že ICC de facto spojuje přístup na bázi IRT (úroveň zvládnutí hodnoceného konstruktu Θ) a přístup na bázi CTT (úspěšnost v řešení testové položky).

- *Obtížnost testové položky*

Obtížnost testové položky je v IRT definována s využitím metriky úrovně zvládnutí hodnoceného konstruktu Θ. Takto obtížnost testové položky odpovídá hodnotě, na které přibližně 50 % testovaných osob zodpovídá testovou položku správně (např. DeMars, 2010; Thorpe a Favia, 2012; De Champlain, 2010; Toland, 2014).[33] Platí, že čím vyšší je hodnota obtížnosti testové položky, tím je také tato testová položka těžší. V tomto se ukazatel obtížnosti testové položky v IRT liší od ukazatele obtížnosti testové položky v CTT, kde platí, že obtížnější testové položky mají nižší hodnotu ukazatele obtížnosti, tj. podílu správně odpovídajících testovaných osob (např. DeMars, 2010). Hodnoty obtížnosti testové položky se teoreticky mohou pohybovat v intervalu od -∞ do +∞, nicméně typický je interval od -2 do +2 tak, aby testové položky nebyly ani příliš těžké, ani příliš jednoduché (např. Toland, 2014; DeMars, 2010). Toland (2014) pak uvádí interval od -3 do + 3 s tím, že hodnoty mimo tento interval ukazují na podezřelé testové položky.

- *Diskriminace testové položky*

V IRT vyjadřuje diskriminace testové položky rychlost změny pravděpodobnosti, že testovaná osoba zodpoví danou testovou položku správně, a to v závislosti na úrovni zvládnutí hodnoceného konstruktu Θ. Hodnota diskriminace tak sděluje, jak dobře je testová položka schopna rozlišit mezi těmi testovanými osobami, které daný konstrukt znají dobře a těmi testovanými osobami, které daný konstrukt znají hůře či neznají vůbec (např. DeMars, 2010). Edelen a Reeve (2007), Toland (2014) doplňují, že diskriminace charakterizuje sílu vztahu testové položky ke škále, která měří hodnocený konstrukt (Θ). V případě IRT, stejně jako v případě CTT, platí, že lepší schopnost rozlišit mezi testovanými osobami s různou úrovní zvládnutí hodnoceného konstruktu Θ mají testové položky s vyšší hodnotou diskriminace (vyšší sklon ICC v obrázku č. 2), přičemž testová položka diskriminuje nejlépe na úrovni odpovídající její obtížnosti (např. DeMars, 2010; Thorpe a Favia, 2012).

Také hodnoty diskriminace testové položky mohou teoreticky spadat do intervalu od -∞ do +∞, obecně jsou však za nekvalitní testové položky považovány ty, které mají nízkou schopnost diskriminace mezi testovanými osobami podle úrovně jejich zvládnutí hodnoceného konstruktu Θ (např. DeMars, 2010). Za takové jsou uváděny hodnoty nižší než 0,4 (např. DeMars, 2010), případně 0,5 (např. Toland, 2014). Thorpe a Favia (2012) pak uvádějí následující klasifikaci testových položek podle hodnoty diskriminace:

- hodnoty nižší než 0,35 – velmi nízká schopnost diskriminace;
- hodnoty v rozmezí 0,35 až 0,64 – nízká schopnost diskriminace;
- hodnoty v rozmezí 0,65 až 1,34 – průměrná schopnost diskriminace;
- hodnoty v rozmezí 1,35 až 1,69 – vysoká schopnost diskriminace;
- hodnoty vyšší než 1,70 – velmi vysoká schopnost diskriminace.

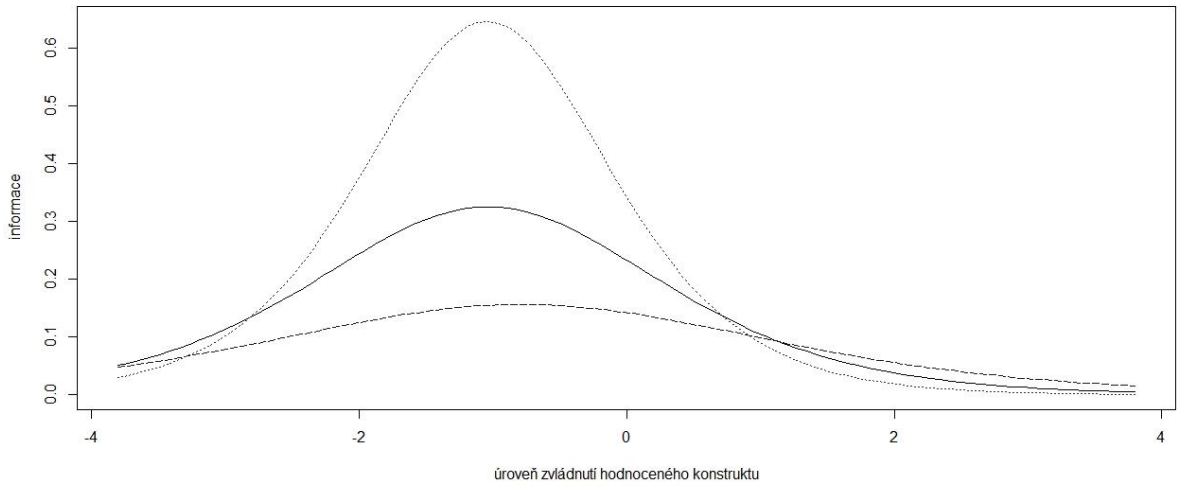
[33] Obtížnost testové položky je tedy měřena stejnou metrikou jako úroveň zvládnutí hodnoceného konstruktu Θ testovanou osobou.

Konečně Edelen a Reeve (2007) zasazují obvyklé hodnoty diskriminace testové položky do intervalu od 0,5 do 2,5, Toland (2014) pak od 0,5 do 3,0.

- *Spolehlivost testové položky a spolehlivost testu*

IRT umožňuje odhadovat nejen spolehlivost celého testu, ale také spolehlivost jednotlivých testových položek. Klíčovou roli v tomto ohledu hrají informační křivky testu (TIC) a testové položky (IIC), které vyjadřují závislost mezi úrovní zvládnutí hodnoceného konstruktu Θ a množstvím poskytnuté informace na dané úrovni Θ_i (např. Edelen a Reeve, 2007). V tomto kontextu platí, že hodnota poskytnuté informace je na různých úrovních zvládnutí hodnoceného konstruktu Θ_i odlišná, přičemž vyšší množství poskytnuté informace znamená vyšší spolehlivost testu či testové položky (např. De Champlain, 2010; DeMars, 2010). De Champlain (2010), DeMars (2010), Hambleton a Jones (1993) doplňují, že nejvyšší množství informace je získáváno na úrovni zvládnutí hodnoceného konstruktu Θ, která odpovídá obtížnosti testové položky, přičemž vyšší diskriminační schopnost testové položky rovněž zvyšuje množství poskytované informace (viz obrázek č. 3 pro příklad IIC; srovnej obrázky č. 2 a č. 3 pro dokreslení uvedených tvrzení).

Obrázek č. 3: Informační křivka testové položky (IIC)



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

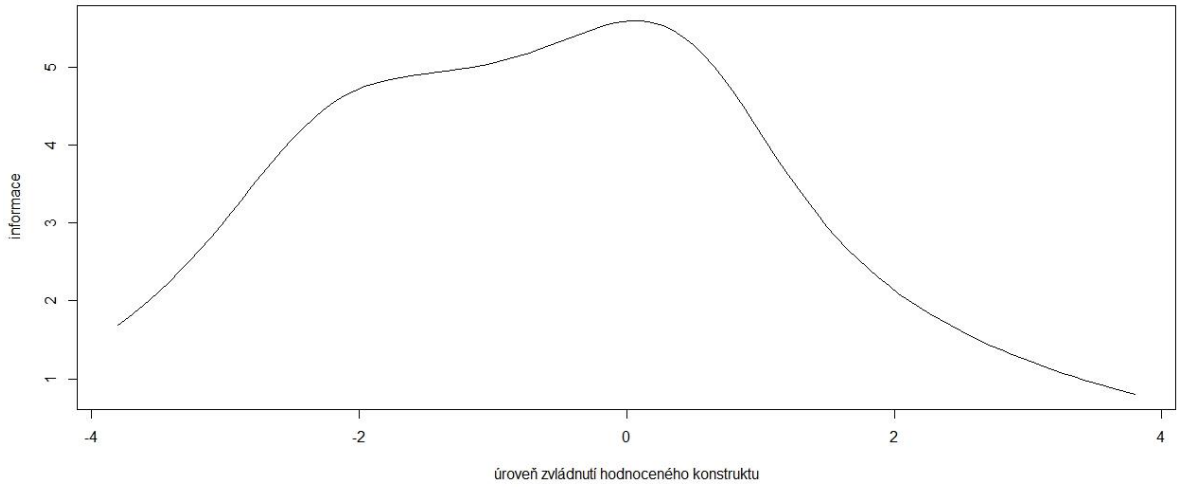
Důležitou vlastností informační funkce utvářející IIC je její aditivita. Takto informační funkce testu odpovídá součtu informačních funkcí testových položek, což lze s výhodou využít při výběru testových položek pro dosažení záměru testu (např. Hambleton a Jones, 1993). De Champlain (2010) uvádí některé možné situace tohoto typu:

- Pokud je záměrem testu stanovit pořadí testovaných osob, pak je potřeba zahrnout do testu testové položky poskytující vysoké množství informace napříč různými úrovněmi zvládnutí hodnoceného konstruktu Θ_i.

- Pokud je záměrem testu určit, zda testovaná osoba zvládla hodnocený konstrukt na určité úrovni Θ_i (*cut-off* úroveň), pak je potřeba zahrnout do testu především testové položky poskytující nejvíce informace v okolí stanovené úrovně zvládnutí hodnoceného konstruktu Θ_i.

V obou uvedených případech jde tedy o cílený výběr testových položek tak, aby konečná podoba informační funkce testu odpovídala jeho záměru (De Champlain, 2010). Obrázek č. 4 zachycuje příklad informační funkce testu, který hůře rozlišuje mezi testovanými osobami s vyšší úrovní zvládnutí hodnoceného konceptu.

Obrázek č. 4: Informační křivka testu (TIC)



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

- *DIF analýza testových položek*

Podstata DIF analýzy testových položek vycházející z IRT je analogií k DIF analýze testových položek vycházející z CTT. Hlavní rozdíl v tomto ohledu spočívá v tom, že DIF analýza v rámci IRT nevyužívá ukazatele dosaženého skóre testované osoby v testu, nýbrž ukazatele úrovně zvládnutí hodnoceného konstruktu Θ (např. Karami, 2012). Takto je například DIF analýza založená na logistické regresi primárně spojena s odhadem úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami (blíže viz podkapitola věnující se odhadům IRT modelů), načež proměnná Θ vstupuje do odhadů logistických regresních modelů (např. Lambert et al., 2018).

3.2.2 Modely vycházející z IRT

Pro analýzu testových položek (stanovení obtížnosti, diskriminace, spolehlivosti) a odhad úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami jsou využívány modely vycházející z IRT (např. Burgos, 2010). V případě dichotomických testových položek se typicky jedná o modely, které vyjadřují vztah mezi pravděpodobnosti správné odpovědi testovaných osob na testovou položku na jedné straně a úrovní jejich zvládnutí hodnoceného

konstruktu Θ na straně druhé, přičemž vztah je doplněn zvoleným počtem parametrů charakterizujících testové položky. Van der Linden (2010) označuje za určitý standard tvorby a hodnocení testů s dichotomickými testovými položkami tzv. 3PL model, který pracuje se třemi parametry a který lze vyjádřit vztahem (např. DeMars, 2010; Hambleton a Jones, 1993):

$$P(x_i = 1; \theta_j) = c_i + (1 - c_i) * \frac{e^{1,7a_i(\theta_j - b_i)}}{1 + e^{1,7a_i(\theta_j - b_i)}},$$

kde a_i je parametr diskriminace testové položky i; b_i je parametr obtížnosti testové položky i; c_i je dolní asymptota testové položky i; Θ_j je úroveň zvládnutí hodnoceného konstruktu testovanou osobou j. Doplňme, že dolní asymptota testové položky odpovídá hodnotě pravděpodobnosti, že testovaná osoba s velmi nízkou úrovní zvládnutí hodnoceného konstruktu Θ správně zodpoví danou testovou položku. DeMars (2010) v tomto kontextu hovoří o parametru „pseudo-hádání“, neboť uvedená osoba by měla být schopna správně zodpovědět testovou položku pouze hádáním. Parametr „pseudo-hádání“ může nabývat hodnoty od 0 do 1, kdy poměrně typické jsou hodnoty kolem 0,3 (např. DeMars, 2010). Za významný faktor, který ovlivňuje hodnotu parametru „pseudo-hádání“, označuje DeMars (2010) kvalitu distraktorů.[34]

Speciálními případy 3PL modelu jsou dva jednodušší modely pracující se dvěma či jedním parametrem. 2PL model odpovídá 3PL modelu, u něhož jsou hodnoty dolní asymptoty testových položek rovny nule. 2PL model tedy nepředpokládá působení parametru „pseudo-hádání“ testované osoby a odhaduje jen dva parametry: (a) parametr obtížnosti testových položek; a (b) parametr diskriminace testových položek (např. DeMars, 2010). 1PL model pak odpovídá 3PL modelu, u něhož je hodnota dolní asymptoty testových položek rovna nule a diskriminace všech testových položek je stejná. 1PL model tedy zohledňuje pouze parametr obtížnosti testových položek a svou podstatou se jedná o nejjednodušší ze tří uvedených modelů (např. DeMars, 2010). Van der Linden (2010), De Champlain (2010) však zároveň uvádějí některé praktické výhody 1PL modelu, které zahrnují:

- méně náročné požadavky na velikost výběrového souboru testovaných osob, a to s ohledem na nižší počet odhadovaných parametrů;
- existenci jednoznačného vztahu mezi úrovní zvládnutí hodnoceného konstruktu Θ (IRT) na jedné straně a dosaženým skóre v testu (CTT) na straně druhé.[35]

Uveďme, že v praxi často používaný Raschův model je ve své podstatě 1PL model, který vychází z odlišné notace a také z odlišného přístupu. DeMars (2010) uvádí, že zatímco 1PL model je odhadován tak, aby co nejlépe odpovídal datům, Raschův model usiluje o přizpůsobení dat modelu s vyloučením těch testových položek, které jsou pro daný model nevhodné.

[34] Uveďme, že hodnota 1,7 se nejen v 3PL modelu vyskytuje historicky. Účelem této hodnoty je aproximovat model tak, aby podoba ICC křivky odpovídala tvaru tzv. normální ogivy. Hodnota může být z 3PL modelu vynechána, což bude mít vliv na hodnotu parametru diskriminace a_i (např. DeMars, 2010; De Champlain, 2010).

[35] Platí tedy, že k danému skóre úspěšnosti v testu existuje jediná hodnota Θ, a také že testové položky se stejným podílem správně odpovídajících testovaných osob jsou v 1PL modelu stejně obtížné. Takové vztahy ovšem nemusí platit v případě 2PL modelu, a to s ohledem na odlišný parametr diskriminace testových položek (např. DeMars, 2010). Tyto skutečnosti mohou být pak pro testované osoby matoucí, když se mohou ptát: „Proč mám jinou hodnotu úrovně zvládnutí hodnoceného konstruktu Θ, když jsem vyřešil stejný počet testových položek?“ (např. De Champlain, 2010).

Vedle 1PL, 2PL a 3PL modelů uvádí van der Linden (2010) ještě další typy modelů vycházející z IRT. Prvním typem jsou tzv. multidimenzionální modely, které řeší problém narušení jednoho z předpokladů 1PL, 2PL a 3PL modelů, předpokladu unidimenzionality. Multidimenzionální modely tak mohou být odhadovány v případě výskytu více dimenzí, tj. hodnocených konstruktů, v testu, přičemž odhadovaná je pravděpodobnost správné odpovědi testovaných osob v závislosti na úrovni jejich zvládnutí vyššího počtu hodnocených konstruktů Θ_1 až Θ_n (např. van der Linden, 2010; Chalmers, 2012). Druhým typem pak jsou tzv. neparametrické modely, které se vyhýbají odhadu parametrů modelů a naopak kladou důraz na ordinální charakter (pořadí) dat (např. van der Linden, 2010).

3.2.3 Odhady modelů vycházejících z IRT

Odhady parametrů modelů vycházejících z IRT jsou založeny na různých metodických přístupech, které však mají některé aspekty společné (např. Rupp, 2005). První společný aspekt je spojený s odhadem dvou skupin parametrů modelů vycházejících z IRT, a to: (a) parametrů vztahujících se k testovým položkám; a (b) parametrů vztahujících se k testovaným osobám, přičemž zároveň jsou hledány parametry, které nejlépe odpovídají pozorovaným datům (např. Burgos, 2010; Rupp, 2005).

Druhý společný aspekt odhadů modelů vycházejících z IRT je vztažený k předpokladu lokální nezávislosti odpovědí testovaných osob na testové položky (např. Rupp, 2005; Burgos, 2010). Naplnění tohoto předpokladu umožňuje počítat podmíněnou pravděpodobnost pozorování daného vzoru odpovědí Y_i testované osoby i s úrovní zvládnutí hodnoceného konstruktu Θ_i na testové položky j prostřednictvím vztahu (např. Rupp, 2005; Harwell, Baker a Zwarts, 1988):

$$P(Y_i|\theta_i) = \prod_J P_j\,(Y_{ij} \mid \theta_i),$$

a pro všechny náhodně vybrané testované osoby pak prostřednictvím vztahu (např. Rupp, 2005; Harwell, Baker a Zwarts, 1988):

$$P(Y|\theta) = \prod_I \prod_J P_j(Y_{ij} \mid \theta).$$

Třetí společný aspekt odhadů modelů vycházejících z IRT je spojený s využitím konceptu věrohodnosti (*likelihood*). Na rozdíl od výše uvedeného výpočtu pravděpodobnosti pozorování daného vzoru odpovědí testovaných osob v závislosti na úrovni jejich zvládnutí hodnoceného konstruktu je věrohodnost funkcí úrovně zvládnutí hodnoceného konstruktu Θ testovanou osobou. Výpočet, který opětovně využívá předpoklad lokální nezávislosti, je analogický (např. Rupp, 2005):

$$L(\theta|Y) = \prod_I \prod_J P_j(Y_{ij} \mid \theta).$$

Zároveň však platí, že z praktického hlediska je snazší pracovat s logaritmickou podobou věrohodnosti (*log-likelihood*), která nabývá podoby (např. Rupp, 2005):

$$logL(\theta|Y) = \sum_{I}\sum_{J} \log\left[P_j\left(Y_{ij}\,\middle|\,\theta\right)\right].$$

Metody maximální věrohodnosti (*maximum likelihood*) jsou tradičním přístupem k odhadům modelů vycházejících z IRT (např. Harwell, Baker a Zwarts, 1988; Burgos, 2010; Wirth a Edwards, 2007), jejich podstatu lze charakterizovat následujícím způsobem.

Předpoklad lokální nezávislosti testových položek umožňuje stanovit pravděpodobnost pozorování daného vzoru odpovědí na testové položky (Y_{ij}) ze strany testovaných osob, jejichž úroveň zvládnutí hodnoceného konstruktu je Θ_i, vztahem (např. Harwell, Baker a Zwarts, 1988; Van der Linden, 2010; DeMars, 2010):

$$P\left(Y_{ij}\middle|\theta_i, \varepsilon\right) = \prod_{J} P_j\,(\theta_i)^{y_{ij}} \times Q_j(\theta_i)^{1-y_{ij}},$$

kde ε je matice skutečných parametrů testových položek a y_{ij} je (správná či nesprávná) odpověď testované osoby i na testovou položku j, a platí $Q_j(\Theta_i) = 1 - P_j(\Theta_i)$. Zdůrazněme, že vztah vyjadřuje podmíněnou pravděpodobnost pozorování daného vzoru odpovědí v závislosti jednak na úrovni zvládnutí hodnoceného konstruktu testovanými osobami (Θ_i) a jednak na matici skutečných parametrů testových položek.

Hodnota věrohodnosti pozorovaného vzoru odpovědí na testové položky všech testovaných osob je pak odvozena vztahem (např. Harwell, Baker a Zwarts, 1988; Burgos, 2010):

$$L = \prod_{I}\prod_{J} P_j\,(\theta_i)^{y_{ij}} \times Q_j(\theta_i)^{1-y_{ij}},$$

případně v logaritmickém tvaru jako (např. Harwell, Baker a Zwarts, 1988; Burgos, 2010):

$$log\ L = \sum_{I}\sum_{J} [y_{ij} log\ P_j(\theta_i) + \left(1 - y_{ij}\right) log Q_j(\theta_i)].$$

S využitím některého z uvedených vztahů jsou pak parametry testových položek odhadovány tak, aby byla maximalizována hodnota věrohodnosti, tj. hodnota L, případně hodnota $log\ L$[36], s cílem najít takovou podobu parametrů testových položek, které nejlépe odpovídají pozorovaným datům (např. Burgos, 2010).

Tradiční způsob nalezení maximální hodnoty věrohodnosti je založen na řešení soustavy rovnic, v nichž první derivace věrohodnosti v logaritmickém tvaru je pro všechny odhadované parametry testové položky j rovna nule, tj. například pro 3PL model platí (např. Harwell, Baker a Zwarts, 1988):

$$\frac{\partial}{\partial a_j}(\log L) = 0;\ \frac{\partial}{\partial b_j}(\log L) = 0;\ \frac{\partial}{\partial c_j}(\log L) = 0,$$

[36] Zjednodušeně platí, že pro daný vzor odpovědí na testové položky (Y_{ij}) všech testovaných osob jsou zkoušeny kombinace parametrů (Θ_i, ε) tak, aby hodnota věrohodnosti byla nejvyšší. Parametry přitom vstupují do příslušného modelu (např. 3PL model, 2PL model, Raschův model), z něhož jsou následně počítány hodnoty $P_j(\Theta_i)$, přičemž zároveň platí $Q_j(\Theta_i) = 1 - P_j(\Theta_i)$.

přičemž do těchto rovnic rovněž vstupuje vyjádření vztahů mezi parametry 3PL modelu (např. Harwell, Baker a Zwarts, 1988). Výpočet logaritmické věrohodnosti lze ovšem vyjádřit také alternativně, kdy je spojitá úroveň zvládnutí hodnoceného konstruktu Θ testovanými osobami shlukována do konečného počtu intervalů vyjadřujících k úrovní zvládnutí hodnoceného konstruktu Θ_k testovanými osobami. Logaritmická věrohodnost je pak vyjádřena vztahem (např. Harwell, Baker a Zwarts, 1988):

$$log\ L = konstanta + \sum_{K}\sum_{J}[r_{jk} log\ P_j(\theta_k) + (n_{jk} - r_{jk}) log Q_j(\theta_k)],$$

kde n_{jk} je počet testovaných osob s úrovní zvládnutí hodnoceného konstruktu v intervalu Θ_k, kteří odpovídají na testovou položku j a r_{jk} je počet těchto testovaných osob, který danou testovou položku zodpovídá správně. I v tomto případě je hledána taková kombinace odhadovaných parametrů modelu, která maximalizuje hodnotu logaritmické věrohodnosti.

Představený postup odhadu parametrů modelu se ovšem potýká s významným nedostatkem v podobě chybějící informace o úrovni zvládnutí hodnoceného konstruktu Θ testovanými osobami. Na tuto skutečnost a také na nedostatky metod sdružené a podmíněné maximální věrohodnosti (viz tabulka č. 8 pro jejich podstatu) reaguje metoda marginální maximální věrohodnosti, kterou Burgos (2010) označuje za nejčastěji používanou metodu tohoto typu.

Tabulka č. 8: Podstata metod sdružené a podmíněné maximální věrohodnosti odhadu modelů vycházejících z IRT

Metoda	Podstata metody
Metoda sdružené maximální věrohodnosti	Metoda sdružené maximální věrohodnosti je nejstarší z uváděných metod maximální věrohodnosti. Hlavním znakem této metody je společný odhad všech parametrů zvoleného modelu, tj. jak parametrů testovaných osob, tak parametrů testových položek (např. Wirth a Edwards, 2007; Harwell, Baker a Zwarts, 1988; Rupp, 2005). Protože však současný odhad všech parametrů modelu není vzhledem k jejich počtu možný, je aplikována postupná strategie: (1) odhadu parametrů testových položek s využitím vstupních hodnot parametrů testovaných osob; a (2) úpravy vstupních hodnot parametrů testovaných osob s využitím nových hodnot parametrů testových položek, s tím, že tento postup je opakován do dosažení stanoveného kritéria konvergence (např. Harwell, Baker a Zwarts, 1988; Wirth a Edwards, 2007; Rupp, 2005). Problémem postupu však je nekonzistentnost odhadu parametrů modelu z důvodu chybějící asymptotické konvergence odhadů parametrů ke skutečným hodnotám populace jak při zvyšování počtu testovaných osob, tak při zvyšování počtu testových položek (např. Harwell, Baker a Zwarts, 1988; Rupp, 2005; Wirth a Edwards, 2007).

Metoda	Podstata metody
Metoda podmíněné maximální věrohodnosti	Metoda podmíněné maximální věrohodnosti řeší problém metody sdružené maximální věrohodnosti tím, že hodnoty úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami nahrazuje alternativní proměnnou, což typicky bývá dosažené skóre testované osoby v testu (např. Burgos, 2010; Rupp, 2005). Tento krok umožňuje odhadovat pouze parametry testových položek prostřednictvím metody maximální věrohodnosti. Burgos (2010) doplňuje, že s využitím takto stanovených parametrů testových položek je následně možné odhadnout také parametry testovaných osob (Θ). Za hlavní nedostatky metody podmíněné maximální věrohodnosti Rupp (2005) především označuje: (a) náročnost požadavků kladených na kvalitu ukazatele dosaženého skóre v testu pro operacionalizaci úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami; a (b) vhodnost metody pouze pro 1PL (Raschův) model, kdy dosažené skóre testovaných osob v testu není dostatečnou statistikou pro odhad parametrů 2PL a 3PL modelů (např. také Harwell, Baker a Zwarts, 1988).

Podstata metody marginální maximální věrohodnosti (dále jen „metoda MML) je založena na následujících východiscích:

- Metoda MML předpokládá, že úroveň zvládnutí hodnoceného konstruktu testovanými osobami (Θ) je náhodná proměnná, u které je možné stanovit tzv. pravděpodobnostně hustotní funkci, a to nejčastěji v podobě normálního rozdělení[37] s průměrem nula a směrodatnou odchylkou σ^2 (např. Burgos, 2010; Harwell, Baker a Zwarts, 1988; Wirth a Edwards, 2007; Rupp, 2005; DeMars, 2010).
- Metoda MML pracuje s maximalizací marginalizované podoby logaritmické věrohodnosti, která je zbavena závislosti na úrovni zvládnutí hodnoceného konstruktu testovanými osobami (Θ) prostřednictvím integrálního počtu (např. Harwell, Baker a Zwarts, 1988; Wirth a Edwards, 2007; Rupp, 2005):

$$logL_M = \sum_I log\left\{\int_{-\infty}^{+\infty} \prod_J P_j(Y_{ij} \mid \theta) \times g(\theta) d\theta\right\},$$

  kde Y_{ij} je vzor odpovědí testovaných osob i na testové položky j, a kde $g(\Theta)$ je apriorní pravděpodobností hustotní funkce úrovně zvládnutí hodnoceného konstruktu osobami v populaci (Θ). Platí tedy, že odhad parametrů testových položek je závislý na apriorní pravděpodobnostní hustotní funkci úrovně zvládnutí hodnoceného konstruktu testovanými osobami (Θ), nikoliv však na samotné úrovni tohoto konstruktu (např. Harwell, Baker a Zwarts, 1988).
- Výpočet výše uvedeného integrálu marginalizované logaritmické věrohodnosti je tradičně nahrazen aproximací křivky pod integrálem, a to na bázi využití konečného počtu kvadraturních bodů, které jsou typicky rovnoměrně rozloženy v metrice například v intervalu od -4 do +4 (např. Rupp, 2005).[38] Pro každý takový bod k je vypočtena

[37] Harwell, Baker a Zwarts (1988), Rupp (2005) nicméně zdůrazňují, že rozdělení nemusí být nutně normální.
[38] Numerická integrace s využitím Gauss-Hermitovy kvadratury (např. Wirth a Edwards, 2007).

aproximovaná hodnota pravděpodobnostní hustotní funkce $A(\Theta_k)$, a to v podobě plochy příslušných obdélníků vymezujících korespondující bod integrální křivky. Marginalizovaná logaritmická věrohodnost následně nabývá tvaru (např. Rupp, 2005):

$$logL_M = \sum_I log\left[\sum_K \left\{\prod_J P_j(Y_{ij} \mid \theta_k)\right\} \times A(\theta_k)\right].$$

- Uveďme, že hodnota $A(\Theta_k)$ de facto odpovídá váze každého kvadraturního bodu ve vazbě na odpovídající výšku a šířku obdélníku pravděpodobnostní hustotní funkce (např. Wirth a Edwards, 2007; Harwell, Baker a Zwarts, 1988), přičemž Wirth a Edwards (2007) upozorňují, že vyšší počet kvadraturních bodů zvyšuje kvalitu odhadu, zároveň však činí výpočetní operace náročnější.
- Metoda MML využívá tzv. Bayesův teorém, který tvrdí, že apriorní pravděpodobnostně hustotní funkci úrovně zvládnutí hodnoceného konstruktu osobami v populaci $g(\Theta)$ lze upravit s využitím známých dat z odpovědí testovaných osob (Y_i) do podoby tzv. aposteriorní pravděpodobnostně hustotní funkce úrovně zvládnutí hodnoceného konstruktu osobami v populaci $P(\Theta_i \mid Yi)$ (např. Rupp, 2005). Následně pro $P(\Theta_i \mid Y_i)$ platí (např. Harwell, Baker a Zwarts, 1988):[39]

$$P(\theta_i \mid Y_i) = \frac{P(Y_i \mid \theta_i)g(\theta)}{\int P(Y_i \mid \theta_i)g(\theta)d\theta},$$

respektive v bodovém vyjádření pro všechny kvadraturní body k (např. Rupp, 2005):

$$P_{ik}(\theta_k \mid Y_i) = \frac{\prod_J\{P_j(Y_{ij} \mid \theta_k)\} \times A(\theta_K)}{\sum_K\{\prod_J P_j(Y_{ij} \mid \theta_k)\} \times A(\theta_k)}.$$

Vlastní odhad parametrů zvoleného modelu může využívat různých metodických přístupů, přičemž tradičním je v tomto ohledu tzv. EM algoritmus (např. Wirth a Edwards, 2007; Harwell, Baker a Zwarts, 1988).[40] Rupp (2005), Wirth a Edwards (2007), Harwell, Baker a Zwarts (1988) charakterizují tři základní kroky EM algoritmu následujícím způsobem:

- V prvním kroku je vypočtena aposteriorní hodnota $P_{ik}(\Theta_i \mid Y_i)$ pro každý vzor odpovědí testovaných osob Y_i a pro každý kvadraturní bod k podle výše uvedeného vztahu, a to s využitím „provizorních“ odhadů parametrů testových položek tak, jak byly vypočteny v předchozím kroku iterace. Tímto způsobem získáváme informaci o pravděpodobnostech, že testovaná osoba i, která na testové položky odpovídá daným způsobem, zvládá hodnocený konstrukt na úrovni odpovídající jednotlivým kvadraturním bodům k.
- Ve druhém kroku jsou pro všechny testové položky j a pro všechny kvadraturní body k vypočítána „umělá data“ $\dot{n}_{jk}$ a $\dot{r}_{jk}$, kde: (a) $\dot{n}_{jk}$ je očekávaný počet testovaných osob v daném kvadraturním bodě k a pro danou testovou položku j; a (b) $\dot{r}_{jk}$ je očekávaný počet správně odpovídajících testovaných osob v daném kvadraturním bodě k a pro danou testovou položku j. Pro tento účel jsou využity aposteriorní hodnoty $P_{ik}(\Theta_k \mid Y_i)$, které byly pro

[39] Aposteriorní pravděpodobnostně hustotní funkce úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami v populaci tedy vyjadřuje pravděpodobnost pozorování daného vzoru odpovědí pro různé úrovně zvládnutí hodnoceného konstruktu testovanými osobami Θ.

[40] Název EM algoritmus je kombinací dvou kroků: (1) kroku očekávání (***E**xpected*); a (2) kroku **M**aximalizace.

všechny vzory odpovědí testovaných osob Y_i a pro všechny kvadraturní body k vypočteny v prvním kroku postupu, přičemž platí:

$$\bar{n}_{jk} = \sum_{I} P_{ik}(\theta_k \mid Y_i),$$

$$\bar{r}_{jk} = \sum_{I} Y_{ij} P_{ik}(\theta_k \mid Y_i).$$

Očekávané hodnoty jsou tedy součtem aposteriorních pravděpodobností, že testovaná osoba s daným vzorem odpovědí má úroveň zvládnutí hodnoceného konstruktu Θ_k, přičemž tento součet je proveden přes všechny odpovídající testované osoby. V rámci druhého vztahu je navíc zohledněno, zda odpověď testované osoby je správná či nikoliv.

- Ve třetím kroku jsou odhadovány parametry testových položek, opětovně na bázi maximalizace hodnoty v tomto případě marginalizované logaritmické věrohodnosti. Protože však pro výpočet $P_j\,(Y_{ij} \mid \Theta_k)$ nejsou, na rozdíl od v úvodu představené výchozí situace, známé skutečné počty testovaných osob s úrovní zvládnutí hodnoceného konstruktu Θ patřící do skupin spojených s jednotlivými kvadraturními body k, jsou alternativně využita „umělá data" $\dot{n}_{jk}$ a $\dot{r}_{jk}$ z druhého kroku postupu. Z praktického hlediska je i v tomto případě založen odhad parametrů testových položek na řešení soustavy rovnic s prvními derivacemi marginalizované logaritmické věrohodnosti všech parametrů rovnými nule, tj. například pro 3PL model:

$$\frac{\partial}{\partial a_j}(\log L_M) = 0;\ \frac{\partial}{\partial b_j}(\log L_M) = 0;\ \frac{\partial}{\partial c_j}(\log L_M) = 0.$$

Uvedený postup je opakován s nově vypočtenými parametry testových položek, a to tak dlouho, dokud není dosaženo požadované úrovně konvergence (např. Rupp, 2005; DeMars, 2010; Wirth a Edwards, 2007).

Metoda MML/EM je primárně spojena s odhadem parametrů testových položek a aposteriorní pravděpodobnostně hustotní funkce úrovně zvládnutí hodnoceného konstruktu Θ osobami v populaci (např. DeMars, 2010; Rupp, 2005), nikoliv tedy s odhadem vlastní úrovně zvládnutí hodnoceného konstruktu Θ testovanými osoby. V tomto ohledu Rupp (2005) uvádí, že úroveň zvládnutí hodnoceného konstruktu testovanými osoby Θ je možné stanovit s využitím odhadů hodnot parametrů testových položek, a to prostřednictvím některého ze tří metodických přístupů: (a) přístupu na bázi maximální věrohodnosti (ML); (b) přístupu na bázi Bayesovského modálního odhadu (MAP); a (c) přístupu na bázi odhadu založeného na střední hodnotě aposteriorního rozdělení (EAP). Tabulka č. 9 představuje podstatu každého z těchto tří metodických přístupů.

Uveďme, že výhody MAP a EAP přístupů oproti ML přístupu spatřuje Thompson (2009) ve skutečnosti, že ML přístup není schopen odhadu úrovně zvládnutí hodnoceného konstruktu Θ těch testovaných osob, které všechny testové položky zodpověděly správně, nebo špatně (viz rovněž Wang, Ma a Chen, 2010). V takovém případě nelze nalézt maximum věrohodnosti, neboť hodnota věrohodnosti konverguje k nekonečnu. Vynásobení věrohodnosti apriorní pravděpodobnostně hustotní funkcí úrovně zvládnutí hodnoceného konstruktu Θ osobami v populaci tento problém řeší. EAP přístup je navíc schopen lépe zohlednit asymetrické aposteriorní rozdělení (např. Thompson, 2009), než je tomu v případě MAP přístupu, který se

zaměřuje pouze na nalezení bodového maxima, což může vést k příliš vysokým odhadům úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami (např. Wang, Ma a Chen, 2010).

Tabulka č. 9: Metodické přístupy ke stanovení úrovně zvládnutí hodnoceného konstruktu Θ testovanými osoby metodou MML/EM

Metoda	Charakteristika metody
ML	ML přístup pro odhad úrovně zvládnutí hodnoceného konstruktu Θ testovanou osobou je založen na hledání maxima věrohodnosti v závislosti na vzoru odpovědí Y_{ij} testované osoby i na všechny testové položky j. Výpočet věrohodnosti přitom sleduje tradiční vztah vynásobení pravděpodobností dosažení správných a nesprávných odpovědí testované osoby v závislosti na úrovni jejího zvládnutí hodnoceného konstruktu Θ_i, přičemž hodnota Θi není známa a naopak známy jsou hodnoty parametrů testových položek. V případě ML přístupu je záměrem nalézt takovou hodnotu Θ_i, která vede k maximalizaci hodnoty věrohodnosti. Thompson (2009) doplňuje, že maximum hodnoty věrohodnosti lze hledat různými způsoby, mezi které patří: (a) odhad hodnoty věrohodnosti pro všechny možné hodnoty Θ_i „brutální silou" s nalezením Θ_i s nejvyšší hodnotou věrohodnosti; a (b) Newton-Raphsonova iterativní metoda založená na zpřesňování hodnoty maxima věrohodnosti prostřednictvím diferenciálního počtu. S ohledem na možný problém výskytu lokálních maxim v případě druhé z uvedených metod doporučuje Thompson (2009) využití více přístupů se srovnáním odhadů.
MAP	MAP přístup je jedna z variant Bayesova přístupu, která primárně násobí věrohodnost apriorní pravděpodobnostně hustotní funkcí úrovně zvládnutí hodnoceného konstruktu osobami v populaci $g(\Theta)$. Maximální hodnota pro úroveň Θ_i je pak hledána z takto konstruovaného aposteriorního rozdělení.
EAP	EAP přístup je druhá z variant Bayesova přístupu, která sleduje stejný postup, jako tomu je v případě MAP přístupu, nicméně u aposteriorního rozdělení nehledá maximální hodnotu, nýbrž počítá střední, tj. očekávanou, hodnotu na bázi různých vah úrovní Θ, které jsou stanoveny prostřednictvím vypočteného aposteriorního rozdělení.

Ale také metoda MML/EM se potýká s některými nedostatky, které především zahrnují: (a) potřebu správného stanovení apriorního rozdělení úrovně zvládnutí hodnoceného konstruktu Θ osobami v populaci (např. Harwell, Baker a Zwarts, 1988) a opomíjení informace o apriorních informacích týkajících se parametrů (např. Burgos, 2010); (b) problémy aplikace metody MML/EM pro odhady modelů měřících vysoký počet dimenzí (např. Burgos, 2010; Wirth a Edwards, 2007); a (c) nutnost řešit otázky spojené s integrací úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami na bázi vymezení kvadraturních bodů (např. Wirth a Edwards, 2007). Tyto nedostatky metody MML/EM stojí v pozadí hledání dalších způsobů odhadu parametrů modelů vycházejících z IRT.

Jednou z moderních alternativ k metodě MML/EM je metoda Monte Carlo s Markovovými řetězci (dále i metoda MCMC), jejíž výhoda spočívá především ve vyhnutí se potřebě práce

s kvadraturními body (např. Wirth a Edwards, 2007; Wang, Ma a Chen, 2010). Východiskem odhadů parametrů modelů vycházejících z IRT pomocí metod MCMC je předpoklad, že všechny odhadované parametry sledují určité apriorní rozdělení hodnot (např. Burgos, 2010). Následně je využita myšlenka Bayesova teorému, která tvrdí, že apriorní rozdělení hodnot parametrů může být zlepšeno s využitím informace obsažené ve vzoru odpovědí testovaných osob. Aposteriorní rozdělení hodnot parametrů pak lze vyjádřit vztahem (např. Burgos, 2010):

$$P(A|B) = \frac{P(B \mid A) \times P(A)}{P(B)},$$

kde *A* je apriorní, typicky normální, rozdělení hodnot parametrů, kde *B* je vzor odpovědí testovaných osob na testové položky, a kde $A \mid B$ je aposteriorní rozdělení hodnot parametrů. Výhodou tohoto postupu je ta skutečnost, že vzor odpovědí testovaných osob pomáhá snižovat nejistotu týkající se podoby apriorního rozdělení (např. Burgos, 2010).

Vlastní MCMC metoda je následně založena na utváření tzv. Markovových řetězců, které představují spojení mezi apriorním a aposteriorním rozdělením hodnot parametrů modelu a které nabývají obecné podoby (např. Burgos, 2010):

$$f(\theta, \beta, \sigma | Y_{ij}) \sim \sum_I \sum_J P_{ij}{}^{y_{ij}} Q_{ij}{}^{1-y_{ij}} \times N(0, \sigma^2) \times f(\beta) \times f(\sigma^2),$$

kde $N(0, \sigma^2)$ je normální rozdělení hodnot úrovně zvládnutí hodnoceného konstruktu testovanými osobami (Θ) s průměrem nula a směrodatnou odchylkou σ^2, kde $f(\beta)$ je apriorní rozdělení hodnot parametrů testových položek, kde $f(\sigma^2)$ je apriorní rozdělení hodnot směrodatné odchylky σ^2, a kde Y_{ij} je odpověď testované osoby *i* na testovou položku *j*. Markovův řetězec pak vychází z myšlenky, že každá událost v bodě $t+1$ je odvozována pouze z události v předchozím bodě *t* a nezávisí na událostech předchozích (např. Wirth a Edwards, 2007). Pokud po úvodní „zahřívací" fázi dochází ke konvergenci hodnot, dostáváme po dosažení konvergence výběrový soubor hodnot parametrů z jejich aposteriorního rozdělení a právě z tohoto výběrového souboru jsou počítány jejich odhady na bázi základních statistik (např. průměr, směrodatné odchylky). Wirth a Edwards (2007), Burgos (2010) takto tvrdí, že konvergence je posun hodnot parametrů z jejich apriorního rozdělení do jejich aposteriorního rozdělení, přičemž pro výpočet statistik nejsou zajímavé hodnoty získané před dosažením konvergence. Tyto jsou z odhadů parametrů modelu vynechány (např. Wirth a Edwards, 2007). Uveďme, že existuje řada technik pro utváření výběrových souborů pro odhady parametrů modelů, přičemž za často využívané označuje Burgos (2010) tzv. Metropolisův-Hastingsův algoritmus, respektive Gibbsův výběrový plán.

Metoda MCMC má vedle svých výhod oproti metodě MML/EM také své nevýhody. Tou hlavní je především potřeba rozhodnutí, zda a kdy dochází k žádoucí konvergenci k aposteriornímu rozdělení pro vytvoření výběrového souboru hodnot parametrů (např. Wirth a Edwards, 2007; Burgos, 2010). Návodné v tomto ohledu může být: (a) pozorování vývoje hodnot parametru graficky s postupující iterací s tím, že žádoucí je absence trendu v datech a existence homogenního rozpětí hodnot; a (b) statistika $\hat{R}$, u které je žádoucí hodnota blížící se jedné a alespoň menší než 1,1 (např. Burgos, 2010).

3.2.4 Požadavky na modely vycházející z IRT

Na hodnocení testů prostřednictvím modelů vycházejících z IRT jsou ve srovnání s CTT kladeny vyšší požadavky, přičemž van der Linden (2010) tvrdí, že teprve kvalitní model, tj. model splňující požadavky na něj kladené, je možné využít jako model pro měření a stanovení úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami.

- ***Požadavky kladené na data***

V odborné literatuře nepanuje obecná shoda týkající se velikosti výběrového souboru testovaných osob pro odhad modelů vycházejících z IRT (např. Thorpe a Favia, 2012; Edelen a Reeve, 2007). Návodné informace v tomto ohledu uvádějí:

- minimální počet 100 až 200 testovaných osob pro odhady testů dichotomických testových položek prostřednictvím 1PL (Raschova) modelu (např. DeMars, 2010; Thorpe a Favia, 2012; Edelen a Reeve, 2007);
- minimální počet 500 testovaných osob pro odhady testů dichotomických testových položek prostřednictvím 2PL modelu (např. De Champlain, 2010; DeMars, 2010; Edelen a Reeve, 2007), přičemž některé zdroje uvádějí také minimální hodnotu 200 testovaných osob (např. Thorpe a Favia, 2012; Edelen a Reeve, 2007);
- minimální počet až 1000 testovaných osob pro odhady testů dichotomických testových položek prostřednictvím 3PL modelu (např. DeMars, 2010).

Z obecnějšího hlediska platí, že: (a) testy s vyšším počtem testových položek a vyšším počtem testovaných osob typicky vedou k lepšímu odhadu parametrů testových položek; (b) více komplexní modely kladou vyšší nároky na počet testovaných osob; a (c) vyžadovaný počet testovaných osob je nižší v případě vyšší kvality testových položek (např. normalita rozdělení hodnot úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami; dobrá diskriminační schopnost testových položek) a reprezentativnosti výběrového souboru testovaných osob vzhledem k populaci (Edelen a Reeve, 2007; Thorpe a Favia, 2012; DeMars, 2010).[41]

[41] DeMars (2010) uvádí, že modely vycházející z IRT nepředpokládají normální rozdělení úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami, ani parametrů testových položek. Zároveň však normalita rozdělení může zvyšovat kvalitu odhadů (např. DeMars, 2010; Toland, 2014; Finch a Habing, 2007). DeMars (2010) dokládá uvedené úvahy na některých příkladech. Pro odhad 2PL modelu s normálním rozdělením hodnot úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami a s vysoce diskriminujícími testovými položkami lze za dostatečný považovat počet 500 testovaných osob a 20 testových položek. V případě 3PL modelů bez normálního rozdělení hodnot úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami a s málo diskriminujícími testovými položkami lze za dostatečný považovat počet 1000 testovaných osob a 40 testových položek (např. DeMars, 2010). DeMars (2010) zároveň uvádí, že zvýšení počtu testovaných osob i počtu testových položek snižuje hodnotu směrodatné chyby odhadů, přičemž taková redukce již může být marginální v případě 2 až 3 tisíc testovaných osob a v případě 50 až 80 testových položek. Konečně platí, že u menších výběrových souborů testovaných osob by měl být obecně preferován odhad jednodušších modelů vycházejících z IRT (např. DeMars, 2010).

- ***Požadavky unidimenzionality a lokální nezávislosti testových položek***

Tradiční modely vycházející z IRT (např. 1PL, 2PL a 3PL model) jsou spojeny s konceptem unidimenzionality, který vyjadřuje požadavek na to, aby test měřil právě jen jeden hodnocený konstrukt (např. Hambleton a Jones, 1993; Edelen a Reeve, 2007; Toland, 2014; Ziegler a Hagemann, 2015; DeMars, 2010). Narušení předpokladu unidimenzionality může vzniknout z různých příčin (např. DeMars, 2010; Ziegler a Hagemann, 2015; Zhang, 2008; Edelen a Reeve, 2007):

- Testové položky uchopují různé konstrukty (např. uvedení matematického testu delším vstupním textem s možným doplněním konstruktu matematické gramotnosti konstruktem čtenářské gramotnosti).
- Testované osoby odpovídají na daný konstrukt odlišně vlivem působících psychologických procesů (např. různá motivace testovaných osob k řešení testu, tj. konstrukt motivace; konstrukt utvářený nedostatkem času pro řešení závěrečných baterií testových položek; konstrukt utvářený neetickým jednáním testovaných osob).

Další konstrukt může být utvářen také faktorem přítomnosti testovaných osob ke skupině (např. pohlaví, socioekonomický původ), což dále opodstatňuje zájem o DIF analýzu testových položek (např. Edelen a Reeve, 2007).

K požadavku unidimenzionality má úzkou vazbu také další požadavek kladený na modely vycházející z IRT, požadavek lokální nezávislosti testových položek, podle něhož nemá mít pravděpodobnost správné odpovědi testovaných osob na jednu testovou položku vztah k pravděpodobnosti správné odpovědi těchto osob na testovou položku jinou (např. Nandakumar, 1994). Platí tedy, že všechny testové položky by měly být při kontrole vlivu hlavního konstruktu testu vzhledem k sobě nezávislé (např. De Champlain, 2010; Zhang, 2013; Toland, 2014) a kromě hlavního konstruktu testu by v něm neměly být přítomné další systematické kovariance testových položek (např. Edelen a Reeve, 2007; DeMars, 2010). Edelen a Reeve (2007), DeMars (2010) dále uvádějí tři typické zdroje narušení předpokladu lokální nezávislosti testových položek:

- testové položky jsou podobného původu a testované osoby s nimi mají zkušenost z dřívějška;
- testové položky jsou charakteristické podobným obsahem nebo jsou odvozovány ze stejného uvozujícího textu;
- testové položky na sebe v testu navazují a odpovědi jsou logicky propojené (např. možnost odvození odpovědi na testovou položku z tesové položky předchozí).

Edelen a Reeve (2007) doplňují, že přítomnost lokální závislosti v testových položkách utváří inflační tendence jejich parametru diskriminace s negativním dopadem na kvalitu odhadovaného modelu.[42]

[42] Edelen a Reeve (2007) uvádějí, že parametr diskriminace testové položky může při porušení požadavku lokální nezávislosti testových položek nabývat hodnot vyšších než 4. Za možnost řešení takového problému označují DeMars (2010), Thorpe a Favia (2012): (a) vynechání jedné testové položky; a (b) spojení testových položek do jedné polytomické testové položky. Potřebu vynechání testové položky je možné ověřit srovnáním kvality dvou alternativních modelů – s podezřelou testovou položkou a bez ní (např. Toland, 2014).

Zhang (2008) upozorňuje na skutečnost, že tradiční modely vycházející z IRT jsou citlivé k přítomnosti více dimenzí (konstruktů) v testu, což může vést k nepřesnostem v řadě odhadů (např. také Bonifay et al., 2015; Finch a Habing, 2007). Zároveň však existuje motivace k preferenci odhadu jednodušších a lépe srozumitelných unidimenzionálních modelů před komplikovanějšími modely multidimenzionálními (např. Zhang, 2008). Bonifay et al. (2015) doplňují tuto úvahu o tvrzení, že naplnění předpokladu unidimenzionality testu je tradičně problém, neboť komplexní konstrukty, které testy typicky měří, jsou velmi nepravděpodobně striktně unidimenzionální. Podobně Nandakumar (1994) hovoří o typické situaci, kdy vedle jedné dominantní dimenze (konstruktu) obsahuje test také vedlejší dimenze (konstrukty), které jsou spojeny jen s omezeným počtem testových položek. Následně vzniká otázka, zda preferovat unidimenzionální nebo multidimenzionální modely, přičemž relevance této otázky je spojena se zjištěními, že i v případě výskytu vedlejších dimenzí (konstruktů) v testu, může být tento efektivně odhadován jednodušími, tzv. esenciálně unidimenzionálními modely (např. Bonifay et al., 2015; Nandakumar, 1994).[43]

Pro zodpovězení otázky, zda vliv vedlejších dimenzí (konstruktů) v testu opomenout a tento odhadovat unidimenzionálním modelem, jsou proto hledány vhodné metody identifikace přítomnosti a síly multidimenzionality v testu (např. Bonifay et al., 2015; Finch a Habing, 2007; Zhang, 2008). Tabulka č. 10 představuje vybrané metodické přístupy využívané za tímto účelem. Z praktického hlediska může mít hodnocení dimenzí obsažených v testu význam také pro ověření souladu mezi expertně stanovenými konstrukty testu a konstrukty identifikovanými ze skutečných odpovědí testovaných osob. Obecně se tedy jedná o zájem posoudit kvalitu expertně stanoveného obsahu testu (např. De Champlain, 2015; Zhang, 2013).

Tabulka č. 10: Metodické přístupy pro hodnocení unidimenzionality testů

Metodický přístup	Charakteristika metodického přístupu
Explorační faktorová analýza	Pro identifikaci počtu faktorů při ex-ante neznalosti konstruktů obsažených v datech doporučují Toland (2014), Edelen a Reeve (2007), Zhang (2013) využít metody explorační faktorové analýzy. Dílčí metody v tomto ohledu zahrnují: (a) posouzení hodnot vlastních čísel (eigenvalues), například Kaiserovým pravidlem; (b) metodu paralelní analýzy; nebo (c) MAP test (např. Ziegler a Hagemann, 2015). V případě testových položek dichotomického charakteru jsou využívány tetrachorické korelace (např. DeMars, 2010; Zhang, 2013). Ziegler a Hagemann (2015) dále doporučují preferovat metodu hlavních os a hodnotit výsledky různých metod rotace faktorů. Pro hodnocení kvality jedno-faktorového řešení lze dále využít indexy souladu skutečných a odhadovaných dat: (a) komparativní index shody (CFI) s preferencí hodnot vyšších než 0,95 (0,90); (b) Tuckerův-Lewisův index (TLI) s preferencí hodnot vyšších než 0,90; a (c) střední kvadratickou chybu aproximace (RMSEA) s preferencí hodnot menších než 0,06 (viz Toland, 2014; De Champlain, 2015).

[43] Uveďme, že esenciální unidimenzionalita se zaměřuje jen na dominantní konstrukt testu, a proto je chápána jako slabší forma lokální nezávislosti testových položek (např. Nandakumar, 1994).

Metodický přístup	Charakteristika metodického přístupu
Stoutův test esenciální unidimenzionality (DIMTEST)	Podle základní myšlenky Stoutova testu esenciální unidimenzionality nastává unidimenzionalita testu tehdy, pokud při kontrole vlivu hlavní dimenze (konstruktu) testu zůstane zachován pouze vliv nekorelovaného chybového efektu. Platí tedy, že při testování osob se stejnou úrovní zvládnutí hodnoceného konstruktu Θ se kovariance dvojic testových položek blíží hodnotě nula (např. DeMars, 2010).[44] Vlastní Stoutův test esenciální unidimenzionality je založen na následujících principech a krocích postupu (např. DeMars, 2010; Finch a Habing, 2007): • Nulová hypotéza předpokládá unidimenzionalitu testu, tj. předpokládá, že průměrná absolutní hodnota kovariancí dvojic testových položek se při kontrole vlivu úrovně hodnoceného konstruktu Θ blíží nule (tzv. podmíněná kovariance testových položek). • Zadaný test je rozdělen na dva dílčí testy, které se od sebe vzhledem k hodnocenému konstruktu co nejvíce odlišují, přičemž jeden z dílčích testů měří hlavní dimenzi (konstrukt) celého testu a druhý z dílčích testů potenciální vedlejší dimenzi (konstrukt) testu. Výběr testových položek pro oba dílčí testy může být založen jak teoreticky (expertní posouzení), tak empiricky s využitím faktorové analýzy s tetrachorickými korelacemi na vzorku testovaných osob[45], případně s využitím dvoustupňového postupu tvořeného hierarchickou klastrovou analýzou na bázi podmíněné kovariance (HCA/CCPROX) pro identifikaci vhodných klastrů testových položek s následným výpočtem DETECT statistik pro výběr nejvhodnějšího dělení (např. Finch a Habing, 2007). Uveďme, že podstata DETECT statistiky je blíže charakterizována na jiném místě této tabulky. • Dílčí test, který měří hlavní dimenzi celého testu (dělící test), je využit pro rozdělení testovaných osob do skupin podle jejich úrovně zvládnutí hodnoceného konstruktu Θ (např. také Zhang, 2008). Dílčí test, který měří potenciální vedlejší dimenzi celého testu (hodnotící test), je využit pro hodnocení toho, zda se průměrná absolutní hodnota kovariancí všech dvojic testových položek uvnitř skupin testovaných osob stejné úrovně zvládnutí hodnoceného konstruktu Θ, tj. při kontrole tohoto konstruktu, blíží nule. Druhý dílčí test je tedy využit pro hodnocení nulové hypotézy testu (např. také Zhang, 2008). Na základě uvedeného postupu je konečně vypočtena DIMTEST statistika testu, která je porovnána s referenčními modely, které mohou být generovány metodou bootstrap.

[44] Předpoklad nulové kovariance je spojen s očekáváním podobných odpovědí na testové položky v případě skupiny testovaných osob, jejichž úroveň zvládnutí hodnoceného konstruktu Θ testu je stejná. Tato skutečnost po kontrole vlivu hodnoceného konstruktu Θ vede k nulové kovarianci testových položek.

[45] DeMars (2010), Finch a Habing (2007) doporučují empiricky vybrat testové položky prvního dílčího testu na vzorku 30 % (popřípadě 50 %) náhodně vybraných testovaných osob a následný vlastní test realizovat pro vzorek zbývajících 70 % (popřípadě 50 %) testovaných osob.

Metodický přístup	Charakteristika metodického přístupu
NOHARM model	NOHARM model představuje nelineární faktorový model, jehož odhad je postaven na hodnocení vztahů mezi dvojicemi dichotomických testových položek, nikoliv tedy všech testových položek najednou, jak tomu bylo v případě Stoutova testu esenciální unidimenzionality (např. Finch a Habing, 2007). Základem NOHARM modelu je matice reziduálních korelací, která má: (a) na hlavní diagonále hodnoty odpovídající vztahům skutečného podílu správných odpovědí na testovou položku a podle modelu očekávaného podílu správných odpovědí na testovou položku; a (b) mimo hlavní diagonálu hodnoty odpovídající vztahům skutečného podílu správných odpovědí na obě testové položky zároveň a očekávaného podílu správných odpovědí na obě testové položky podle modelu. V návaznosti na takto konstruovanou matici reziduálních korelací byla následně navržena řada statistik založených na reziduálních korelacích, mezi které patří také Gessaroli – De Champlain χ^2 statistika (např. DeMars, 2010). Ta je využívána pro testování nulové hypotézy NOHARM modelu, že hodnoty elementů matice reziduálních korelací mimo hlavní diagonálu jsou rovny nule a tedy že model je jedno-faktorový. Pokud není nulová hypotéza zamítnuta, pak je jedno-faktorový model adekvátní aproximací pozorovaných korelací mezi testovými položkami a platí předpoklad unidimenzionality testu.
DETECT index	DETECT index jako stále oblíbenější metrika multidimenzionality testu vychází z myšlenky klesající vhodnosti využití unidimenzionálních modelů při rostoucí úrovni multidimenzionality testu (např. Bonifay et al., 2015). DETECT index předpokládá, že testová položka měří (Zhang, 2013): • testový kompozit, tj. lineární kombinaci hlavního hodnoceného konstruktu, • reziduální část, která je ortogonální (nekorelovaná) k testovému kompozitu. DETECT index dále vychází z tzv. jednoduché dimenzionální struktury testu, která odpovídá rozdělení testu na určitý počet klastrů, kdy každý klastr obsahuje testové položky, které jsou vzájemně dimenzionálně homogenní a zároveň dimenzionálně heterogenní k testovým položkám dalších klastrů. DETECT index je tedy spojen s hledáním optimální podoby dimenzionálně založeného rozdělení testových položek do klastrů, kdy každý klastr odpovídá jiné dimenzi testu, ať již dimenze má daný obsahový význam či nikoliv (např. Zhang, 2013). Platí, že optimální rozdělení testových položek maximalizuje hodnotu DETECT indexu[46] (např. Zhang, 2013; Bonifay et al., 2015) a právě tato hodnota umožňuje hodnotit multidimenzionalitu testu, kdy: (a) hodnoty nižší než 0,1 (příp. 0,2) indikují unidimenzionalitu testu; (b) hodnoty v intervalu 0,1 až 0,5 (příp. 0,2 až 0,4) indikují slabou multidimenzionalitu testu; (c) hodnoty v intervalu 0,5 až 1,0 (příp. 0,4 až 1,0) indikují středně silnou multidimenzionalitu testu; a (d) hodnoty vyšší než 1,0 indikují silnou multidimenzionalitu testu.

[46] Optimální rozdělení testových položek usiluje o maximalizaci podmíněné kovariance testových položek stejného klastru a o hodnoty podmíněné kovariance blížící se nule pro testové položky odlišných klastrů. Tyto požadavky naplňuje právě optimální rozdělení testových položek homogenních klastrů testových položek, které jsou nezávislé (nekorelované) s testovými položkami dalších klastrů.

Metodický přístup	Charakteristika metodického přístupu
Q_3 test pro lokální nezávislost testových položek	Podstata Q_3 testu pro lokální nezávislost testových položek je primárně založena na odhadu parametrů unidimenzionálního modelu vycházejícího z IRT. Následně je pro každou odpověď testované osoby vypočtena hodnota reziduí jako rozdíl mezi predikovanou a skutečnou odpovědí na danou testovou položku v závislosti na úrovni zvládnutí hodnoceného konstruktu testu Θ. Statistika Q_3 testu je následně počítána v podobě korelace mezi rezidui každé dvojice testových položek, přičemž v takto konstruované matici reziduálních korelací jsou předmětem zájmu jednak statisticky významné korelace a jednak korelace vyšší než hodnota 0,20 (např. DeMars, 2010).

Zhang (2008) se detailně zabýval otázkou, za jakých okolností jsou unidimenzionální modely odhadovány s akceptovatelnou mírou nepřesností, i když test vykazuje multidimenzionální charakter. Zhang (2008) v tomto ohledu uvádí, že unidimenzionální modely jsou vhodné především v případech, kdy: (a) jen malý počet testových položek měří sekundární dimenzi (konstrukt) testu; (b) existuje vysoká úroveň korelace hlavní a sekundární dimenze (konstruktu) testu; a (c) středně velký podíl testových položek měří sekundární dimenzi (konstrukt) testu, a to se středně vysokou úrovní korelace s hlavní dimenzí.

Konečně doplňme, že v případě detekce problémů s předpokladem unidimenzionality či lokální nezávislosti testových položek existuje několik možností řešení, a to především: (a) vynechání problémových testových položek; (b) rozdělení testových položek tak, aby utvářely dvě škály; (c) odhad multidimenzionálních modelů vycházejících z IRT; a (d) převedení dichotomických testových položek narušujících požadavek lokální nezávislosti na polytomické testové položky (např. DeMars, 2010; Toland, 2014; Zhang, 2008; Naumenko, 2014).

- ***Soulad modelu a dat***

Hodnocení souladu modelových a empirických dat se zajímá o to, zda je odhadovaný model vycházející z IRT vhodně specifikován (např. DeMars, 2010). Přirozenou motivací k takovému hodnocení je zájem o dosažení co nejvyššího souladu mezi empirickými daty na jedné straně a daty generovanými modelem na straně druhé (např. Orlando a Thissen, 2000; DeMars, 2010; Stone a Zhang, 2003; Stone, 2000; Chalmers a Ng, 2017), neboť takový soulad podporuje validitu modelu a předchází hrozbám formulace nepřesných závěrů (např. Reise, 1990; Maydeu-Olivares, 2015; Maydeu-Olivares a García-Forero, 2010). Toland (2014), Stone a Zhang (2003), DeMars (2010), Stone (2000), Chalmers a Ng (2017) přitom uvádějí, že pro hodnocení souladu modelových a empirických dat byla odvozena řada statistik na úrovni testu (modelu), testové položky a testované osoby. Doplňme, že s ohledem na komplexnost řešené problematiky není důvodné předpokládat plný soulad mezi empirickými a modelovými daty (např. Toland, 2014), proto Maydeu-Olivares (2015) hovoří o výhodách přístupu, který nehodnotí úplný soulad empirických a modelových dat, nýbrž pouze soulad přibližný.

(A) Úroveň testové položky

Základní informace pro hodnocení modelových a empirických dat na úrovni testových položek poskytují ICC křivky, přičemž žádoucí je situace, kdy empirická data leží v blízkosti predikovaných modelových dat a zároveň je naplněn předpoklad lepších predikovaných výsledků testovaných osob s vyšší úrovní zvládnutí hodnoceného konstruktu Θ (např. DeMars, 2010). Vedle vizuálního posouzení (např. srovnání skutečných a modelem predikovaných odpovědí testovaných osob) je zde možnost využití řady indexů a statistik dobré shody na úrovni testové položky (např. Orlando a Thissen, 2000; DeMars, 2010; Toland, 2014; Reise, 1990).[47]

Orlando a Thissen (2000), Stone a Zhang (2003), Stone (2000), Reise (1990), Chalmers a Ng (2017) charakterizují typický postup hodnocení dobré shody testových položek 2PL a 3PL modelů:

(a) Primárně je odhadován model vycházející z IRT, tj. parametry testových položek a úroveň zvládnutí hodnoceného konstruktu Θ testovanými osobami.

(b) Testované osoby jsou roztříděny do malého počtu skupin podle podobnosti úrovně zvládnutí hodnoceného konstruktu Θ.

(c) Pro každou skupinu testovaných osob a pro každou testovou položku jsou vypočteny podíly testovaných osob, které zvolily správnou, respektive nesprávnou odpověď na testovou položku – vytvoření rozdělení skutečných (empirických) četností odpovědí testovaných osob.

(d) Pro každou skupinu testovaných osob a pro každou testovou položku jsou vypočteny podíly testovaných osob, které by měly podle modelu zvolit správnou, respektive nesprávnou odpověď na testovou položku – vytvoření rozdělení očekávaných (modelových) četností odpovědí testovaných osob.

(e) Empirické a modelové četnosti odpovědí testovaných osob na testovou položku jsou srovnány vzhledem ke všem skupinám testovaných osob, přičemž typicky je využívána některá z forem χ^2 statistiky, která po zobecnění nabývá podoby:

$$\chi^2 = \sum_K \sum_J \frac{n_k(O_{kj} - E_{kj})^2}{E_{kj}},$$

kde O_{kj} je empirické rozdělení četností odpovědí testovaných osob; E_{kj} je modelové rozdělení četností odpovědí testovaných osob pro skupinu testovaných osob k na dané úrovni zvládnutí hodnoceného konstruktu Θ a pro testovou položku j; a n_k je počet testovaných osob ve skupině k. Nižší hodnoty χ^2 statistiky indikují vyšší kvalitu testové položky, tj. vyšší soulad empirických a modelových dat (např. Reise, 1990). Doplňme, že statisticky významná p-hodnota χ^2 statistiky vede k zamítnutí nulové hypotézy, že odhadovaný model je dobrým odhadem skutečně pozorovaných empirických dat (např. DeMars, 2010; Toland, 2014).

[47] Uveďme, že Maydeu-Olivares (2015) považuje statistiky dobré shody za zvláštní případ indexů dobré shody, které jsou požívány pro testování hypotéz, a je k nim tedy k dispozici teoretické rozdělení.

Tabulka č. 11 uvádí speciální případy χ^2 statistiky užívané pro hodnocení dobré shody testové položky vzhledem k empirickým datům.

Tabulka č. 11: Speciální případy χ^2 statistiky pro hodnocení dobré shody testové položky

Speciální případ χ^2 statistiky	Charakteristika speciálního případu χ^2 statistiky
Yenovo Q_1	Yenovo Q_1 rozděluje testované osoby do 10 skupin podle úrovně jejich zvládnutí hodnoceného konstruktu Θ, a to s přibližně stejným počtem testovaných osob v každé skupině. Modelový podíl dané odpovědi na testovou položku je počítán jako predikovaná pravděpodobnost této odpovědi na úrovni průměrné hodnoty zvládnutí hodnoceného konstruktu Θ testovanými osobami dané skupiny (např. Reise, 1990; Stone, 2000; Orlando a Thissen, 2000; Stone a Zhang, 2003; Chalmers a Ng, 2017).
Bockův χ^2	Bockův χ^2 se od Yenova Q_1 odlišuje: (a) možností různého počtu utvářených skupin testovaných osob podle úrovně jejich zvládnutí hodnoceného konstruktu Θ; a (b) využitím mediánu úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami dané skupiny pro stanovení modelového podílu dané odpovědi na testovou položku (např. Stone a Zhang, 2003; Orlando a Thissen, 2000; Stone 2000; Chalmers a Ng, 2017).
G^2 statistika	G^2 statistika je počítána s využitím věrohodnostního poměru, že se empirická hodnota vyskytne vzhledem k modelové hodnotě. Výpočet G^2 statistiky pak je založen na vztahu (např. Orlando a Thissen, 2000): $$G_i{}^2 = 2\sum_{k=1}^{10} n_k \left[O_{ki} \times ln\left(\frac{O_{ki}}{E_{ki}}\right) + (1 - O_{ki}) \times ln\left(\frac{1 - O_{ki}}{1 - E_{ki}}\right)\right],$$ kde O_{ki} je empirické rozdělení četností odpovědí testovaných osob; E_{ki} je modelové rozdělení četností odpovědí testovaných osob pro skupinu k s úrovní jejich zvládnutí hodnoceného konstruktu Θ a pro odpověď na testovou položku i; a kde n_k je počet testovaných osob ve skupině testovaných osob k. Uveďme, že výpočet je podobný jako v případě Yenova Q_1.

Orlando a Thissen (2000), Stone a Zhang (2003), Stone (2000), Chalmers a Ng (2017) uvádějí, že tradiční statistiky dobré shody testových položek se potýkají s některými opakujícími se problémy: (a) Výpočet statistik dobré shody testových položek závisí na modelu, pomocí něhož je odhadována úroveň zvládnutí hodnoceného konstruktu Θ testovanými osobami, což má následně dopad na nejasný počet stupňů volnosti pro výpočet χ^2 statistik. (b) Výpočet statistik dobré shody testových položek závisí na způsobu, který byl použit pro vytvoření skupin testovaných osob vzhledem k úrovni zvládnutí hodnoceného konstruktu Θ. (c) χ^2 statistika je citlivá jednak k velmi nízkým modelovým hodnotám a jednak k počtu testovaných osob, kdy problémy lze pozorovat především v případě velkých výběrových souborů. Ke třem uváděným problémům pak Stone (2000) doplňuje nepřesnosti spojené s chybnou klasifikací testované osoby do některé ze skupin v důsledku nepřesného odhadu úrovně jejího zvládnutí

hodnoceného konstruktu Θ, přičemž silněji je takový problém pozorován v případě krátkých testů s nízkým počtem testových položek.

Uvedené problémy jsou motivací k hledání alternativních statistik dobré shody na úrovni testové položky. Jeden z možných přístupů je založený na shlukování testovaných osob podle empirických dat, nikoliv podle úrovně zvládnutí hodnoceného konstruktu Θ, přičemž často uváděným je v tomto ohledu přístup využívající úspěšnost těchto osob v testu pro definování jejich skupin na bázi všech možných hodnot celkového skóre, kterých je možné dosáhnout (např. DeMars, 2010; Orlando a Thissen, 2000). V rámci každé skupiny je opětovně počítán rozdíl mezi empirickým a modelovým rozdělením odpovědí testovaných osob, kdy empirické rozdělení odpovědí je odvozeno přímo z pozorovaných dat, zatímco odvození modelového rozdělení odpovědí je založeno na úrovni zvládnutí hodnoceného konstruktu Θ (např. Orlando a Thissen, 2000). Z uvedeného postupu vycházejí modifikované verze χ^2 a G^2 statistiky, které jsou označovány jako $S\text{-}\chi^2$ a $S\text{-}G^2$ statistiky, přičemž řada studií ukázala lepší schopnost těchto statistik správně hodnotit úroveň dobré shody testových položek (např. DeMars, 2010; Orlando a Thissen, 2000).

Další z alternativních přístupů je zaměřen na zohlednění problému nejistoty spojené s odhadem úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami. Takto Stone (2000), Stone a Zhang (2003) navrhli metodiku založenou na výpočtu aposteriorní pravděpodobnosti, že testovaná osoba dosahuje různé úrovně zvládnutí hodnoceného konstruktu Θ v závislosti na svém vzoru odpovědí na testové položky. Pravděpodobnost odpovědi testované osoby na danou testovou položku pak není přisouzena jediné úrovni zvládnutí hodnoceného konstruktu Θ, nýbrž je rozdělena mezi několik různých, i když sobě blízkých, úrovní. Předmětem zájmu metodiky pak je výpočet aposteriorního rozdělení empirických odpovědí testovaných osob v rámci množiny několika definovaných úrovní zvládnutí hodnoceného konstruktu Θ, kdy aposteriorní pravděpodobnost rozděluje testované osoby mezi několik úrovní zvládnutí hodnoceného konstruktu Θ (tzv. pseudo-četnosti). Pro danou testovou položku je pseudo-četnost odpovědi j na úrovni zvládnutí hodnoceného konstruktu Θ_k počítána s využitím Bayesova přístupu k aposteriorní pravděpodobnosti, a to vztahem (např. Stone, 2000):

$$r_{j\theta_k} = \sum_{N} x_{jn} \frac{P\left(x_n \middle| X_{\theta_k}\right) P(X_{\theta_k})}{P(x_n)},$$

kde N je počet testovaných osob; x_{jn} je rovno jedna, pokud testovaná osoba n vybrala odpověď j na danou testovou položku a jinak je rovno nule; $P(x_n \mid X_{\Theta k})$ je podmíněná pravděpodobnost daného vzoru odpovědí testované osoby n na množinu testových položek na úrovni zvládnutí hodnoceného konstruktu Θ_k; $P(X_{\theta_k})$ je váha úrovně zvládnutí hodnoceného konstruktu Θ_k testovanými osobami v rámci apriorního rozdělení úrovně zvládnutí hodnoceného konstruktu Θ testovaných osob – typicky v rámci normálního rozdělení[48]; a $P(x_n)$ odpovídá nepodmíněné pravděpodobnosti výskytu daného vzoru odpovědí testované osoby n na množinu testových

[48] Váhu jednotlivých úrovní zvládnutí hodnoceného konstruktu testovanými osobami lze ovšem odvozovat i přímo z aposteriorního rozdělení skutečných (empirických) dat (např. Stone, 2000).

položek. Uveďme, že celkový součet pseudo-četností $r_{j\theta_k}$ odpovídá počtu testovaných osob (např. Stone, 2000).

S využitím výše popsaného metodického přístupu jsou počítány hodnoty empirického aposteriorního rozdělení odpovědí testovaných osob (pseudo-četnosti) vzhledem k možným odpovědím na testovou položku *j* a vzhledem k úrovním zvládnutí hodnoceného konstruktu Θ_k testovanými osobami (např. Stone, 2000). Modelové rozdělení odpovědí testovaných osob je primárně založeno na výpočtu pravděpodobnosti dosažení odpovědi *j* na danou testovou položku při úrovni zvládnutí hodnoceného konstruktu Θ_k testovanými osobami, přičemž tato pravděpodobnost je pak pro každou odpověď na testovou položku *j* a pro každou úroveň zvládnutí hodnoceného konstruktu Θ_k testovanými osobami vynásobena korespondující pseudo-četností pro danou úroveň zvládnutí hodnoceného konstruktu Θ_k testovanými osobami (např. Stone, 2000). Stone (2000) následně uvádí, že srovnáním empirických a modelových odpovědí testovaných osob je možné vypočítat hodnotu χ^2 statistiky, která je označována jako χ^{2*}. Zároveň však vzájemná závislost četností vznikající ze zahrnutí odpovědí testovaných osob do více úrovní jejich zvládnutí hodnoceného konstruktu Θ neumožňuje použití tradičního χ^2 rozdělení pro stanovení p-hodnoty. Stone a Zhang (2003), Stone (2000) uvádějí dvě možná řešení tohoto problému:

- První řešení je založeno na odvození teoretického rozdělení na bázi Monte-Carlo simulací, které využívají znalosti modelu vycházejícího z IRT a rozdělení úrovní zvládnutí hodnoceného konstruktu Θ.
- Druhé řešení využívá QDH rozdělení pro testování (např. Donoghue a Hombo, 2001), a to bez ohledu na přítomnost malých četností v rozdělení. Stone a Zhang (2003) však poukazují na nízkou kvalitu tohoto rozdělení ve vazbě na statistické ověřování robustnosti závěrů.

Konečně uveďme, že Stone (2000) zdůrazňuje výhody aposteriorního rozdělení odpovědí testovaných osob především v kontextu krátkých testů, zároveň však poukazuje na možné problémy související s nízkým počtem odpovědí při extrémních hodnotách úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami. Za možné řešení pak Stone (2000) navrhuje výpočet χ^{2*} statistiky pro interval Θ od -2 do +2.

Konečně Chalmers a Ng (2017) navrhli řešit problém spolehlivosti odhadů úrovně zvládnutí hodnoceného konstruktu Θ testovanými osobami jiným způsobem, který označili jako *PV-Q₁* statistika. Podstata tohoto přístupu je založena na využití algoritmu pro generování dostatečně velkého počtu množin, které obsahují 10 hodnot úrovně zvládnutí hodnoceného konstruktu Θ_m testovanými osobami, které jsou utvářené s využitím odhadovaného modelu. Pro takto vytvořené množiny hodnot Θ_m je následně odhadováno Yenovo Q_1, přičemž *PV-Q₁* statistika je počítána pomocí vztahu (Chalmers a Ng, 2017):

$$PV - Q1 = \frac{Q_1(\theta_1{}^*) + \cdots + Q_1(\theta_M{}^*)}{M}.$$

Výpočet statistiky je tak založen na myšlence konvergence vysokého počtu hodnot $Q_1(\theta_M{}^*)$ ke skutečné hodnotě, přičemž Chalmers a Ng (2017) hovoří o stabilitě řešení již při 30 iteracích. Příslušná p-hodnota je počítána z χ^2 rozdělení. V tomto ohledu však Chalmers a Ng (2017)

upozorňují na skutečnost, že $PV\text{-}Q_1$ statistika nesleduje přesně χ^2 rozdělení, a proto doporučují řešit tento problém na bázi Monte-Carlo simulací. Takto je s využitím parametrů příslušného modelu generován dostatečně velký počet datových souborů možných odpovědí testovaných osob, pro které je počítána $PV\text{-}Q_1$ statistika a k nim příslušná p-hodnota z χ^2 rozdělení s tím, že p-hodnota pro hodnocené odpovědi testovaných osob je porovnána s množinou generovaných p-hodnot pro ověření dobré shody testové položky.

Celkově Chalmers a Ng (2017) doporučují využití $PV\text{-}Q_1$ statistiky a S- χ^2 statistiky pro rychlé, ale méně přesné, hodnocení dobré shody testové položky. Pro hlubší poznání dobré shody testové položky pak Chalmers a Ng (2017) doporučují aplikovat přístup χ^{2*} statistiky a $PV\text{-}Q_1^*$ statistiky, který využívá bootstrap a Monte-Carlo simulace. Konečně uveďme, že $PV\text{-}Q_1$, $PV\text{-}Q_1^*$ a χ^{2*} statistiky se ukazují být efektivní i v situacích, kdy je databáze odpovědí poznamenána vysokým počtem chybějících hodnot (např. Chalmers a Ng, 2017).

(B) Úroveň testované osoby

Hodnocení dobré shody modelových a empirických dat na úrovni testované osoby je spojeno se záměrem detekovat korektnost měřené úrovně zvládnutí hodnoceného konstruktu Θ touto osobou. Motivace k takovému kroku může být obdobná jako v případě hledání neobvyklého vzoru odpovědí testovaných osob na testované položky v případě přístupů CTT, tedy (např. Tendeiro, Meijer a Niessen, 2016; DeMars, 2010):

- neetické chování testovaných osob, včetně znalosti testových položek dopředu;
- nízká motivace testovaných osob pro vyplňování testu vedoucí k hádání odpovědí;
- nedostatečné znalosti testovaných osob v některé z testovaných oblastí.

DeMars (2010) poukazuje na vysoký význam hodnocení dobré shody modelu na úrovni testované osoby především v případě odhadů Raschova modelu, kdy významný nesoulad může být motivací k vynechání odpovědí testované osoby z vyhodnocení.

Pro hodnocení souladu empirických a modelových dat na úrovni testované osoby bylo rovněž v přístupech vycházejících z IRT navrženo několik statistik dobré shody, jejichž konstrukce vychází z posouzení, zda je empirický vzor odpovědí testované osoby na testové položky neobvyklý ve srovnání s modelem odhadovaným na základě vzoru odpovědí všech testovaných osob. Předpokládá se také vyšší pravděpodobnost správné odpovědi testované osoby na méně obtížné testové položky a naopak vyšší pravděpodobnost nesprávné odpovědi testované osoby na více obtížné testové položky (např. Tendeiro, Meijer a Niessen, 2016). Hodnocení souladu empirických a modelových dat na úrovni testované osoby pak může probíhat buď vizuálně, nebo s využitím navržených statistik dobré shody, které Tendeiro, Meijer a Niessen (2016) řadí do dvou typů:

- statistiky vycházející z věrohodnostního poměru v logaritmické míře, tj. z odhadu modelu vycházejícího z IRT;
- statistiky založené na odpovědích skupin testovaných osob, tj. neparametrické statistiky nevyžadující odhad parametrických modelů vycházejících z IRT (blíže viz podkapitola 3.1.1).

Tabulka č. 12 poskytuje přehled hlavních statistik dobré shody empirických a modelových dat na úrovni testované osoby, které vycházejí z modelů IRT.

Tabulka č. 12: Statistiky dobré shody empirických a modelových dat na úrovni testované osoby (modely vycházející z IRT)

Statistika	Charakteristika statistiky
l_o	Statistika l_o je vyjádřena vztahem: $$l_0 = logL(\theta_n) = \sum_I x_{ni} \times lnP_i(\theta_n) + (1 - x_{ni}) \times \ln\big(1 - P_i(\theta_n)\big),$$ kde x_{ni} je vzor odpovědí testované osoby n na testové položky i a P_i je model vycházející z IRT (např. 1PL, 2PL, 3PL model). Protože podstata metodického přístupu usiluje o maximalizaci hodnot věrohodnostního poměru v logaritmické míře, je horší shoda empirických a modelových dat na úrovni testované osoby spojená s nízkou hodnotou statistiky l_o (např. Tendeiro, Meijer a Niessen, 2016).
l_z a l_z^*	Statistika l_z vylepšuje statistiku l_o o prvek standardizace a normalizace, klade však v praxi nereálný požadavek na znalost skutečné úrovně zvládnutí hodnoceného konstruktu testovanými osobami. Statistika l_z^* proto zavádí korekci statistiky l_z tak, aby bylo možné pracovat s modelovými úrovněmi zvládnutí hodnoceného konstruktu Θ testovanými osobami. Protože podstata metodického přístupu usiluje o maximalizaci hodnot věrohodnostního poměru v logaritmické míře je horší shoda empirických a modelových dat na úrovni testované osoby spojená s nízkou hodnotou statistiky l_z a l_z^* (např. Tendeiro, Meijer a Niessen, 2016).

(C) Úroveň testu (modelu)

Hlavním záměrem výpočtu indexů a statistik dobré shody na úrovni testu (modelu) je posoudit, jak dobře model odhadovaný na základě IRT charakterizuje empirická data (např. Maydeu-Olivares a García-Forero, 2010), neboli zda empirická data mohla být vygenerována vlastním modelem (např. Maydeu-Olivares, 2015). Maydeu-Olivares a García-Forero (2010), Chen, de la Torre a Zhang (2013) v tomto ohledu rozlišují dva odlišné typy evaluací, které využívají indexy a statistiky dobré shody na úrovni testu (modelu):

- Relativní evaluace sleduje cíl vybrat z hodnocených modelů model s nejlepším souladem s empirickými daty, podstata evaluace tedy funguje na bázi srovnání indexů a statistik dobré shody množiny odhadovaných modelů.
- Absolutní evaluace naopak sleduje cíl posoudit míru souladu empirických a modelových dat, přičemž statistika dobré shody by měla s vysokou pravděpodobností zamítnout modely, jejichž soulad s empirickými daty je omezený.

Tradiční statistiky dobré shody na úrovni testu (modelu) opětovně vycházejí z porovnání empirických a modelových četností odpovědí testovaných osob na testové položky (např. DeMars, 2010; Bartholomew a Tzamourani, 1999). Takto je možné odpovědi testovaných osob na n dichotomických testových položek rozdělit do 2^n množin různých vzorů odpovědí (p_c). Analogické rozdělení četností je možné vytvořit pro modelem generovaná data (π_c).

Pro srovnání takto konstruovaných rozdělení četností odpovědí testovaných osob, tj. pro testování nulové hypotézy o souladu modelových a empirických dat, jsou následně využívány dvě tradiční statistiky (např. Maydeu-Olivares, 2015; Maydeu-Olivares a García-Forero, 2010; Bartholomew a Tzamourani, 1999; Maydeu-Olivares, Cai a Hernández, 2011):

- χ^2 statistika počítána vztahem $\chi^2 = N \sum_C \frac{(p_c - \pi_c)^2}{\pi_c}$, [49]
- věrohodnostní poměr $G^2 = 2N \sum_C p_c \times ln\left(\frac{p_c}{\pi_c}\right)$,

kde *C* odpovídá počtu možných kombinací odpovědí testovaných osob (např. Maydeu-Olivares, 2015).

Tradiční statistiky dobré shody empirických a modelových dat se na úrovni testu (modelu) potýkají s problémy, které jsou charakteristické pro testy založené na χ^2 rozdělení a k nimž především patří požadavek na minimální hodnoty očekávaných četností v buňkách rozdělení četností (např. Maydeu-Olivares a García-Forero, 2010). Naplnění právě tohoto požadavku bývá v případě modelů vycházejících z IRT obtížné, neboť řada možných vzorů odpovědí nabývá velmi malé četnosti výskytu (např. Bartholomew a Tzamourani, 1999). Tato skutečnost následně má negativní dopad na přesnost odhadu statistik (např. Maydeu-Olivares, 2015).[50] Maydeu-Olivares (2015), Maydeu-Olivares a García-Forero (2010), Bartholomew a Tzamourani (1999) uvádějí dva možné přístupy k řešení problému vysokého počtu nízkých hodnot v matici očekávaných četností χ^2 rozdělení:

- První přístup je založený na metodě Monte Carlo simulací (bootstrap), které jsou konstruovány s využitím parametrů modelu vycházejícího z IRT. Tyto simulace generují rozdělení *p*-hodnot jako referenční rozdělení pro posouzení *p*-hodnoty odhadovaného modelu (např. Bartholomew a Tzamourani, 1999). Za problém tohoto přístupu však Maydeu-Olivares (2015) označuje přesnost odhadů z takto utvářených simulací.
- Druhý přístup řeší problém vysokého počtu nízkých hodnot v rozdělení četností prostřednictvím slučování sousedních kategorií, což zároveň vede k redukci jejich počtu. Takový přístup však nezaručuje, že problém bude nutně vyřešen a zároveň může docházet ke ztrátě podstatné informace empirických dat (např. Bartholomew a Tzamourani, 1999).

S ohledem na uvedené problémy doporučují Maydeu-Olivares (2015), Maydeu-Olivares a García-Forero (2010) sledovat třetí přístup, který je založený na využití statistik s tzv. omezenou informací. Podstata tohoto přístupu je de facto založena na slučování informací tak, aby byl odstraněn vliv nízkých hodnot v rozdělení četností. Prakticky je využita konstrukce četnostních tabulek na několika úrovních (srovnej s Maydeu-Olivares, 2015; Maydeu-Olivares a García-Forero, 2010):

[49] Platí, že χ^2 rozdělení má *C-q-1* stupňů volnosti, kde *q* je počet odhadovaných parametrů modelu.

[50] Uveďme, že Maydeu-Olivares a García-Forero (2010) doporučují porovnat *p*-hodnoty získané výpočtem Pearsonovy χ^2 statistiky a věrohodnostního poměru s tím, že pokud se tyto hodnoty od sebe odlišují, je odůvodněné předpokládat, že jejich odhad není správný.

- Úroveň 1 je spojena s pravděpodobnosti výskytu odpovědi na testovou položku bez ohledu na hodnoty testových položek ostatních s tím, že pravděpodobnost je počítána pro všechny odpovědi na testové položky. Zdůrazněme, že celkový součet pravděpodobností je jedna.
- Úroveň 2 je spojena s pravděpodobnosti výskytu kombinace dvou odpovědí na dvě testové položky, opětovně bez ohledu na hodnoty testových položek ostatních s tím, že pravděpodobnost je počítána pro všechny dvojice odpovědí testových položek. Zdůrazněme, že celkový součet pravděpodobností je jedna.
- Úroveň *n*, tj. nejvyšší úroveň, je spojena s pravděpodobností výskytu všech možných vzorů odpovědí na všech *n* testových položek, přičemž celkový součet těchto pravděpodobností je i v tomto případě roven jedné.

Je zřejmé, že nejvyšší úroveň konstrukce četnostních tabulek v sobě obsahuje veškerou informaci, kterou rozdělení četností nabízí a reálně odpovídá výpočtu χ^2 statistiky. Zároveň je ale potřeba si uvědomit, že právě v tomto případě se vyskytují situace, kdy výskyt řady vzorů odpovědí na testové položky je velmi vzácný, neboť je jen málo pravděpodobné, že by testované osoby daným způsobem testové položky zodpovídali. Řešení na bázi výpočtu statistik s omezenou informací vychází z úvahy, že jsou do výpočtu vzaty pouze nižší úrovně konstrukce četnostních tabulek, čímž je řešen problém nízkých četností tím, že jsou spojeny s kategoriemi jinými (srovnej s Maydeu-Olivares a García-Forero, 2010; Maydeu-Olivares, 2015).

Maydeu-Olivares (2015) shrnuje výše uvedenou úvahu v tvrzení, že testování založené na statistikách dobré shody, které vycházejí z konceptu omezené informace, využívá data četnostních tabulek úrovně *r* nižší, než je nejvyšší úroveň *n* ($r < n$), a to za účelem lepšího odhadu *p*-hodnoty, přičemž vlastní statistiky jsou označovány jako M_r, kde hodnota *r* odpovídá počtu úrovní využívaných ve výpočtu statistiky. Maydeu-Olivares (2015), Maydeu-Olivares a García-Forero (2010) zároveň zmiňují doporučení využívat statistiku počítanou ze dvou úrovní (momentů), tj. statistiku M_2. Uveďme, že s ohledem na podobu konstrukce statistiky M_2 s její vazbou na χ^2 rozdělení jsou pro dosažení dobré shody empirických a modelových dat žádoucí nízké hodnoty vlastní statistiky a nevýznamné *p*-hodnoty (např. DeMars, 2010; Toland, 2014).

Maydeu-Olivares (2015) upozorňuje na spojení testu s takovým množství informací, že není realistické předpokládat jejich přesný odhad jakýmkoliv modelem vycházejícím z IRT. Z tohoto důvodu Maydeu-Olivares (2015) hovoří o možné preferenci hodnocení dobré shody na bázi nikoliv přesného, ale přibližného souladu empirických a modelových dat. Podstata takového přístupu je založena na využití indexů, které nejsou využívány pro testování hypotéz, nýbrž jsou spojeny se stanovením tzv. *cut-off* hodnot, kterou model nemůže překročit, pokud má být naplněna dobrá shoda empirických a modelových dat. Tabulka č. 13 představuje příklady takových indexů, přičemž doplněny jsou indexy vhodné pro srovnání dobré shody vyššího počtu modelů vycházejících z IRT.

Tabulka č. 13: Indexy hodnotící dobrou shodu empirických a modelových dat na úrovni testu (modelu)

Index	Charakteristika indexu
RMSEA₂	Index *RMSEA₂* vychází ve svém výpočtu ze statistiky M_2 a vztahu: $$RMSEA_2 = \sqrt{\frac{M_2 - df_2}{N \times df_2}},$$ kde df_2 je počet stupňů volnosti modelu a *N* je počet testovaných osob. *Cut-off* hodnota je v případě indexu *RMSEA₂* stanovena ve výši 0,05 (dichotomické testové položky), přičemž pro dobrou shodu empirických a modelových dat jsou žádoucí nižší hodnoty (např. Maydeu-Olivares, 2015).
SRMSR	Index *SRMSR* vychází z výpočtu tzv. standardizovaných reziduí dvojic testových položek, které odpovídají rozdílu empirické a modelové korelace těchto položek, přičemž modelová korelace je počítána jako modelová kovariance dělená modelovou směrodatnou odchylkou. Index *SRMSR* je následně odmocninou průměru umocněných standardizovaných reziduí. *Cut-off* hodnota dobré shody je stanovena ve výši 0,05, přičemž žádoucí jsou nižší hodnoty indexu *SRMSR* (např. Maydeu-Olivares, 2015).
-2LL, AIC, BIC	Pro srovnání dobré shody vyššího počtu modelů vycházejících z IRT je možné využít indexy počítané s využitím věrohodnostního poměru v logaritmické míře. Tyto indexy především zahrnují (např. Toland, 2014; DeMars, 2010; Maydeu-Olivares a García-Forero, 2010; Chen, de la Torre a Zhang, 2013): • index *-2LL* odpovídající hodnotě mínus dvojnásobku hodnoty věrohodnostního poměru v logaritmické míře; • index Akaikeho informačního kritéria (*AIC*); • index Bayesova (Schwarzova) informačního kritéria (*BIC*). V případě těchto tří indexů je vybírán jako nejlepší model, který má nejnižší hodnotu *-2LL*, *AIC* nebo *BIC* (např. Chen, de la Torre a Zhang, 2013; DeMars, 2010; Maydeu-Olivares a García-Forero, 2010; Toland, 2014). Doplňme, že Maydeu-Olivares a García-Forero (2010) charakterizují *AIC* a *BIC* jako indexy, které penalizují modely s vyšším počtem odhadovaných parametrů a tímto způsobem preferují úspornější modely vycházející z IRT.

Konečně uveďme, že indexy a statistiky dobré shody mají široké praktické využití, které mimo jiné zahrnují poskytnutí argumentů: (a) pro rozhodnutí o vynechání problematické testové položky z testu; (b) pro identifikaci problémové testované osoby ve výsledku testu; a (c) pro rozhodnutí o nejvhodnější podobě odhadovaného modelu.

3.2.5 Společná škála testů, equating – přístup vycházející z IRT

Základní podstata propojení skóre testů na společnou škálu (*equating*) byla představena v části věnované CTT, v této podkapitole jsou proto zdůrazněna především specifika přístupu vycházejícího z IRT, který je v obecné podobě založen na třech dílčích krocích (např. Cook a Eignor, 1991; Sansivieri, Wiberg a Matteucci, 2018):

- vytvoření plánu sběru dat (např. SG, EG či NEAT *equating design*) a odhad parametrů zvoleného modelu (IRT) pro nový a referenční test;
- transformace škály nového a referenčního testu v případě využití NEAT přístupu ke sběru dat;
- equating skóre testů vycházející z IRT.

- **Transformace škály úrovně zvládnutí hodnoceného konstruktu Θ**

Transformovat škálu úrovně zvládnutí hodnoceného konstruktu Θ je potřebné při volbě NEAT přístupu ke sběru dat, protože v tomto případě nejsou skupiny testovaných osob ekvivalentní.[51] Sansivieri, Wiberg a Matteucci (2018) blíže charakterizují postup transformace škály úrovně zvládnutí hodnoceného konstruktu Θ, a to při sledování NEAT přístupu ke sběru dat pro odhad 3PL modelu.[52]

Východiskem transformace škály úrovně zvládnutí hodnoceného konstruktu Θ je lineární vztah pro transformaci škály K nového testu na škálu J referenčního testu, kdy v případě referenčního testu jsou fixovány parametry odhadovaného modelu:

$$\theta_K = A^*\theta_J + B^*.$$

Předmětem zájmu transformace škály úrovně zvládnutí hodnoceného konstruktu Θ tak je nalezení nejlepších hodnot dvou parametrů lineárního vztahu A^* a B^* (Cook a Eignor, 1991), které lze určit prostřednictvím vztahů (např. Sansivieri, Wiberg a Matteucci, 2018):

$$A^* = \frac{\sigma(b_K)}{\sigma(b_J)} = \frac{\overline{a_J}}{\overline{a_K}} = \frac{\sigma\big(\theta(b_K)\big)}{\sigma\left(\theta\big(b_J\big)\right)},$$

$$B^* = \overline{b_K} - A^*\overline{b_J} = \overline{\theta_K} - A^*\overline{\theta_J}.$$

Z těchto vztahů je zřejmé, že transformace škály úrovně zvládnutí hodnoceného konstruktu Θ může být založena: (a) na obtížnosti testových položek b; (b) na diskriminaci testových položek a; a (c) na úrovni zvládnutí hodnoceného konstruktu Θ, nejčastěji je však v tomto ohledu využíván parametr obtížnosti testových položek – směrodatná odchylka pro nalezení parametru A^* a průměr pro nalezení parametru B^* (např. Cook a Eignor, 1991; Sansivieri, Wiberg a Matteucci, 2018). Uveďme, že s transformací škály úrovně zvládnutí hodnoceného konstruktu Θ je spojena také kalibrace (nový odhad) parametrů testových položek škály K vzhledem ke škále J, a to prostřednictvím následujících vztahů (např. Cook a Eignor, 1991; Sansivieri, Wiberg a Matteucci, 2018):

$$a_K = \frac{a_J}{A^*};\ b_K = A^*b_J + B^*;\ c_K = c_J,$$

[51] Naopak v případě sledování SG a EG plánu sběru dat není transformace škály úrovně zvládnutí hodnoceného konstruktu Θ potřebná, protože používaná škála je stejná s ohledem na vzájemně ekvivalentní výběrové soubory testovaných osob (např. Sansivieri, Wiberg a Matteucci, 2018; Cook a Eignor, 1991).

[52] Cook a Eignor (1991) doporučují minimální velikost výběrového souboru v intervalu 2,5 až 3 tisíce testovaných osob.

kde *a* je parametr diskriminace, *b* parametr obtížnosti a *c* dolní asymptota, tj. parametr „pseudo-hádání“, testových položek (např. Sansivieri, Wiberg a Matteucci, 2018).

Praktickou otázkou, která vede k řešení výše uvedených vztahů, je, jakým způsobem odhadovat hodnoty parametrů A^* a B^*. V tomto ohledu se nabízí několik možností, jejichž typickým znakem je využití informací z kotvícího testu (např. Cook a Eignor, 1991; Sansivieri, Wiberg a Matteucci, 2018; Kilmen a Demirtasli, 2012):

- Metoda průměr/sigma odhaduje hodnoty parametrů A^* a B^* s využitím hodnot průměru a směrodatné odchylky obtížnosti testových položek kotvícího testu b_{jJ} a b_{jK}, přičemž metoda usiluje o to, aby po transformaci byly průměr a směrodatná odchylka obtížnosti těchto testových položek v obou testech stejné.
- Metoda průměr/průměr je analogií k metodě průměr/sigma, nicméně hodnoty parametrů A^* a B^* jsou odhadovány s využitím hodnot průměrů diskriminace testových položek kotvícího testu a_{jJ} a a_{jK} a obtížnosti testových položek kotvícího testu b_{jJ} a b_{jK}.

Nedostatkem uvedených metodických přístupů je ta skutečnost, že neodhaduje všechny parametry testových položek najednou. Pro řešení tohoto nedostatku bylo navrženo několik metodických přístupů vycházejících z podoby ICC křivek testových položek a základního vztahu (např. Sansivieri, Wiberg a Matteucci, 2018):

$$P_{ij}(\theta_{iK}; a_{jK}; b_{jK}; c_{jK}) = P_{ij}\left(A^*\theta_{iJ} + B^*; \frac{a_{jJ}}{A^*}; A^*b_{jJ} + B^*; c_{jJ}\right),$$

přičemž platí, že uvedená rovnost nebude platit dokonale. Záměrem proto je nalézt takové hodnoty parametrů A^* a B^*, které minimalizují kritérium odlišnosti obou vztahů, přičemž za tímto účelem jsou opětovně využívány testové položky kotvícího testu. Využívány jsou především Haebarův a Stocking-Lordův přístup, které se odlišují v podobě minimalizačního kritéria (např. Sansivieri, Wiberg a Matteucci, 2018).

- **Propojení škál testů (*equating*)**

Přístup založený na IRT je konečně možné využít k propojení skóre nového a referenčního testu (alternativa k přístupu CTT). Metodický postup se v tomto ohledu skládá z následujících kroků (např. Sansivieri, Wiberg a Matteucci, 2018):

- V prvním kroku je vytvořeno podmíněné pravděpodobnostní rozdělení pozorovaných skóre testovaných osob v novém testu *X*, a to v závislosti na úrovni jejich zvládnutí hodnoceného konstruktu Θ. Takto jsou pro různé hodnoty úrovně zvládnutí hodnoceného konstruktu Θ stanoveny pravděpodobnosti dosažení možných skóre v novém testu *X*.
- Ve druhém kroku je podmíněné pravděpodobnostní rozdělení pozorovaných skóre testovaných osob v novém testu *X* využito pro odvození marginálního rozdělení skóre v testu *X*, tj. skóre, které nezávisí na úrovni zvládnutí hodnoceného konstruktu Θ testovaných osob. Využit je vztah:

$$f(x) = \int f(X|\theta)h(\theta)d\theta,$$

kde $h(\Theta)$ je rozdělení počtu testovaných osob s danou úrovní zvládnutí hodnoceného konstruktu Θ v testu X.

- Ve třetím kroku metodického postupu je odvozené marginální rozdělení skóre v novém testu X využito pro vytvoření kumulativní distribuční funkce skóre v testu. Analogický postup je sledován pro odvození kumulativní distribuční funkce skóre referenčního testu Y.
- V posledním kroku je pro propojení skóre nového testu X a referenčního testu Y využit ekvipercentilní přístup k propojování škál testů.

Analogicky k CTT lze propojení skóre nového a referenčního testu rozšířit o vztah ke třetí škále, která může být využita pro reporting výsledků (např. Cook a Eignor, 1991).

3.3 Vyhodnocení a reporting výsledků testu

Nedílnou součástí ověřovacího testování v počátečním vzdělávání je vyhodnocení a reporting výsledků. Základní vyhodnocení je přirozeně spojeno s výpočtem jednoduchých ukazatelů deskriptivní a inferenční statistiky (např. střední hodnota, rozptyl dat, rozdělení četností) a jejich následným reportingem. Základní otázky, které je potřeba v tomto kontextu řešit, zahrnují:

- výběr škály, na které je test vyhodnocován a výsledky následně reportovány (např. dosažené skóre, úspěšnost v testu, škála vycházející z IRT, percentilové pořadí), a to včetně případného rozlišení dílčích škál ověřovacího testu (např. multidimenzionální modely vycházející z IRT);
- výběr úrovně hodnocení zahrnující, a to například úrovně: (a) testované osoby; (b) třídy; (c) školy; (d) území; a (e) systému;
- výběr dílčích charakteristik hodnocených úrovní (např. testovaná osoba, třída, škola, území a systém), vůči kterým je test vyhodnocován a výsledky následně reportovány.

Pro komplexní vyhodnocení a reporting výsledků jsou odbornou literaturou akcentované další metodické přístupy. V tomto ohledu je jednou z hlavních otázka, které charakteristiky různých úrovní hodnocení (např. testovaná osoba, třída, škola, území a systém) jsou nejvíce vztaženy k dosaženým výsledkům testovaných osob v testech. O'Dwyer a Parker (2014) uvádějí, že využití tradičních regresních modelů je zde komplikováno narušením předpokladu nezávislosti pozorování, které je typicky dáno umístěním testovaných osob v dané třídě a škole, kdy vzniká závislost působením řady faktorů na těchto úrovních (např. Tabachnick a Fidell, 2007). Tradiční řešení uvedeného problému vychází z odhadů hierarchických regresních modelů.

3.3.1 Hierarchické regresní modely

Tabachnick a Fidell (2007) hovoří o hierarchických regresních modelech jako o struktuře dat organizované na vyšším počtu úrovní, například: (a) testovaná osoba; (b) třída; a (c) škola, přičemž pro tyto úrovně jsou definovány různé proměnné. Díky struktuře utvářené hierarchickými regresními modely platí, že je možné upustit od striktního požadavku

jednoúrovňových regresních modelů na nezávislost chybového členu.[53] Takto hierarchické regresní modely umožňují, aby se hodnoty parametrů navzájem lišily na vyšší úrovni analýzy, kdy se například vztah mezi dosaženým skóre testovaných osob a jejich motivací může odlišovat mezi různými školami. Zároveň platí, že právě školy typicky jsou – na rozdíl od testovaných osob – vybírány z populace náhodně. Celkově tak hierarchické regresní modely obsahují proměnné různých hierarchických úrovní a umožňují modelovat vztahy mezi proměnnými stejné úrovně i vztahy napříč úrovněmi (Tabachnick a Fidell, 2007).

Jednoduchý příklad hierarchického regresního modelu může být definován vztahem (např. O'Dwyer a Parker, 2014; Finch, Bolin a Kelley, 2014; Tabachnick a Fidell, 2007):

$$Y_{ij} = \gamma_{oo} + \gamma_{10}X_{1ij} + \gamma_{20}X_{2ij} + \gamma_{01}W_{1j} + r_{ij} + u_{oj},$$

kde Y_{ij} je predikovaná hodnota vysvětlované proměnné (např. dosažené skóre) pro testovanou osobu i školy j; X_{1ij} a X_{2ij} jsou dvě vysvětlující proměnné pro testovanou osobu i školy j (proměnné 1. úrovně); W_{1j} je vysvětlující proměnná školy j (proměnná 2. úrovně); r_{ij} je chyba vznikající uvnitř škol a u_{oj} je chyba vznikající mezi školami. Z výše uvedeného vztahu je zřejmé, že hierarchický regresní model pracuje s více komplexním vzorem chyby, která je utvářen: (a) jednak na 1. úrovni, tj. na úrovni testované osoby; a (b) jednak na 2. úrovni, tj. na úrovni školy. Tato skutečnost následně umožňuje počítat tzv. vnitrotřídní koeficient korelace (ICC), který odpovídá podílu rozptylu vysvětlovanému na úrovni školy a jehož hodnota tedy vysvětluje, jaký podíl rozdílů ve vysvětlované proměnné připadá na úroveň školy a jaký podíl na úroveň testované osoby. Hodnoty ICC se pohybují v intervalu od 0 do 1 (např. O'Dwyer a Parker, 2014; Finch, Bolin a Kelley, 2014; Tabachnick a Fidell, 2007). Uveďme, že předmětem zájmu odhadů hierarchických regresních modelů je dále – z hlediska vyhodnocení a reportingu výsledků – především hodnocení statistické významnosti regresních koeficientů na úrovni testované osoby i školy.

[53] Tabachnick a Fidell (2007), O'Dwyer a Parker (2014) poukazují na problémy analýzy hierarchických dat, které jsou realizovány na stejné úrovni. Takto analýza dat na vyšší úrovni (např. škola) vede ke ztrátě informace na úrovni nižší (např. testované osoby). Naopak analýza dat na úrovni nižší (např. testované osoby) vede k chybě odhadů v důsledku narušení předpokladu nezávislosti pozorování.

4. Modelové případové studie

Teoreticko-metodická východiska, která byla představena v předchozí kapitole, umožňují identifikovat řadu modelových případových studií, které jsou řešeny v kontextu tvorby testů a jejich vyhodnocení. V této kapitole jsou některé z nich blíže charakterizovány, a to při sledování řetězce informací týkajících se: (a) představení modelové případové studie; (b) charakteristiky postupu řešení modelové případové studie; a (c) interpretace výsledků modelové případové studie.

4.1 Modelové případové studie řešené s využitím klasické teorie testů

Dále uvedené modelové případové studie vycházejí z hodnocení simulovaných dat testu, který je tvořený odpověďmi tří tisíc žáků na 38 dichotomických testových položek nabízejících výběr ze čtyř možných odpovědi. Test je zaměřen na hodnocení konstruktu statistické gramotnosti žáků.

4.1.1 Hodnocení kvality testových položek

Přirozeným záměrem hodnocení testů je stanovit dosaženou úroveň hodnoceného konstruktu, v tomto případě statistické gramotnosti žáků. Kvalita škály, na níž je úroveň statistické gramotnosti žáků prostřednictvím testu měřena, ovšem závisí na kvalitě hodnocených testových položek. Před stanovením dosažené úrovně statistické gramotnosti žáků je proto záměrem této modelové případové studie zhodnotit kvalitu testových položek a na tomto základě rozhodnout o dalším postupu.

- ***Řešení***

Hodnocení kvality testových položek, které vychází z CTT, je založeno na výpočtu jejich základních charakteristik, konkrétně pak:

- obtížnosti testové položky, tj. podílu testovaných žáků, kteří zodpověděli danou testovou položku správně;
- úrovně diskriminace testové položky, tj. schopnosti testové položky rozlišit mezi žáky s různou úrovní zvládnutí hodnoceného konstruktu (statistické gramotnosti), a to s využitím upravené bodově biseriální korelace;
- hodnoty spolehlivosti testu, tj. Cronbachova alfa, při vynechání dané testové položky, a to při srovnání s referenční hodnotou spolehlivosti celého testu, který obsahuje také danou testovou položku.

Tabulka č. 14 zachycuje hodnoty uvedených charakteristik 38 testových položek hodnoceného testu statistické gramotnosti žáků. Pro označení testových položek hvězdičkou jsou využity referenční hodnoty základních charakteristik testových položek, tj. hodnota obtížnosti 0,10 pro velmi obtížné testové položky, hodnota 0,90 pro velmi jednoduché testové položky a hodnota 0,20 pro testové položky s nízkou schopností diskriminace (blíže viz kapitola 3).

Tabulka č. 14: Kvalita testových položek – základní charakteristiky

Testová položka	Obtížnost (itemMean)	Diskriminace (pBis)	Spolehlivost testu při vynechání testové položky (celý test = 0,8715)	itemMean		pBis
				vysoká	nízká	nízká
ID1	0,71	0,34	0,8689			
ID2	0,52	0,28	0,8703			
ID3	0,79	0,44	0,8673			
ID4	0,55	0,27	0,8705			
ID5	0,63	0,32	0,8693			
ID6	0,72	0,37	0,8684			
ID7	0,44	0,34	0,8690			
ID8	0,84	0,44	0,8675			
ID9	0,54	0,33	0,8692			
ID10	0,78	0,49	0,8663			
ID11	0,81	0,47	0,8668			
ID12	0,54	0,35	0,8689			
ID13	0,75	0,51	0,8658			
ID14	0,45	0,34	0,8691			
ID15	0,62	0,47	0,8663			
ID16	0,45	0,31	0,8696			
ID17	0,42	0,35	0,8687			
ID18	0,48	0,32	0,8694			
ID19	0,64	0,54	0,8649			
ID20	0,50	0,35	0,8688			
ID21	0,55	0,40	0,8677			
ID22	0,43	0,39	0,8680			
ID23	0,36	0,28	0,8701			
ID24	0,55	0,53	0,8649			
ID25	0,49	0,41	0,8676			
ID26	0,73	0,42	0,8674			
ID27	0,16	0,14	0,8721			*
ID28	0,40	0,31	0,8696			
ID29	0,45	0,32	0,8693			
ID30	0,41	0,29	0,8700			
ID31	0,37	0,22	0,8713			
ID32	0,32	0,16	0,8724			*
ID33	0,64	0,37	0,8683			
ID34	0,49	0,39	0,8680			
ID35	0,62	0,48	0,8662			
ID36	0,58	0,47	0,8663			
ID37	0,50	0,42	0,8674			
ID38	0,39	0,26	0,8707			

Zdroj: vlastní zpracování s využitím CTT package (Willse, 2018)

- ***Interpretace***

Hodnocení kvality testových položek neukazuje na přítomnost velmi snadných testových položek, neboť obtížnost žádné testové položky není vyšší než 0,90, ani velmi obtížných testových položek, neboť obtížnost žádné testové položky není nižší než 0,10 (např. Krishnan, 2013). Dvě testové položky (ID27 a ID32) jsou charakteristické nízkou schopností diskriminace mezi žáky vzhledem k jejich statistické gramotnosti, neboť jejich hodnota diskriminace je nižší než 0,20 (např. Thompson, 2016). Vynechání těchto dvou testových položek by rovněž zvýšilo hodnotu spolehlivosti testu (Cronbachovo alfa). Celkově lze kvalitu testových položek hodnotit pozitivně.

4.1.2 Hodnocení kvality testových položek – distraktory

Hodnocení kvality testových položek testu statistické gramotnosti žáků ukázalo na existenci dvou testových položek s nízkou schopností diskriminace mezi žáky vzhledem k jejich úrovni statistické gramotnosti. Jeden z důvodů této skutečnosti může být nízká kvalita distraktorů, kdy některý z nich může působit na žáky s vyšší úrovní statistické gramotnosti matoucím způsobem. Tato modelová případová studie se proto zaměřuje na hodnocení kvality distraktorů dvou testových položek (ID27 a ID32, viz tabulka č. 14), jejichž vynechání z testu statistické gramotnosti rovněž zvyšuje spolehlivost celého testu.

- ***Řešení***

Hodnocení kvality distraktorů testových položek ID27 a ID32 je založeno na postupu, který je analogií k postupu předchozí modelové případové studie s tím, že počítána je pouze upravená bodově biseriální korelace pro distraktory testových položek. Specificky tak je správná odpověď na danou testovou položku postupně nahrazena třemi nesprávnými odpověďmi a s využitím této změny je spočítáno šest hodnot bodově biseriální korelace pro tři distraktory každé z testových položek. Tabulka č. 15 zachycuje výsledky tohoto postupu.

Tabulka č. 15: Upravená bodově biseriální korelace distraktorů testových položek ID27 a ID32 (*multi-choice* testová položka s výběrem ze čtyř možností)

Testová položka	Distraktor	Diskriminace (pBis)
ID27	ID27_1	-0,11
	ID27_2	-0,02
	ID27_3	0,10
	ID27_4 (správná odpověď)	0,14
ID32	ID32_1	-0,13
	ID32_2	-0,15
	ID32_3	0,17
	ID32_4 (správná odpověď)	0,16

- ***Interpretace***

Tabulka č. 15 ukazuje na podobné zdůvodnění nízké schopnosti obou testových položek (ID27 a ID32) diskriminovat mezi žáky podle úrovně jejich statistické gramotnosti. Takto první a druhý distraktor mají žádoucí vlastnost záporné hodnoty upravené bodově biseriální korelace. Naopak třetí distraktor má kladnou hodnotu upravené bodově biseriální korelace a může tak svou podobou z určitého důvodu motivovat žáky s vyšší úrovní statistické gramotnosti k výběru. Žádoucí je proto důkladná analýza podoby třetího a čtvrtého distraktoru obou testových položek.

4.1.3 Hodnocení kvality testových položek – DIF analýza

Kvalita testových položek může být nepříznivě ovlivněna jejich „nespravedlivým chováním" vzhledem k charakteristikám žáků, kdy jejich dvě různé skupiny, které mají stejnou úroveň statistické gramotnosti, odpovídají na danou testovou položku různě. V takovém případě nejsou odlišnosti v odpovědích na testovou položku spojeny s konstruktem statistické gramotnosti, nýbrž s charakteristikou žáků, která utváří nežádoucí šum pro interpretaci výsledků. Z uvedeného důvodu se tato modelová případová studie zaměřuje na identifikaci problémových testových položek vzhledem k jejich chování v rámci dvou skupin žáků: (a) dívky; a (b) chlapci.

- ***Řešení***

Z řady možností DIF analýzy, které vycházejí z CTT, jsou v této modelové případové studii využity dva metodické postupy.

Parametrický přístup je založen na logistické regresi a statistické významnosti odhadovaných parametrů modelu (zde LRT DIF statistika). Vedle statistické významnosti je hodnocena rovněž úroveň DIF s využitím srovnání hodnot *pseudo-*R^2 dvou regresních modelů: (a) modelu, který vysvětluje pravděpodobnost správné odpovědi na testovou položku úrovní statistické gramotnosti žáků; a (b) modelu, který vysvětluje pravděpodobnost správné odpovědi na testovou položku úrovní statistické gramotnosti žáků a příslušností žáka ke skupině dívek či chlapců, a to včetně interakčního členu. Hodnoty úrovně DIF nižší než 0,035 jsou považovány za nízké – kategorie A, hodnoty vyšší než 0,070 pak za vysoké – kategorie C (např. Lambert et al., 2018; Wiberg, 2007). Důležitost hodnoceni úrovně DIF je dána vysokou pravděpodobností statistické významnosti parametrů v případě velkých výběrových souborů žáků.

Neparametrický přístup je tzv. Mantel-Haenszelův (MH) přístup k DIF analýze, který umožňuje detekovat uniformní DIF v testových položkách s využitím $MH_{\chi 2}$ statistiky a hodnoty ΔMH. MH přístup klasifikuje testové položky do tří kategorií: (a) kategorie A s nízkou úrovní DIF (statisticky nevýznamné hodnoty $MH_{\chi 2}$ statistiky a absolutní hodnota ΔMH menší než 1); (b) kategorie B se střední úrovní DIF (statisticky významné hodnoty $MH_{\chi 2}$ statistiky a absolutní hodnota ΔMH menší než 1,5) a kategorie C s vysokou úrovní DIF (statisticky významné hodnoty $MH_{\chi 2}$ statistiky a absolutní hodnota ΔMH vyšší než 1,5 (např. Yavuz et al., 2018; Wiberg, 2007; Michaelides, 2008). Tabulka č. 16 zachycuje výsledky DIF analýzy 38 testových položek testu statistické gramotnosti pro dvě skupiny žáků – dívky a chlapce.

Tabulka č. 16: DIF analýza testových položek – dívky a chlapci

Testová položka	Logistická regrese			MH přístup		
	LRT statistika významnost	Úroveň DIF	Kategorie DIF	$MH_{\chi 2}$ významnost	ΔMH	Kategorie DIF
ID1	ne	0,000	A	ne	-0,06	A
ID2	ne	0,000	A	ne	-0,13	A
ID3	ne	0,001	A	ne	-0,15	A
ID4	ne	0,000	A	ne	0,08	A
ID5	ne	0,001	A	ne	-0,34	A
ID6	ne	0,000	A	ne	0,12	A
ID7	ne	0,001	A	ne	-0,06	A
ID8	ne	0,000	A	ne	-0,37	A
ID9	ano	0,004	A	ano	0,63	A
ID10	ne	0,001	A	ne	-0,43	A
ID11	ne	0,001	A	ne	-0,39	A
ID12	ne	0,000	A	ne	-0,09	A
ID13	ne	0,000	A	ne	0,16	A
ID14	ne	0,000	A	ne	-0,22	A
ID15	ne	0,000	A	ne	-0,09	A
ID16	ne	0,001	A	ne	0,00	A
ID17	ne	0,000	A	ne	0,13	A
ID18	ne	0,000	A	ne	-0,03	A
ID19	ne	0,001	A	ne	0,42	A
ID20	ne	0,001	A	ne	-0,24	A
ID21	ne	0,000	A	ne	0,13	A
ID22	ne	0,001	A	ne	0,06	A
ID23	ne	0,002	A	ne	0,37	A
ID24	ne	0,001	A	ne	0,30	A
ID25	ne	0,000	A	ne	-0,14	A
ID26	ne	0,000	A	ne	0,07	A
ID27	ne	0,000	A	ne	-0,22	A
ID28	ne	0,001	A	ne	-0,06	A
ID29	ano	0,002	A	ne	-0,30	A
ID30	ne	0,001	A	ne	0,35	A
ID31	ne	0,001	A	ne	0,11	A
ID32	ne	0,001	A	ne	0,15	A
ID33	ne	0,000	A	ne	-0,28	A
ID34	ne	0,000	A	ne	0,12	A
ID35	ne	0,001	A	ne	-0,28	A
ID36	ne	0,000	A	ne	0,04	A
ID37	ne	0,000	A	ne	0,05	A
ID38	ne	0,000	A	ne	0,07	A

Zdroj: vlastní zpracování s využitím difR package (Magis, Beland a Raiche, 2020)

- ***Interpretace***

Logistická regrese ani MH přístup neukázaly přítomnost DIF v některé z testových položek, tj. testové položky se chovají spravedlivě vůči dívkám i chlapcům. Nejsilněji je vliv hodnocené charakteristiky žáka pozorován v případě testové položky ID9, nicméně i v tomto případě se jedná o nízkou úroveň DIF (viz zařazení testové položky ID9 do kategorie A při využití parametrického i neparametrického metodického přístupu).

4.1.4 Hodnocení neobvyklého vzoru odpovědí žáka na testové položky

Pro odpovědi žáka na testové položky testu statistické gramotnosti by mělo platit, že jednodušší testové položky jsou zodpovídány spíše správně, zatímco obtížnější testové položky spíše nesprávně. Pokud se vzor odpovědí žáka od tohoto předpokladu odlišuje, vyvstává otázka po důvodech této skutečnosti, které mohou zahrnovat: (a) nízkou motivaci žáka, která vede k efektu hádání a rychlého vyplnění odpovědí na testové položky; (b) časový tlak na žáka v pozdější fázi řešení testových položek; (c) neetické chování žáka; či (d) prostou neznalost některých z testovaných oblastí. Identifikace žáků s neobvyklou strukturou odpovědí je proto prvním krokem pro další diskusi příčin tohoto jevu, a to včetně případného přijetí souvisejících opatření.

- ***Řešení***

Hodnocení neobvyklého vzoru odpovědí žáka na testové položky testu statistické gramotnosti, které vychází z CTT, je založeno na výpočtu statistik (indexů) charakterizujících soulad vzoru odpovědí daného žáka s odpověďmi žáků ostatních. Takto může být vzor odpovědí žáka na testové položky testu statistické gramotnosti posuzován především prostřednictvím následujících statistik (indexů), tj.:

- statistiky *r.pbis*, tj. bodově biseriální korelace mezi odpověďmi žáka na testové položky testu statistické gramotnosti a obtížností testových položek;
- statistik *C, C** a *U3*, které jsou založeny na srovnání vzoru odpovědí žáka se strukturou tzv. Guttmanova vzoru odpovědí, tj. s ideálním vzorem odpovědí žáka za předpokladu, že nejvíce obtížné testové položky žák zodpovídá chybně a naopak nejméně obtížné testové položky zodpovídá správně, přičemž počet správně zodpovězených testových položek obou vzorů odpovědí je shodný.

V případě statistiky r.pbis je nejvíc neobvyklá struktura odpovědí žáka spojena s vysokými zápornými hodnotami, v případě statistik *C, C** a *U3* je ideální vzor odpovědí žáků indikován hodnotou 0, rostoucí kladné hodnoty pak znamenají zvyšující se míru neobvyklosti struktury odpovědí žáka v testu statistické gramotnosti. Tabulka č. 17 zachycuje výsledky hodnocení neobvyklého vzoru odpovědí žáků na testové položky testu statistické gramotnosti.

Tabulka č. 17: Žáci s nejméně obvyklou strukturou odpovědí na testové položky (prvních 10 žáků podle statistiky *r.pbis*)

ID žáka	Hodnocená statistika							
	r.pbis	Pořadí	C	Pořadí	C*	Pořadí	U3	Pořadí
1199	-0,4416	1	1,56	2	0,83	2	0,83	4
2588	-0,4098	2	1,52	5	0,81	5	0,83	5
1805	-0,3993	3	1,55	3	0,83	3	0,85	1
1986	-0,3889	4	1,60	1	0,85	1	0,84	3
2351	-0,3766	5	1,54	4	0,82	4	0,84	2
2605	-0,3626	6	1,45	8	0,77	8	0,78	8
2900	-0,3557	7	1,46	7	0,79	6	0,79	6
1989	-0,3384	8	1,43	10	0,76	9	0,76	9
2367	-0,3153	9	1,43	9	0,77	7	0,79	7
2304	-0,2853	10	1,37	12	0,73	11	0,76	10

Zdroj: vlastní zpracování s využitím PerFit package (Tendeiro, 2018)

- ***Interpretace***

Interpretace tabulky č. 17 umožňuje identifikovat žáky (viz ID žáka) s neobvyklým vzorem odpovědí na testové položky testu statistické gramotnosti, přičemž analogická zjištění poskytují všechny čtyři hodnocené statistiky (indexy). Platí také, že uvedení žáci vyřešili správně spíše malý počet testových položek, což naznačuje pravděpodobný vliv faktoru hádání na výsledný vzor odpovědí identifikovaných žáků.

4.1.5 Hodnocení spolehlivosti testu

Podle jednoho ze základních vztahů CTT platí, že žákem dosažená úspěšnost v testu statistické gramotnosti (testové skóre) je součtem jeho skutečného skóre, tj. skutečného zvládnutí hodnoceného konstruktu (úroveň statistické gramotnosti) a chybového skóre daného podobou testu statistické gramotnosti. Právě kvůli existenci chybového skóre, tj. rozdílu mezi testovým a skutečným skóre žáka, je obecným zájmem hodnocení testů také ověření jejich spolehlivosti.

- ***Řešení***

CTT nabízí možnost využít řadu ukazatelů spolehlivosti testu statistické gramotnosti, které jsou počítány na bázi vzoru odpovědí žáků na testové položky. Tyto ukazatelé zahrnují:

- šest Guttmanových ukazatelů λ_1 až λ_6, které zahrnují i jeden z nejvíce používaných ukazatelů spolehlivosti Cronbachovo alfa (λ_3);
- ukazatele ω_h a ω_t, kde ukazatel ω_h odhaduje spolehlivost testu vzhledem k hlavnímu konstruktu (úroveň statistické gramotnosti) a ukazatel ω_t odhaduje celkovou spolehlivost testu, když bere do úvahy také vliv skupinových konstruktů dílčích testových položek.

Nízká hodnota ω_h ve srovnání s hodnotou ω_t naznačuje slabší vliv hlavního konstruktu testu, což také narušuje spolehlivost odhadu základních ukazatelů spolehlivosti testu. Revelle (2019) následně hovoří o tom, že ω_h je lepším ukazatelem spolehlivosti hlavního faktoru testu než nejčastěji používané Cronbachovo alfa. Tabulka č. 18 zachycuje hodnoty odhadů ukazatelů spolehlivosti testu statistické gramotnosti s tím, že pozitivně je potřeba hodnotit vysoké hodnoty těchto odhadů.

Tabulka č. 18: Hodnoty odhadů ukazatelů spolehlivosti testu

	Ukazatel spolehlivosti testu							
	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	ω_h	ω_t
Hodnota ukazatele	0,87	0,89	0,89	0,92	0,88	0,91	0,70* 0,75**	0,88

Pozn.: Hodnoty ω_h a ω_t stanoveny pro očekávaný počet 2*, respektive 3** skupinových faktorů.

Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

- ***Interpretace***

Hodnoty ukazatelů spolehlivosti testu statistické gramotnosti nabývají hodnot vyšších, než jsou referenční hodnoty 0,70 i 0,80 (např. Krishnan, 2013; Thompson, 2016). Výsledky tak naznačují dobrou spolehlivost celého testu. Srovnání hodnot ω_h a ω_t ovšem naznačuje slabší vliv dalších skupinových konstruktů dílčích testových položek, které snižují důvěryhodnost vypočtených hodnot ukazatelů spolehlivosti testu. Hodnota ω_h je nižší než hodnoty dalších ukazatelů spolehlivosti (λ_1 až λ_6).

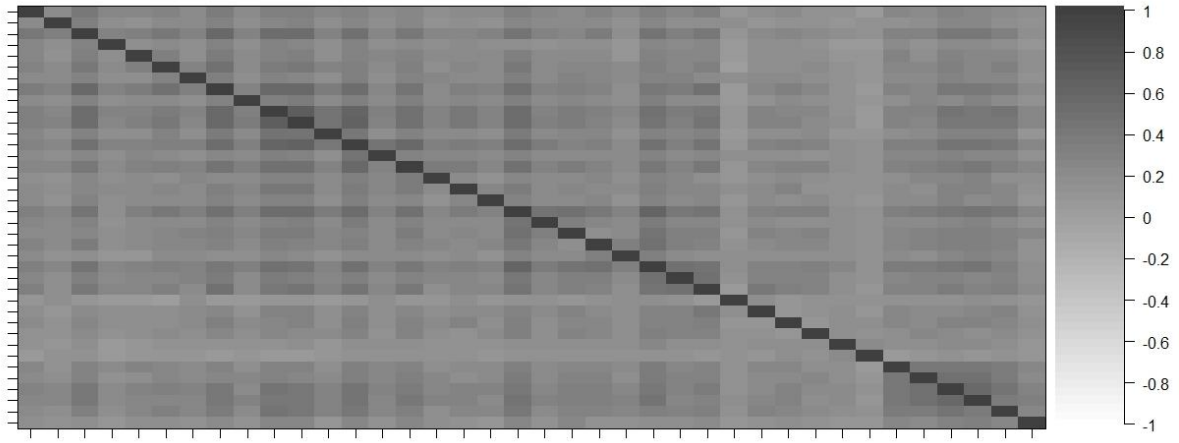
4.1.6 Hodnocení unidimenzionality testu

Řada metodických postupů CTT (ale i IRT) je založena na předpokladu unidimenzionality testu, který měří právě jeden hlavní konstrukt, tj. úroveň statistické gramotnosti žáků. Při splnění předpokladu unidimenzionality není kvalita vypočítaných statistik narušována přítomností jiných konstruktů, respektive přítomností vzájemně souvisejících testových položek. Z těchto důvodů je hodnocení unidimenzionality důležitou součástí hodnocení testů.

- ***Řešení***

Pro hodnocení naplnění předpokladu unidimenzionality testu statistické gramotnosti se nabízí několik metodických postupů. První využitý teoreticko-metodický postup poskytuje vstupní pohled na vztahy mezi dvojicemi testových položek testu statistické gramotnosti prostřednictvím výpočtu tetrachorických korelací odpovědí žáků na ně. V souladu s tradičním hodnocením úrovně korelace indikují vysoké hodnoty silnější vztah mezi testovými položkami, což působí proti předpokladu unidimenzionality testu. Obrázek č. 5 zachycuje hodnoty tetrachorických korelací mezi dvojicemi testových položek testu statistické gramotnosti.

Obrázek č. 5: Tetrachorické korelace testových položek



Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

Druhý využitý teoreticko-metodický přístup k hodnocení unidimenzionality testu statistické gramotnosti je založen na explorační faktorové analýze, a to s řešením základní otázky optimálního počtu konstruktů utvářených odpověďmi žáků na testové položky. Pro zodpovězení uvedené otázky lze sledovat několik metodických postupů.

Z hodnocení vlastních čísel (*eigenvalues*) faktorů extrahovaných z testu statistické gramotnosti žáků vycházejí tři metodické postupy, kdy: (a) Kaiserovo kritérium indikuje optimální počet konstruktů jako počet vlastních čísel faktorů vyšších než 1; (b) sutinový graf (*scree plot*) určuje optimální počet konstruktů ve vazbě na umístění významného zlomu v křivce vlastních čísel faktorů; a (c) metoda paralelní analýzy srovnává hodnoty vlastních čísel faktorů s hodnotami vlastních čísel faktorů extrahovaných ze simulovaných dat, přičemž hlavní myšlenka této metody tvrdí, že pro test statistické gramotnosti jsou smysluplné ty faktory (konstrukty), jejichž vlastní číslo je vyšší než korespondující hodnota vlastního čísla extrahovaného z faktorů simulovaných dat.

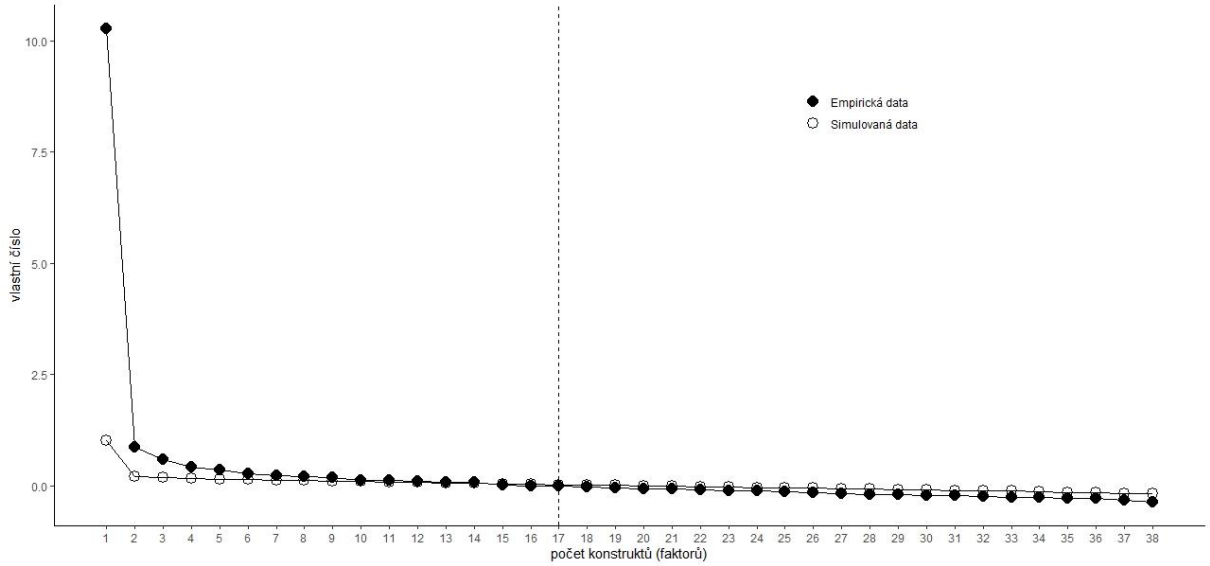
Tabulka č. 19 zachycuje hodnoty vlastních čísel pro prvních pět konstruktů (faktorů) faktorové analýzy odpovědí žáků na testové položky testu statistické gramotnosti založené na metodě hlavních os a maximální věrohodnosti. Obrázek č. 6 pak znázorňuje sutinový graf vlastních čísel v kontextu srovnání empirických a simulovaných dat paralelní analýzy s tím, že tento postup doporučuje jako optimální počet 14 konstruktů (faktorů).

Tabulka č. 19: Hodnoty vlastních čísel pro osm konstruktů (faktorů) faktorové analýzy

Počet konstruktů (faktorů)	1	2	3	4	5
Vlastní číslo	10,27	0,86	0,59	0,42	0,37

Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

Obrázek č. 6: Sutinový graf vlastních čísel konstruktů (faktorů) faktorové analýzy



Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

Metoda velmi jednoduché struktury (dále i „VSS“) stanovuje optimální počet konstruktů (faktorů) srovnáním korelační matice odpovědí žáků na testové položky testu statistické gramotnosti se zjednodušenou maticí, která zohledňuje pouze nejvyšší faktorové zátěže každé testové položky (VSS komplexity 1), a to pro výzkumníkem stanovený maximální počet konstruktů (faktorů). Nejvyšší hodnota VSS kritéria indikuje optimální počet faktorů. Ukazatel VSS komplexity 2 pracuje s více komplexní faktorovou strukturou odpovědí žáků na testové položky testu statistické gramotnosti, protože zohledňuje nejen nejvyšší faktorové zátěže každé testové položky, ale také druhé nejvyšší faktorové zátěže. I v tomto případě je optimální počet faktorů vybrán prostřednictvím nejvyšší hodnoty VSS kritéria.

Tabulka č. 20 zachycuje hodnoty kritérií VSS komplexity 1 a VSS komplexity 2 pro různý počet konstruktů (faktorů) faktorové analýzy odpovědí žáků na testové položky testu statistické gramotnosti. Kritérium VSS komplexity 1 indikuje optimální počet jednoho konstruktu (faktoru), kritérium VSS komplexity 2 indikuje optimální počet dvou konstruktů (faktorů).

Tabulka č. 20: Hodnoty VSS kritérií podle počtu konstruktů (faktorů) faktorové analýzy

VSS kritérium	Počet konstruktů (faktorů)					
	1	2	3	4	5	6
VSS1	0,84	0,47	0,41	0,28	0,27	0,23
VSS2	0,00	0,85	0,76	0,62	0,57	0,49

Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

Velicerův MAP test je založen na hodnocení série matic korelací odpovědí žáků na testové položky testu statistické gramotnosti. V nultém kroku je počítána průměrná hodnota korelací

mimo hlavní diagonálu (MAP). V prvním kroku je odstraněn vliv prvního konstruktu a analogicky je počítána průměrná hodnota korelací (MAP) mimo hlavní diagonálu. Tento postup je pak opakován pro *k-1* kroků, kde *k* je celkový počet testových položek testu statistické gramotnosti. Nejnižší hodnota MAP představuje optimální počet konstruktů (faktorů).

Tabulka č. 21 zachycuje hodnoty Velicerova MAP testu pro různý počet konstruktů (faktorů) faktorové analýzy odpovědí žáků na testové položky testu statistické gramotnosti. Velicerovo MAP indikuje optimální počet dvou konstruktů (faktorů).

Tabulka č. 21: Hodnoty Velicerova MAP podle počtu konstruktů (faktorů)

Počet konstruktů (faktorů)	1	2	3	4	5	6
Velicerovo MAP	0,0049	0,0047	0,0050	0,0057	0,0063	0,0073

Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

Konečně DETECT index hodnotí úroveň multidimenzionality odpovědí žáků na testové položky testu statistické gramotnosti, přičemž předpokládá, že každá testová položka měří jednak hlavní konstrukt (statistická gramotnost), jednak reziduální část, která je nekorelovaná s hlavním konstruktem. Výpočet DETECT indexu je pak spojený s hledáním optimální podoby rozdělení testových položek do skupin (klastrů), kdy každá skupina odpovídá jinému konstruktu, ať již tento má význam či nikoliv. Podle DETECT indexu je pak úroveň multidimenzionality testu hodnocena takto (např. Zhang, 2013; Bonifay et al., 2015):

- Hodnoty nižší než 0,1(0,2) naznačují unidimenzionalitu testu.
- Hodnoty mezi 0,1-0,5 (0,2-0,4) naznačují slabou multidimenzionalitu testu.
- Hodnoty mezi 0,5-1,0 (0,4-1,0) naznačují středně silnou multidimenzionalitu testu.
- Hodnoty vyšší než 1,0 naznačují silnou multidimenzionalitu testu.

Tabulka č. 22 zachycuje hodnoty DETECT indexu pro různé počty skupin (klastrů) testových položek. Nejvyšší hodnoty dosahuje DETECT index pro rozdělení testových položek do 10 skupin (klastrů), tj. hodnoty 0,519.

Tabulka č. 22: Hodnoty DETECT indexu podle počtu skupin (klastrů) testových položek

	Počet skupin (klastrů) testových položek							
	2	3	4	5	6	7	8	9
DETECT index	-0,065	0,264	0,423	0,461	0,478	0,487	0,504	0,510
	10	11	12	13	14	15	16	17
DETECT index	0,519	0,518	0,499	0,503	0,499	0,501	0,506	0,506

Zdroj: vlastní zpracování s využitím sirt package (Robitzsch, 2020)

- ***Interpretace***

Využité metodické postupy pro hodnocení unidimenzionality v odpovědích žáků na testové položky testu statistické gramotnosti přinášejí následující zjištění:

- Vysoká hodnota vlastního čísla prvního konstruktu (viz tabulka č. 19), která je rovněž spojena s prvním významný zlomem v sutinovém grafu (viz obrázek č. 6), naznačuje dominantní vliv hlavního konstruktu, tj. statistické gramotnosti žáků, a to v souladu s předpokladem unidimenzionality zadaného testu. Podporu tomuto zjištění poskytuje také optimální počet jednoho faktoru při využití kritéria VSS komplexity 1 (viz tabulka č. 20).
- Vedle dominantního hlavního faktoru naznačuje Velicerovo MAP přítomnost jednoho vedlejšího konstruktu (faktoru) a rovněž hodnota DETECT indexu je na hranici slabé a středně silné multidimenzionality testu (viz tabulka č. 21 a č. 22). Tetrachorické korelace dvojic testových položek ukazují v tomto ohledu přítomnost několika shluků testových položek s o něco vyšší úrovní korelace mezi nimi (viz obrázek č. 5).

Tabulka č. 23 ukazuje optimální počet konstruktů (faktorů) v odpovědích žáků na testové položky testu statistické gramotnosti při aplikaci různých metodických přístupů. Takto se předpoklad unidimenzionality testu statistické gramotnosti jeví jako naplněný s tím, že pozornost lze věnovat také dvou-faktorovému řešení ve vazbě na hodnocení silnějších asociací mezi testovými položkami (hodnocení lokální nezávislosti testových položek).

Tabulka č. 23: Optimální počet konstruktů (faktorů) – různé metodické přístupy

	Kaiserovo kritérium	Sutinový graf	Paralelní analýza	VSS1	VSS2	MAP	DETECT index
Počet konstruktů (faktorů)	1	1	14*	1	2	2	0,519

* Vysoký počet navrhovaných konstruktů (faktorů) metodou paralelní analýzy je spojený s citlivostí této metody k velikosti hodnoceného výběrového souboru žáků, která se projevuje v nízkých hodnotách vlastních čísel pro faktory simulovaných dat. Důsledkem pak je návrh příliš vysokého optimálního počtu 14 konstruktů (faktorů).

4.1.7 Hodnocení konstruktů obsažených v testu

V modelové případové studii představené v podkapitole 4.1.6 bylo doporučeno věnovat pozornost hodnocení vztahů testových položek při předpokladu přítomnosti dvou konstruktů v testu statistické gramotnosti žáků. Takové hodnocení úkol lze rovněž dát do souvislostí se dvěma dalšími záměry hodnocení testů:

(a) ověření existence očekávaných dílčích konstruktů (subdimenzí) testu, pokud tyto byly definovány dopředu;

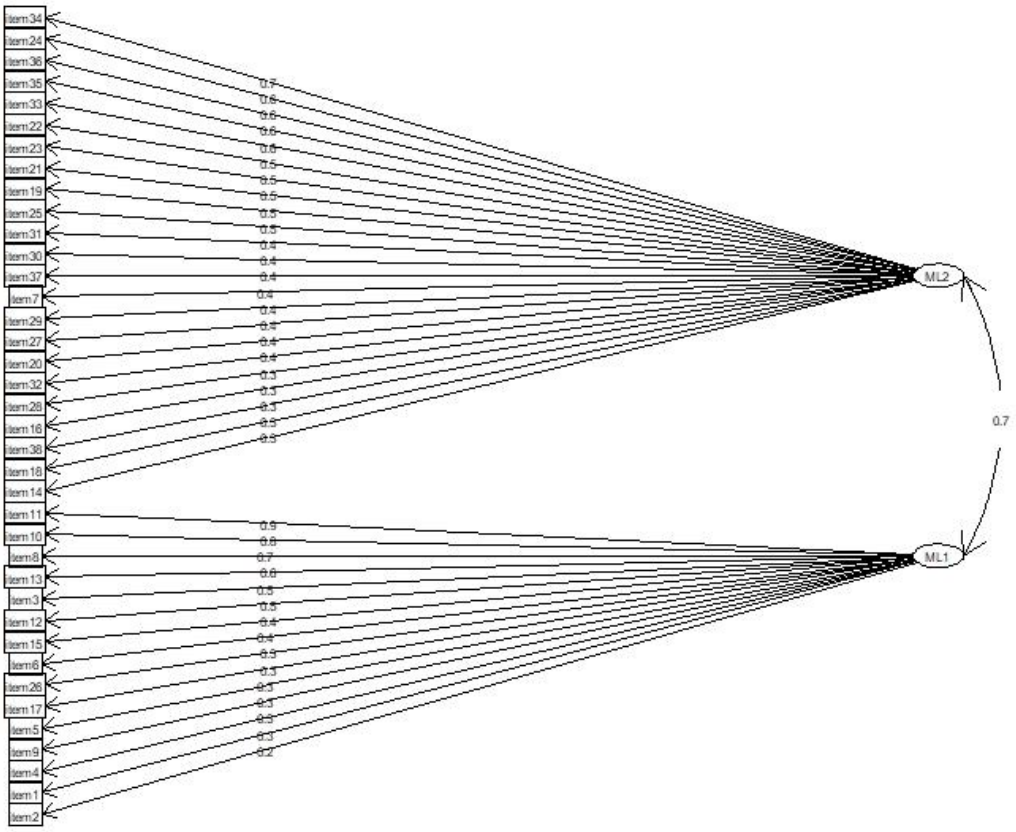
(b) identifikace neznámých dílčích konstruktů (subdimenzí) testu a hodnocení výsledků žáků při řešení testových položek utvářejících tyto konstrukty.

Do tohoto kontextu je zasazena tato modelová případová studie.

- ***Řešení***

Pro hodnocení konstruktů obsažených v odpovědích žáků na testové položky testu statistické gramotnosti byl využit postup hledající ty testové položky, které jsou s těmito konstrukty nejsilněji asociovány. V návaznosti na modelovou případovou studii představenou v podkapitole 4.1.6 byla řešena situace se dvěma konstrukty, k nimž byly odhadovány faktorové zátěže testových položek.[54] Obrázek č. 7 znázorňuje řešení.

Obrázek č. 7: Vztahy testových položek, řešení se dvěma konstrukty (faktory)



Pozn.: V tabulce jsou uvedeny faktorové zátěže vyšší než 0,20.

Zdroj: vlastní zpracování s využitím psych package (Revelle, 2020)

- ***Interpretace***

Obrázek č. 7 ukazuje příslušnost dílčích testových položek ke dvěma konstruktům dvoufaktorového řešení faktorové analýzy testu statistické gramotnosti. Detailnější analýza utvářených konstruktů může vést k jejich vhodné interpretace. Zároveň je potřeba upozornit na silnou vazbu mezi oběma konstrukty, což opodstatňuje rovněž vhodnost řešení s jedním hlavním faktorem.

[54] Konkrétně byla využita metoda hlavních os a maximální věrohodnosti s tetrachorickými korelacemi.

4.1.8 Propojení dosaženého skóre žáků na společnou škálu

Vzorový test statistické gramotnosti byl zadán ve dvou verzích tvořených 38 testovými položkami s odpověďmi zajištěnými od tří tisíc žáků. Obě verze testů přitom mají 20 testových položek společných. V tomto kontextu vzniká otázka, jakým způsobem lze výsledek žáků v obou verzích testů vyjádřit na stejné škále tak, aby ani jedna skupina žáků nebyla znevýhodněna odlišnou obtížnosti testu. Záměrem modelové případové studie je tedy propojit dosažené skóre žáků na společnou škálu.

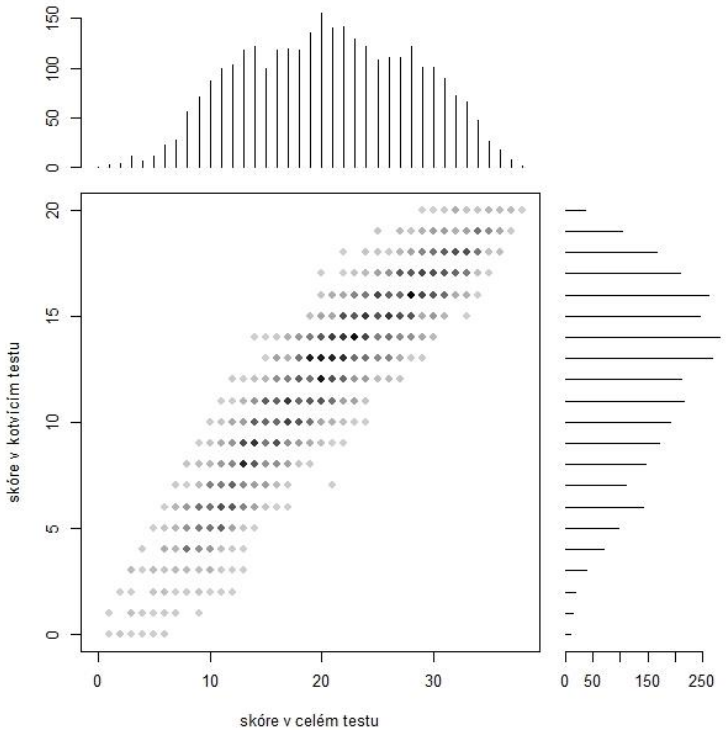
- ***Řešení***

Pro řešení modelové případové studie je potřeba propojit škály obou testů (*equating*) s tím, že za tímto účelem je vhodné aplikovat přístup založený na dvou neekvivalentních skupinách žáků, které společně řeší kotvící test tvořený 20 testovými položkami (NEAT přístup). Pro propojení obou skóre testů je potřeba mít pro obě skupiny testovaných žáků informace o:

- dosaženém skóre žáků v řešení kotvícího testu statistické gramotnosti;
- dosaženém skóre žáků v řešení celého testu statistické gramotnosti.

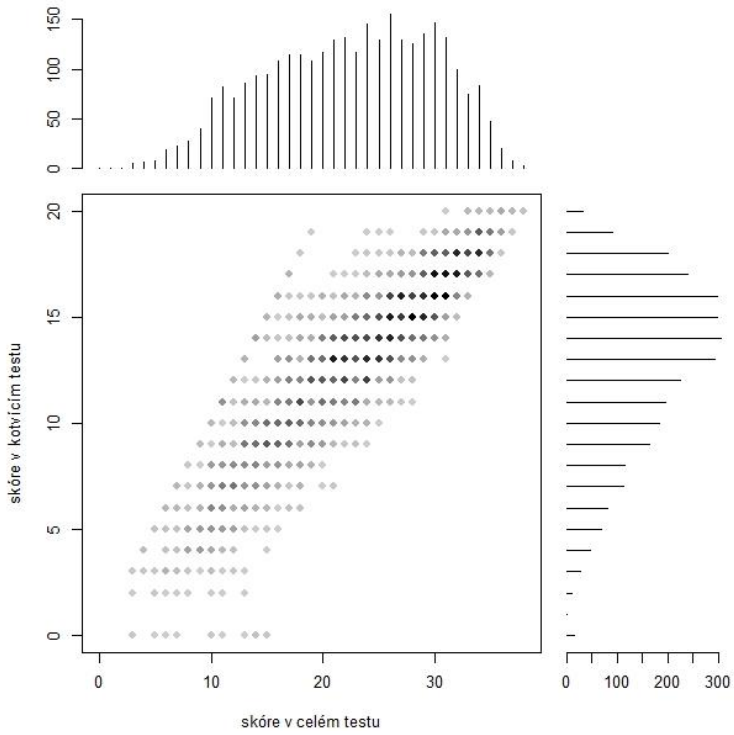
Tyto informace jsou následně využity pro výpočet četností výskytu dvojic hodnot dosaženého skóre žáků v řešení jednak kotvícího a jednak celého testu statistické gramotnosti, přičemž tento výpočet je opětovně proveden pro obě skupiny žáků (viz obrázky č. 8 a č. 9).

Obrázek č. 8: Rozdělení četností výskytu skóre žáků v první verzi testu statistické gramotnosti – celý test, kotvící test a vazby mezi oběma typy testů



Zdroj: vlastní zpracování s využitím equate package (Albano, 2018)

Obrázek č. 9: Rozdělení četností výskytu skóre žáků ve druhé verzi testu statistické gramotnosti – celý test, kotvící test a vazby mezi oběma typy testů



Zdroj: vlastní zpracování s využitím equate package (Albano, 2018)

Z rozdělení četností výsledků celého i kotvícího testu tak, jak je zachycují obrázky č. 8 a č. 9, je zjevné, že první verze testu je obtížnější než verze druhá. Z tohoto důvodu je propojení škál obou testů žádoucí. Za tímto účelem je využit ekvipercentilní přístup, který je založen na percentilovém umístění skóre žáků v kotvícím a celém testu pro jejich propojení. Předpokládáme stejnou velikost populací obou výběrových souborů řešících první, respektive druhý test. Výsledkem, který tímto způsobem dostáváme, jsou korespondující skóre obou verzí testu statistické gramotnosti (viz tabulka č. 24).

Tabulka č. 24: Korespondující skóre prvního a druhého testu statistické gramotnosti (vybraná skóre)

Test	Korespondující skóre									
Test 1	11	12	13	14	15	16	17	18	19	20
Test 2	11,1	12,2	13,3	14,5	15,6	16,5	17,5	18,5	19,7	20,9
Test	**Korespondující skóre**									
Test 1	21	22	23	24	25	26	27	28	29	30
Test 2	22,1	23,3	24,4	25,5	26,3	27,2	28,2	29,2	30,1	31,0

Zdroj: vlastní zpracování s využitím equate package (Albano, 2018)

- ***Interpretace***

Řešení modelové případové studie umožňuje reportovat dosažené skóre žáků řešících odlišné verze testu statistické gramotnosti na stejné škále, a to s využitím převodní tabulky korespondujících skóre (viz tabulka č. 24). Z té je dobře patrné zohlednění vyšší obtížnosti první verze testu statistické gramotnosti, neboť skóre žáků řešících tuto verzi testu je po propojení obou škál vyšší. Tato skutečnost se projevuje také v hodnocení průměrného skóre žáků, kdy rozdíl průměrného skóre žáků v řešení obou verzí testu statistické gramotnosti se snižuje po propojení obou škál (viz tabulka č. 25).

Tabulka č. 25: Průměrné skóre žáků v různých verzích testu statistické gramotnosti

Skupina žáků	Kotvící test	Celý test bez propojení škál	Celý test na společné škále
Žáci řešící první verzi testu	12,2	20,6	21,4
Žáci řešící druhou verzi testu	12,8	22,4	22,4

4.2 Modelové případové studie řešené s využitím teorie odpovědi na položku

Dále uvedené modelové případové studie opětovně vycházejí z hodnocení vzorového testu statistické gramotnosti tvořeného 38 dichotomickými testovými položkami s odpověďmi tří tisíc žáků.

4.2.1 Stanovení výsledku žáka na škále vycházející z IRT

V modelových případových studiích řešených přístupy, které vycházejí z CTT, byl výsledek žáků v testu statistické gramotnosti stanoven jednak v podobě dosaženého skóre žáka (počet správně zodpovězených testových položek), jednak v podobě dosažené úspěšnosti (podíl správně zodpovězených testových položek). Přístupy, které vycházejí z IRT, nabízejí další možnosti, jak výsledek žáků v testu statistické gramotnosti vyjádřit, a to v podobě úrovně statistické gramotnosti měřené jako latentní konstrukt Θ. Záměrem této modelové případové studie je proto stanovit výsledek žáka v testu statistické gramotnosti na třech škálách vycházejících z IRT, které odpovídají: a) 1PL modelu; b) 2PL modelu; a c) 3PL modelu. Výsledky jsou následně mezi sebou porovnány.

- ***Řešení***

Stanovení výsledku žáků v testu statistické gramotnosti na škálách vycházejících z IRT je založeno na dvou základních krocích postupu:

- V prvním kroku jsou odhadovány parametry testových položek testu statistické gramotnosti v závislosti na vybraném modelu. V případě 1PL modelu tak je odhadována pouze obtížnost testových položek, v případě 2PL modelu obtížnost a diskriminace testových položek, v případě 3PL modelu obtížnost, diskriminace a parametr „pseudo-hádání“.

- Ve druhém kroku je odhadována úroveň statistické gramotnosti žáků, tj. hodnota Θ, a to v návaznosti na jejich vzor odpovědí na testové položky (viz tabulka č. 26 pro příklad tohoto vztahu). V modelové případové studii jsou využity dva přístupy odhadu statistické gramotnosti žáků: (a) MAP přístup s hledáním maximální věrohodnosti odhadu hodnoty Θ z aposteriorního rozdělení s apriorním předpokladem normálního rozdělení úrovně statistické gramotnosti žáků; a (b) EAP přístup hledající střední (očekávanou) hodnotu Θ se zohledněním různých vah úrovní statistické gramotnosti žáků s předpokladem jejich apriorního normálního rozdělení.

Při odhadu úrovně statistické gramotnosti žáků je potřeba rovněž rozhodnout o metodě odhadu parametrů modelů. V modelové případové studii je sledována metoda marginální maximální logaritmické věrohodnosti s využitím hybridního EM – Newton-Raphsonova algoritmu.

Tabulka č. 26: Příklady vztahů vzoru odpovědí žáků na testové položky testu statistické gramotnosti a dosažené úrovně statistické gramotnosti; 2PL model (EAP)

Vzor odpovědí žáků na testové položky („1“ – správná odpověď; „0“ – nesprávná odpověď)	Úroveň statistické gramotnosti žáků (Θ)
1,0,0,0,1,1,1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,0,1,1,1,0,0,1,0,0,1,1,1,1,1,0,1	0,607
1,0,1,0,1,1,0,1,1,1,1,1,1,1,1,1,0,0,1,1,0,1,1,1,0,1,0,1,0,0,1,0,1,1,0,0,1,0	0,394
1,0,0,1,0,0,0,1,1,1,1,1,1,1,0,1,1,0,0,1,1,0,0,1,1,0,1,0,0,0,1,0,1,1,0,0,1,0,1	-0,106
1,0,0,1,0,0,1,1,1,0,1,0,1,0,1,0,1,1,0,1,0,1,1,1,1,1,1,0,0,1,0,0,0,1,0,1,0,0,0	-0,310
1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,0,0,1,1,1,0,0,0,1,1,1,0,0,1,0,0,1,0,0,1,1,0,1	0,502
0,0,0,0,0,1,1,1,1,1,0,0,1,0,0,0,0,0,0,0,0,0,1,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,0	-1,246
1,0,1,1,1,1,0,1,0,1,1,1,0,0,0,1,1,0,1,1,1,1,0,1,0,1,0,0,0,0,0,1,0,1,1,1,1,1	0,163
1,0,1,1,1,1,1,1,0,1,1,0,0,0,1,0,0,1,0,1,0,0,0,0,0,0,0,0,0,0,0,0,1,0,1,1,1,0,1	-0,558
……	……..

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

Tabulka č. 27: Odhady úrovně statistické gramotnosti žáků s využitím různých metodických přístupů; hodnoty pro vybrané žáky

ID	Skóre	1PL model (MAP)	1PL model (EAP)	2PL model (MAP)	2PL model (EAP)	3PL model (MAP)	3PL model (EAP)
1	26	0,590	0,614	0,562	0,607	0,653	0,664
2	24	0,355	0,365	0,365	0,394	0,386	0,401
3	20	-0,094	-0,080	-0,159	-0,106	-0,058	-0,072
4	19	-0,204	-0,189	-0,326	-0,310	-0,241	-0,262
5	20	-0,094	-0,080	-0,182	-0,130	-0,100	-0,079
---	---	---	---	---	---	---	---

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

Tabulka č. 27 pak uvádí odhady úrovně statistické gramotnosti žáků s využitím různých metodických přístupů, a to včetně hodnoty dosaženého skóre jako počtu správných odpovědí na testové položky.

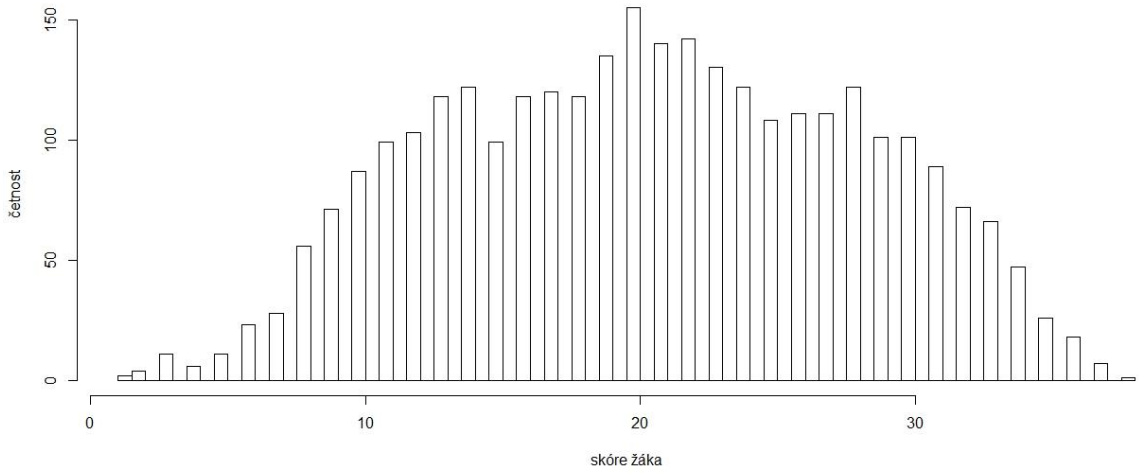
- ***Interpretace***

Hodnoty úrovně statistické gramotnosti žáků, které jsou zachyceny v tabulce č. 27, umožňují prezentovat některá specifika využitých metodických přístupů k jejich odhadu. Odhad úrovně statistické gramotnosti žáků prostřednictvím 1PL modelu ukazuje přímou vazbu mezi dosaženým skóre žáka a úrovní statistické gramotnosti odhadované 1PL modelem. Takto je každé hodnotě skóre přiřazena právě jedna hodnota úrovně statistické gramotnosti žáků odhadované tímto modelem. Tato skutečnost je způsobena stejnou hodnotou parametru diskriminace všech testových položek.

V případě 2PL modelu není předpoklad stejné hodnoty parametru diskriminace všech testových položek zachován. Důsledkem této skutečnosti jsou odlišné hodnoty úrovně statistické gramotnosti žáků, kteří dosáhli stejného celkového skóre v testu, nicméně jejich vzor odpovědí na testové položky se liší (viz žáci ID3 a ID5). Zjevnou nevýhodou odhadu úrovně statistické gramotnosti žáků prostřednictvím 2PL modelu je nižší stupeň intuitivního porozumění reportovaných hodnot, naopak výhodou tohoto přístupu je vyšší diferenciace mezi žáky, kdy proměnná získává více spojitý charakter.

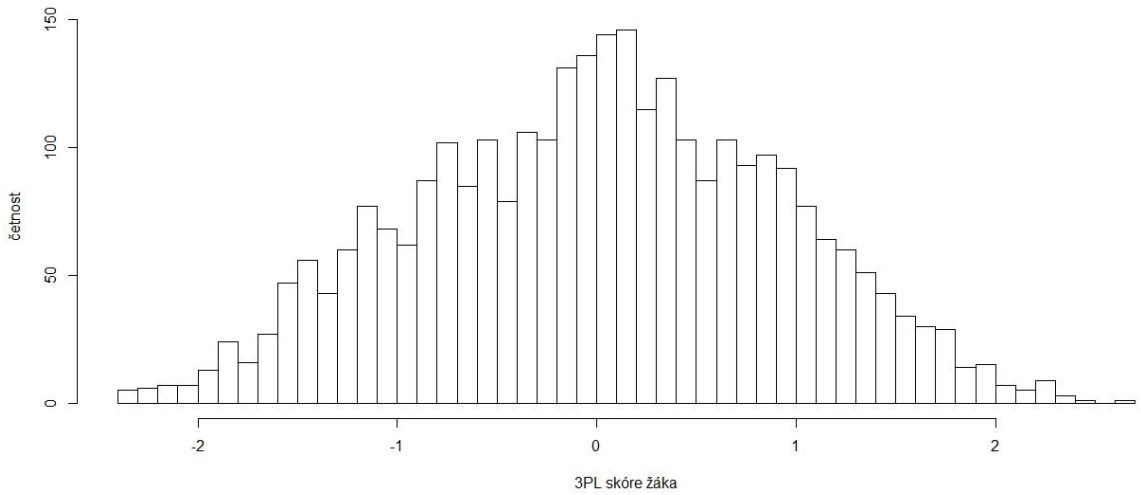
Analogické poznámky jako v případě 2PL modelu lze formulovat také pro odhad úrovně statistické gramotnosti žáků prostřednictvím 3PL modelu. Tento bere navíc do úvahy odlišný parametr „pseudo-hádání“ jednotlivých testových položek. Volba podoby metriky statistické gramotnosti žáků by měla vzít uvedené skutečnosti do úvahy.

Obrázek č. 10: Histogram hodnot úrovně statistické gramotnosti žáků (skóre žáka)



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

Obrázek č. 11: Histogram hodnot úrovně statistické gramotnosti žáků (3PL model, MAP)



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

Obrázky č. 10 a č. 11 zachycují prostřednictvím histogramů charakteristiky dvou způsobů hodnocení úrovně statistické gramotnosti žáků – dosažené skóre žáka (počet správných odpovědí) a 3PL model (MAP). Při srovnání těchto histogramů je dobře patrný rozdíl ve spojitosti hodnot 3PL modelu ve srovnání s diskrétními hodnotami dosaženého skóre žáků. Takto tako vzniká potenciál 3PL modelu lépe aproximovat normální rozdělení hodnot úrovně statistické gramotnosti žáků díky svému spojitému charakteru. Poukažme také na hodnoty blížící se 1 korelací mezi proměnnými charakterizujícími různé způsoby odhadů statistické gramotnosti žáků (viz tabulka č. 28).

Tabulka č. 28: Korelace mezi proměnnými charakterizujícímu různé způsoby odhadů úrovně statistické gramotnosti žáků

	Skóre	1PL (EAP)	1PL (MAP)	2PL (EAP)	2PL (MAP)	3PL (EAP)	3PL (MAP)
Skóre	1,000	1,000	1,000	0,991	0,9991	0,987	0,986
1PL (EAP)		1,000	1,000	0,991	0,991	0,987	0,986
1PL (MAP)			1,000	0,991	0,991	0,987	0,986
2PL (EAP)				1,000	1,000	0,998	0,998
2PL (MAP)					1,000	0,998	0,998
3PL (EAP)						1,000	1,000
3PL (MAP)							1,000

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

4.2.2 Převedení výsledku žáka na alternativní bodovou škálu

V předchozí modelové případové studii byl představen postup pro odhad úrovně statistické gramotnosti žáků s využitím různých škál vycházejících z IRT. Běžnou praxí při hodnocení testů však je reporting výsledků nikoliv na IRT škále, ale na alternativní a lépe srozumitelné škále (např. počet dosažených bodů, percentilová škála). V této modelové případové studii je našim zájmem převést hodnoty úrovně statistické gramotnosti žáků 3PL modelu (MAP) na škálu se středem 500 bodů a směrodatnou odchylkou 100 bodů.

- ***Řešení***

Řešení modelové případové studie je primárně založeno na aplikaci jednoduché transformace tak, aby průměr hodnot dosažené úrovně statistické gramotnosti žáků odhadované 3PL modelem (MAP) odpovídal hodnotě 500 bodů a směrodatná odchylka těchto hodnot hodnotě 100 bodů. Tabulka č. 29 zachycuje výsledek takové transformace. Současně je stanoveno percentilové pořadí dosaženého výsledku na škále odhadované 3PL modelem (MAP).

Tabulka č. 29: Transformace hodnot úrovně statistické gramotnosti žáků odhadované 3PL modelem (MAP) na bodovou škálu s průměrnou hodnotou 500 bodů a směrodatnou odchylkou 100 bodů

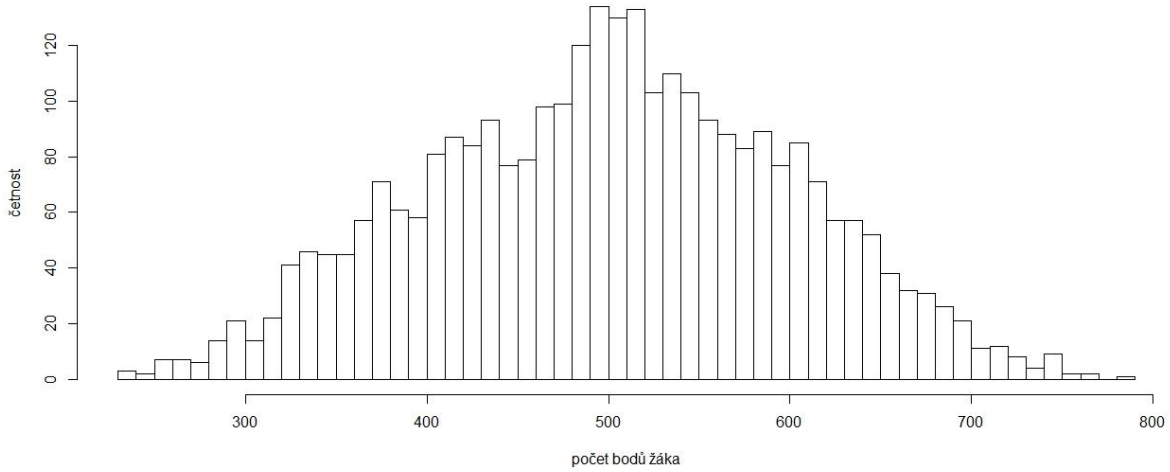
ID	3PL model (MAP)	Bodová škála	Percentilová škála
1	0,653	570	0,745
2	0,386	541	0,654
3	-0,058	492	0,455
4	-0,241	472	0,381
5	-0,100	547	0,673
---	---------	---------	---------

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018), respektive CTT package (Willse, 2018)

- ***Interpretace***

Škála 3PL modelu a bodová škála uvedené v tabulce č. 29 jsou ekvivalentní, jejich rozdílnost spočívá ve velikosti střední hodnoty (průměru) a směrodatné odchylky. Percentilová škála představuje percentilové pořadí výsledku daného žáka. Obrázek č. 12 zachycuje rozdělení hodnot dosažené úrovně statistické gramotnosti žáků pro bodovou škálu.

Obrázek č. 12: Histogram hodnot úrovně statistické gramotnosti žáků – 3PL model (MAP), škála s průměrem 500 bodů a směrodatnou odchylkou 100 bodů



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018), respektive CTT package (Willse, 2018)

4.2.3 Hodnocení kvality testových položek

Hodnocení kvality testových položek testu statistické gramotnosti, které vychází z CTT, bylo představeno ve třetí kapitole. Tato modelová případová studie má stejný cíl, nicméně využívá odlišné metodické přístupy založené na IRT, přičemž pro hodnocení kvality testových položek testu statistické gramotnosti sleduje 2PL a 3PL model. Volba modelů je dána jednak šíři parametrů, které oba modely odhadují a jednak možností ilustrace rozdílů mezi oběma modely.

- ***Řešení***

Podoba hodnocení kvality testových položek testu statistické gramotnosti je primárně ovlivněna odlišnostmi v množině parametrů, které 2PL a 3PL modely odhadují. V případě 2PL modelu jsou odhadovány:

- parametr obtížnosti testových položek testu statistické gramotnosti;
- parametr diskriminace testových položek testu statistické gramotnosti.

V případě 3PL modelu jsou odhadovány:

- parametr obtížnosti testových položek testu statistické gramotnosti;
- parametr diskriminace testových položek testu statistické gramotnosti;
- parametr „pseudo-hádání“ testových položek testu statistické gramotnosti.

Tabulka č. 30 zachycuje odhady parametrů testových položek testu statistické gramotnosti založené na 2PL a 3PL modelu.

Tabulka č. 30: Odhady parametrů testových položek testu statistické gramotnosti

Testová položka	2PL model		3PL model		
	Obtížnost	Diskriminace	Obtížnost	Diskriminace	Pseudo-hádání
ID1	-1,11	0,95	-1,13	0,92	0,001
ID2	-0,13	0,65	0,32	0,78	0,135
ID3	-1,17	1,67	-1,21	1,58	0,000
ID4	-0,35	0,62	-0,35	0,63	0,000
ID5	-0,72	0,82	-0,73	0,81	0,001
ID6	-1,07	1,09	-1,10	1,05	0,000
ID7	0,33	0,82	0,70	1,11	0,143
ID8	-1,29	1,99	-1,35	1,86	0,000
ID9	-0,21	0,83	0,21	1,02	0,152
ID10	-1,01	2,10	-0,90	2,16	0,094
ID11	-1,11	2,20	-1,07	2,16	0,056
ID12	-0,24	0,87	-0,24	0,88	0,000
ID13	-0,91	2,09	-0,78	2,17	0,100
ID14	0,27	0,82	0,55	1,02	0,107
ID15	-0,49	1,40	-0,25	1,61	0,121
ID16	0,28	0,76	0,86	1,32	0,220
ID17	0,45	0,84	0,44	0,86	0,000
ID18	0,14	0,77	0,54	0,98	0,139
ID19	-0,52	1,84	-0,18	2,55	0,181
ID20	-0,02	0,89	0,55	1,41	0,220
ID21	-0,22	1,04	0,34	1,67	0,234
ID22	0,33	1,01	0,72	1,79	0,185
ID23	0,90	0,72	1,19	2,20	0,225
ID24	-0,21	1,67	0,12	2,53	0,173
ID25	0,03	1,07	0,52	1,84	0,214
ID26	-0,97	1,34	-0,88	1,34	0,065
ID27	4,03	0,44	2,27	1,85	0,111
ID28	0,62	0,75	0,90	1,01	0,117
ID29	0,29	0,80	0,83	1,40	0,214
ID30	0,58	0,71	1,06	1,27	0,199
ID31	1,11	0,51	1,57	1,29	0,239
ID32	2,01	0,40	1,74	2,33	0,259
ID33	-0,71	1,01	0,26	2,12	0,382
ID34	0,06	1,01	0,62	2,19	0,253
ID35	-0,50	1,43	0,06	2,47	0,273
ID36	-0,34	1,36	0,17	2,31	0,239
ID37	0,01	1,09	0,30	1,42	0,126
ID38	0,79	0,59	1,08	0,75	0,101

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

- ***Interpretace***

Pro hodnocení obtížnosti testových položek platí, že typické hodnoty se pohybují v intervalu od -2 do +2. Hodnoty mimo interval od -3 do +3 lze považovat za velmi snadné či velmi těžké (např. Toland, 2014; DeMars, 2010). Tabulka č. 30 ukazuje, že za problematickou lze považovat především testovou položku ID27, která se ukazuje být relativně obtížná (viz také srovnání obtížnosti této testové položky s testovými položkami ostatními s využitím přístupu CTT). Za pozornost rovněž stojí vliv zohlednění parametru „pseudo-hádání“, tj. odhad 3PL modelu, na změnu obtížnosti testových položek, a to včetně testové položky ID27.

Obvyklé hodnoty parametru diskriminace se pohybují v rozpětí od 0,5 do 2,5(3,0), přičemž platí, že vyšší hodnoty jsou spojeny s lepší schopností diskriminace (Edelen a Reeve, 2007; Toland, 2014). V našem případě mají relativně nižší schopnost diskriminace především testové položky ID27 a ID32, tj. testové položky, na něž poukázalo již hodnocení vycházející z CTT. Zároveň však odhad 3PL modelu a zohlednění parametru „pseudo-hádání“ významně zvyšuje schopnost testových položek diskriminovat mezi žáky vzhledem k jejich úrovni statistické gramotnosti.

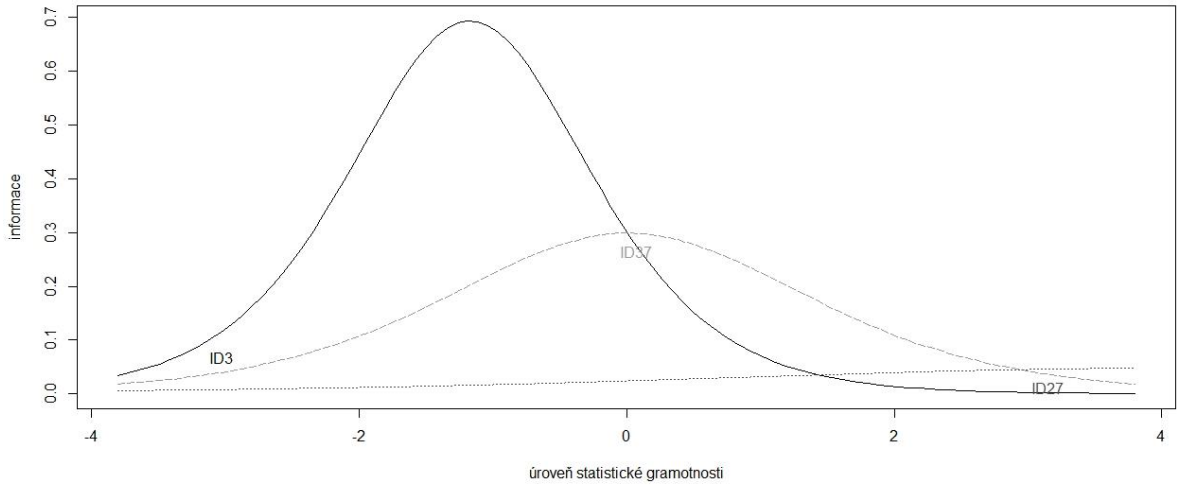
4.2.4 Hodnocení kvality testových položek – informační křivka a spolehlivost

Vedle parametrů testových položek testu statistické gramotnosti může být jejich kvalita posuzována také tzv. informační funkcí, která říká, jaké množství informace poskytuje testová položka na dané úrovni statistické gramotnosti žáka, přičemž platí, že vyšší množství poskytnuté informace zvyšuje spolehlivost testové položky. Uveďme, že informační funkce umožňuje konstruovat pro každou testovou položku informační křivku, která poskytuje grafickou informaci o spolehlivosti testové položky. Tato modelová případová studie, která vychází z odhadu 3PL modelu, se právě otázkou hodnocení spolehlivosti testových položek prostřednictvím konstrukce informační křivky zabývá.

- ***Řešení***

Konstrukce informační křivky testových položek testu statistické gramotnosti primárně vychází z odhadu 2PL modelu tak, jak byl popsán v modelové případové studii podkapitoly 4.2.1. Následně je z tohoto modelu odvozena informační funkce pro každou testovou položku testu statistické gramotnosti a s využitím informační funkce potom konstruovány jejich informační křivky. Obrázek č. 13 znázorňuje informační křivky vybraných testových položek testu statistické gramotnosti.

Obrázek č. 13: Informační křivky vybraných testových položek testu statistické gramotnosti žáků – 2PL model



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

- ***Interpretace***

Množství informace, kterou informační křivka poskytuje, je spojena se spolehlivostí testové položky. Takto je z obrázku č. 13 patrné, že nejvyšší spolehlivost je spojena s testovou položkou ID3, o něco nižší pak s testovou položkou ID37 a nejnižší s „více problémovou" testovou položkou ID27. Obecně platí, že nejvyšší množství informace poskytují testové položky na úrovni statistické gramotnosti, která odpovídá jejich obtížnosti. Důležitým souvisejícím aspektem jsou pak také hodnoty parametru diskriminace, kdy vyšší hodnoty diskriminace jsou spojeny s vyšší úrovní poskytnuté informace (srovnej také s údaji tabulky č. 30)

4.2.5 Hodnocení kvality testových položek – míra dobré shody

Kvalita testových položek testu statistické gramotnosti může být hodnocena také na základě souladu mezi empirickými a modelovými daty s tím, že žádoucí je v tomto ohledu vyšší míra shody obou typů dat. Tato modelová případová studie tedy řeší právě tuto otázku, a to pro parametry testových položek odhadovaných prostřednictvím 2PL modelu.

- ***Řešení***

Hodnocení dobré shody empirických dat a modelových dat, která jsou generována testovými položkami testu statistické gramotnosti, je založeno na využití tří statistik: (a) Yenovo Q_1; (b) *S-*χ^2 statistiky; a (c) *PV-Q1* statistiky. Předmětem hodnocení je jednak statistická významnost těchto statistik, kdy statisticky významná hodnota odmítá nulovou hypotézu o souladu empirických a modelových hodnot, jednak síla případného nesouladu hodnocená hodnotou RMSEA. *Cut-off* hodnota vysokého nesouladu je zde stanovena na 0,05. Tabulka č. 31 uvádí hodnoty tří statistik, a to pro odhad 2PL modelu testu statistické gramotnosti.

Tabulka č. 31: Odhady parametrů testových položek testu statistické gramotnosti

Testová položka	Yenovo Q_1	Yenovo Q_1 RMSEA	S-χ^2	S-χ^2 RMSEA	PV-Q1	PV-Q1 RMSEA
ID1	23,3*	0,025	51,7*	0,016	11,8	0,013
ID2	6,4	0,000	42,6	0,010	6,5	0,000
ID3	20,7*	0,023	38,5	0,012	6,9	0,000
ID4	10,4	0,010	26,0	0,000	9,1	0,007
ID5	10,4	0,010	39,8	0,010	9,3	0,007
ID6	12,6	0,014	41,3	0,011	10,6	0,010
ID7	17,1	0,019	39,5	0,010	10,8	0,011
ID8	28,0*	0,032	50,3*	0,019	10,1	0,012
ID9	20,1	0,022	24,8	0,000	10,0	0,009
ID10	28,9*	0,032	48,2*	0,018	10,7	0,013
ID11	18,6*	0,024	22,9	0,000	7,6	0,006
ID12	31,2*	0,031	42,4	0,012	17,5	0,020
ID13	19,8*	0,025	34,4	0,011	11,2	0,014
ID14	9,4	0,008	30,7	0,000	7,1	0,000
ID15	12,4	0,014	29,4	0,002	9,5	0,008
ID16	25,0*	0,027	28,6	0,000	13,8	0,016
ID17	22,7*	0,025	64,7*	0,019	18,8	0,021
ID18	9,4	0,008	26,2	0,000	7,7	0,000
ID19	40,7*	0,037	30,2	0,006	19,0	0,021
ID20	20,9*	0,023	36,0	0,008	14,3	0,016
ID21	19,2	0,022	58,6*	0,018	17,3	0,020
ID22	46,9*	0,040	31,4	0,005	16,2	0,019
ID23	64,6*	0,049	70,8*	0,021	46,1*	0,040
ID24	33,0*	0,032	24,3	0,000	16,6	0,019
ID25	34,2*	0,033	27,9	0,000	19,1	0,022
ID26	21,4*	0,024	40,8	0,012	9,8	0,009
ID27	35,8*	0,034	42,2	0,011	31,2*	0,031
ID28	19,5	0,022	44,2	0,012	8,5	0,004
ID29	19,0	0,021	40,3	0,010	13,7	0,015
ID30	19,9	0,022	50,0	0,014	11,8	0,013
ID31	23,5*	0,025	18,1	0,000	11,5	0,012
ID32	40,0*	0,037	53,7*	0,016	27,8*	0,029
ID33	34,8*	0,033	45,0	0,014	24,7*	0,026
ID34	54,0*	0,044	43,7	0,012	29,7*	0,030
ID35	26,7*	0,028	34,0	0,008	19,5	0,022
ID36	41,3*	0,037	51,9*	0,017	22,1*	0,024
ID37	18,5	0,021	34,2	0,008	10,3	0,010
ID38	12,0	0,013	40,2	0,010	12,0	0,013

* statisticky významná hodnota na 0,01 hladině významnosti

Zdroj: vlastní zpracování s využitím mirt package (Chalmers, 2020)

- ***Interpretace***

Hodnoty Yenova Q_1 jsou statisticky významné v případě vyššího počtu testových položek testu statistické gramotnosti, a některé testové položky vykazují statisticky významnou hodnotu také v případě *S-*χ^2 statistiky a *PV-Q1* statistiky. Zároveň však hodnota RMSEA je ve všech případech nižší, než je *cut-off* hodnota 0,05, což ukazuje na dobrý soulad empirických a modelových dat. Rozdíly ve zjištěních jsou dány nedostatky statistik založených na χ^2 rozdělení v případě velkých souborů dat. Uveďme, že pro další hodnocení kvality testových položek lze rovněž využít přesnější statistiky *PV-Q1** *či* χ^{2*}, které jsou však založeny na výpočetně náročné metodě parametrického bootstrapingu.

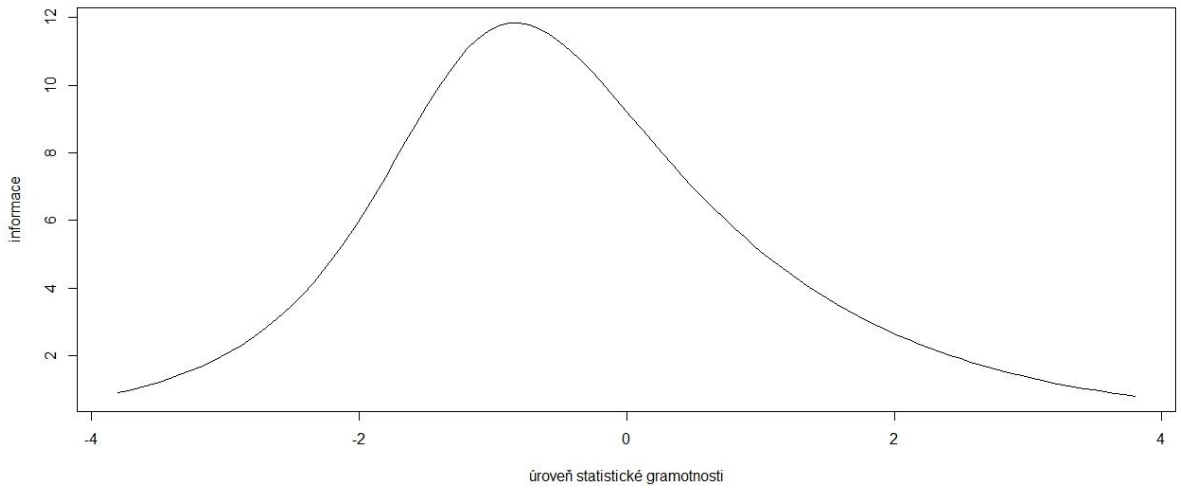
4.2.6 Hodnocení spolehlivosti testu

V případě modelů vycházejících z IRT je spolehlivost testu statistické gramotnosti žáků hodnocena jeho informační funkcí a informační křivkou. Informační funkce říká, jaké množství informace poskytuje test na dané úrovni statistické gramotnosti žáků, přičemž platí, že vyšší množství poskytnuté informace zvyšuje spolehlivost testu. Informační funkce zároveň umožňuje konstruovat informační křivku testu statistické gramotnosti, která poskytuje grafickou informaci o jeho spolehlivosti. Modelová případová studie se věnuje této otázce, a to specificky pro odhad 2PL modelu.

- ***Řešení***

Konstrukce informační křivky testu statistické gramotnosti primárně vychází z odhadu příslušného 2PL modelu tak, jak byl popsán v modelové případové studii podkapitoly 4.2.1. Následně je z tohoto modelu odvozena informační funkce testu statistické gramotnosti a z ní informační křivka zachycená na obrázku č. 14.

Obrázek č. 14: Informační křivka testu statistické gramotnosti žáků – 2PL model



Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018)

- ***Interpretace***

Množství informace, kterou informační křivka testu statistické gramotnosti poskytuje, je spojena se spolehlivostí testu. Takto je z obrázku č. 14 patrné, že test vykazuje nejvyšší spolehlivost v intervalu úrovně statistické gramotnosti mezi hodnotami -2 a 1 a jde tedy spíše o lehčí test. Pokud by bylo záměrem testu diskriminovat žáky na celé škále jejich úrovně statistické gramotnosti, bylo by žádoucí doplnit test o některé obtížnější testové položky. Zohledněn může být také vliv parametru „pseudo-hádání".

4.2.7 Hodnocení kvality testu – míra dobré shody

Volba modelu vycházejícího z IRT pro odhad úrovně statistické gramotnosti žáků je spojena s otázkou, do jaké míry je model schopen reprodukovat empirická data testu statistické gramotnosti, tj. zda skutečná empirická data mohou být vygenerována tímto modelem. Hovoříme o míře dobré shody na úrovni testu (modelu). V případě, že identifikujeme nízkou míru shody mezi naším modelem a empirickými daty, pak je zpochybněna validita našeho přístupu. Modelová případová studie se věnuje této otázce, a to specificky pro odhad 2PL modelu.

- ***Řešení***

Pro hodnocení míry dobré shody empirických dat na jedné straně a modelových dat 2PL testu statistické gramotnosti využíváme postup založený na odhadu dvou dílčích indexů:

- indexu *RMSEA*,
- indexu *SRMSR*,

přičemž pro oba indexy je stanovena *cut-off* hodnota ve výši 0,05. Tabulka č. 32 zachycuje hodnoty těchto indexů pro odhadovaný 2PL model.

Tabulka č. 32: Hodnoty indexů RMSEA a SRMSR testu statistické gramotnosti (2PL model)

	RMSEA	SRMSR
Hodnota	0,026	0,028

Zdroj: vlastní zpracování s využitím mirt package (Chalmers, 2020)

- ***Interpretace***

Hodnota indexů RMSEA i SRMSR je nižší než *cut-off* hodnota dobré shody skutečných a modelových dat. Tato skutečnost opodstatňuje vhodnost využití 2PL modelu testu statistické gramotnosti.

4.2.8 Výběr modelu srovnáním souladu empirických a modelových dat

V předchozích modelových případových studiích byly odhadovány různé typy modelů vycházejících z IRT, které byly založeny na empirických datech testu statistické gramotnosti, tj.: (a) 1PL model; (b) 2PL model; a (c) 3PL model. Přirozenou otázkou je, který z uvedených modelů je pro odhad úrovně statistické gramotnosti žáků nejvhodnější. Touto otázkou se zabývá tato modelová případová studie.

- ***Řešení***

Řešení modelové případové studie je založeno na srovnání tří indexů dobré shody pro odhady 1PL, 2PL a 3PL modelu testu statistické gramotnosti, a to:

- testu poměru věrohodností zahnízděných modelů (LRT),
- Akaikeho informačního kritéria (AIC),
- Bayesova (Schwarzova) informačního kritéria (BIC).

Tabulka č. 33 zachycuje hodnoty těchto tří indexů pro jednotlivé odhadované modely.

Tabulka č. 33: Srovnání hodnot indexů dobré shody pro 1PL, 2PL a 3PL modely testu statistické gramotnosti

Model	LRT	AIC	BIC
1PL model	-	133 965	134 199
2PL model	1 830**	132 209	132 665
3PL model	473,0**	131 812	132 497

* statisticky významný rozdíl na 0,001 hladině významnosti

- ***Interpretace***

Hodnoty indexů dobré shody, které jsou uvedeny v tabulce 4-7, ukazují na preferenci 3PL modelu před 2PL modelem a 2PL modelu před 1PL modelem (viz tabulka č. 33 a statistická významnost rozdílů modelů, respektive preferovaně nižší hodnoty indexů AIC a BIC). Pro modelování empirických dat testu statistické gramotnosti je nejvhodnější zvolit 3PL model, a to rovněž v kontextu již dříve uváděného vlivu parametru „pseudo-hádání“ na odhad parametrů testových položek.

4.2.9 Hodnocení neobvyklého vzoru odpovědí žáka na testové položky

V modelové případové studii v podkapitole 4.1.4 bylo popsáno několik postupů pro identifikaci žáků s neobvyklým vzorem odpovědí na testové položky testu statistické gramotnosti. Zatímco v modelové případové studii v podkapitole 4.1.4 byly sledovány metodické přístupy založené na CTT, v této modelové případové studii jsou představeny metodické přístupy vyžadující odhad modelů vycházejících z IRT (opětovně odhad 2PL modelu).

- ***Řešení***

Hodnocení neobvyklé struktury odpovědí žáka na testové položky, které vychází z IRT, je založeno na výpočtu statistiky lz^*. Platí, že špatný soulad empirických a modelových dat je spojen s nízkými hodnotami statistiky lz^*. Tabulka č. 34 zachycuje hodnoty lz^* těch žáků, u kterých byl v modelové případové studii v kapitole 4.1.4 identifikován nejvyšší nesoulad mezi skutečným a ideálním vzorem odpovědí na testové položky testu statistické gramotnosti. Doplněno je také pořadí těchto žáků vzhledem ke statistice lz^* (ve vzestupném pořadí).

Tabulka č. 34: Hodnota lz^* žáků s nejméně obvyklou strukturou odpovědí na testové položky v modelové případové studii kapitoly 4.1.4

ID žáka	lz^*	Pořadí podle lz^*	Pořadí podle r.pbis	Pořadí podle C*	Pořadí podle U3
1199	-3,42	9	1	2	4
2588	-3,98	3	2	5	5
1805	-3,41	10	3	3	1
1986	-2,73	33	4	1	3
2351	-2,98	25	5	4	2
2605	-3,90	4	6	8	8
2900	-3,08	20	7	6	6
1989	-3,10	18	8	9	9
2367	-2,82	30	9	7	7
2304	-2,82	29	10	11	10

Zdroj: vlastní zpracování s využitím PerFit package (Tendeiro, 2018)

- ***Interpretace***

Vzor odpovědí žáků, který byl hodnocen jako neobvyklý přístupy vycházejícími z CTT, lze označit za poměrně neobvyklý také při využití přístupu založeného na IRT. Přesto závěry obou metodických přístupů vykazují určité rozdíly, když statistika lz^* indikuje nejvíce neobvyklý vzor odpovědí u jiných žáků než přístupy vycházející z CTT. Takto lze považovat za přínosné využití obou metodických přístupů pro identifikaci neobvyklého vzoru odpovědí žáků na testové položky testu statistické gramotnosti.

4.2.10 Hodnocení lokální nezávislosti testových položek

Lokální nezávislost testových položek je jeden z významných aspektů, který ovlivňuje kvalitu škál, jejich tvorba je založena jak na přístupech vycházejících z CTT (viz otázka týkající se unidimenzionality testu), tak na přístupech založených na IRT. Záměrem této modelové případové studie je proto představit způsob hodnocení lokální nezávislosti testových položek.

- ***Řešení a interpretace***

Řešení modelové případové studie je primárně založeno na odhadu 1PL modelu, z něhož jsou využity informace o hodnotách úrovně statistické gramotnosti žáků (Θ) a obtížnosti testových položek testu statistické gramotnosti. Tyto informace jsou následně využity pro odhad statistiky *Q3* pro každou dvojici testových položek testu statistické gramotnosti, kdy potenciálně problémové jsou především hodnoty vyšší než 0,2. V matici hodnot *Q3* byla zaznamenána jediná hodnota *Q3* vyšší než 0,2, a to pro dvojici testových položek ID10 a ID11 (*Q3* = 0,233). Vztahu mezi těmito dvěma testovými položkami je vhodné věnovat pozornost.

4.2.11 Propojení dosaženého skóre žáků na společnou škálu

Modelová případová studie v podkapitole 4.1.8 představuje metodický postup, jak propojit dosažené skóre dvou skupin žáků, kteří řešili dva různě obtížné testy statistické gramotnosti, tj. jak převést tato skóre na společnou škálu. Tato modelová případová studie řeší stejný problém, využitý přístup k řešení ovšem nevychází z CTT, jako je tomu v podkapitole 4.1.8, nýbrž je založen na IRT. Záměrem modelové případové studie tedy je opětovně propojit dosažené skóre žáků na společnou škálu tak, aby ani jedna skupina žáků nebyla znevýhodněna odlišnou obtížnosti testu. Za tímto účelem je využit odhad 1PL a 3PL modelů pro oba propojované testy statistické gramotnosti žáků.

- ***Řešení***

Řešení modelové případové studie primárně vychází z NEAT přístupu ke sběru dat, a proto vyžaduje transformaci škál testů na společnou škálu. V případě 3PL modelu je za tímto účelem využit vztah:

$$P_{ij}(\theta_{iK}; a_{jK}; b_{jK}; c_{jK}) = P_{ij}\left(A^*\theta_{iJ} + B^*; \frac{a_{jJ}}{A^*}; A^*b_{jJ} + B^*; c_{jJ}\right),$$

kde *a* je parametr diskriminace, *b* je parametr obtížnosti a *c* je parametr „pseudo-hádání“ testových položek a kde záměrem je nalézt takové hodnoty A^* a B^*, které minimalizují odlišnost obou vztahů. Pro nalezení hodnot A^* a B^* je možné využít různé přístupy, mezi které patří také Haeberův či Stocking-Lordův přístup. V případě 1PL modelu se výše uvedený vztah přirozeně zjednodušuje fixací parametrů diskriminace a „pseudo-hádání“.

Vlastní postup řešení modelové případové studie vychází z odhadu parametrů zvoleného modelu (1PL a 3PL model), a to pro oba propojované testy a následně z výpočtu hodnot A^* (1,046 pro 1PL model a 1,165 pro 3PL) a B^* (-0,159 pro 1PL model a -0,147 pro 3PL model) pro transformaci škály prvního testu na škálu druhého testu. Pro výpočet je využit Haeberův přístup. Na závěr jsou parametry modelů a hodnoty A^* a B^* využity pro propojení skóre obou testů.

Tabulka č. 35 zachycuje korespondující skóre obou testů po převedení na společnou škálu, a to jak s využitím metodického přístupu vycházejícího z CTT (viz modelová případová studie v podkapitole 4.1.8), tak s využitím metodického přístupu vycházejícího z IRT (1PL a 3PL model).

Tabulka č. 35: Korespondující skóre prvního a druhého testu statistické gramotnosti (vybraná skóre; vybrané metodické přístupy)

Test	Korespondující skóre									
Test 1	11	12	13	14	15	16	17	18	19	20
Test 2 (CTT)	11,1	12,2	13,3	14,5	15,6	16,5	17,5	18,5	19,7	20,9
Test 2 (IRT – 1PL)	11,3	12,3	13,4	14,4	15,5	16,5	17,6	18,6	19,7	20,7
Test 2 (IRT – 3PL)	10,9	12,0	13,0	14,1	15,2	16,3	17,3	18,4	19,6	20,7
Test	**Korespondující skóre**									
Test 1	21	22	23	24	25	26	27	28	29	30
Test 2 (CTT)	22,1	23,3	24,4	25,5	26,3	27,2	28,2	29,2	30,1	31,0
Test 2 (IRT – 1PL)	21,7	22,8	23,8	24,8	25,9	26,9	27,9	28,9	30,9	31,9
Test 2 (IRT – 3PL)	21,9	23,0	24,1	25,2	26,3	27,4	28,5	29,5	30,5	31,5

Zdroj: vlastní zpracování s využitím ltm package (Rizopoulos, 2018) a equateIRT package (Battauz, 2018)

- ***Interpretace***

Tabulka č. 35 zachycuje odpovídající si hodnoty skóre na společné škále dvou hodnocených testů. Obecně jsou výsledky získané metodickými přístupy vycházejícími z CTT a IRT obdobné, pozorovat však lze vyšší variabilitu odchylek v případě využití 3PL modelu, než v případě využití 1PL modelu. Tato skutečnost je přirozeně dána vyšším počtem odhadovaných parametrů 3PL modelu. Doplňme, že pomocí parametrů A^* a B^* lze snadno dopočítat korespondující hodnoty také pro úroveň statistické gramotnosti žáků (hodnota Θ pro oba testy na společné škále).

4.2.12 Odhad parametrů multidimenzionálních modelů

Na několika místech této knihy byla zdůrazněna důležitost předpokladu unidimenzionality pro správný odhad parametrů modelů vycházejících z IRT. V případě narušení tohoto předpokladu je jednou z možností, jak tento nedostatek řešit, odhad multidimenzionálních modelů. A právě představením tohoto metodického přístupu pro test statistické gramotnosti žáků se zabývá tato modelová případová studie.

- ***Řešení***

Řešení modelové případové studie je založeno na odhadu faktorového modelu vycházejícího z paradigmatu IRT (zde 2PL model), a to s využitím metody maximální věrohodnosti a EM algoritmu (blíže viz Chalmers, 2020). Předpokládejme přitom, že v testu statistické gramotnosti jsou obsaženy dva konstrukty (faktory, dimenze). Odhad příslušného modelu nám následně podává informaci o:

- hodnotách faktorových zátěží vyjadřujících vztah mezi konstrukty (faktory, dimenzemi) a testovými položkami (viz tabulka č. 36);

- úrovní zvládnutí obou hodnocených konstruktů žáky (hodnota Θ pro oba konstrukty).

Doplňme informaci o velmi vysoké úrovni korelace (0,95) mezi hodnotami Θ obou konstruktů testu statistické gramotnosti žáků.

Tabulka č. 36: Faktorové zátěže testových položek ke dvěma konstruktům testu statistické gramotnosti žáků

Testová položka	Konstrukt (faktor) 1	Konstrukt (faktor) 2	Testová položka	Konstrukt (faktor) 1	Konstrukt (faktor) 2
ID1	0,479	0,014	ID20	0,456	0,016
ID2	0,342	0,021	ID21	0,453	0,089
ID3	0,655	0,057	ID22	0,433	0,097
ID4	0,372	-0,027	ID23	0,278	0,133
ID5	0,401	0,044	ID24	0,628	0,095
ID6	0,525	0,022	ID25	0,567	-0,032
ID7	0,374	0,074	ID26	0,614	0,013
ID8	0,764	0,002	ID27	0,136	0,133
ID9	0,435	0,009	ID28	0,369	0,046
ID10	0,821	-0,046	ID29	0,313	0,136
ID11	0,885	-0,107	ID30	0,391	-0,002
ID12	0,531	-0,077	ID31	0,159	0,154
ID13	0,732	0,059	ID32	0,204	0,029
ID14	0,353	0,101	ID33	0,062	0,545
ID15	0,580	0,073	ID34	-0,074	0,705
ID16	0,329	0,095	ID35	0,060	0,715
ID17	0,472	-0,024	ID36	0,041	0,709
ID18	0,368	0,055	ID37	0,240	0,358
ID19	0,691	0,061	ID38	0,079	0,299

Zdroj: vlastní zpracování s využitím mirt package (Chalmers, 2020)

- ***Interpretace***

Odhad multidimenzionálních modelů je jednou z cest řešení narušení předpokladu unidimenzionality testu, kdy je v něm obsažen vyšší počet konstruktů. Modelová případová studie představená v podkapitole 4.1.6 ukázala na dominantní hlavní konstrukt (faktor) testu statistické gramotnosti žáků, přičemž doporučila se věnovat také hodnocení řešení se dvěma konstrukty (faktory). V souladu se zjištěními analýzy tetrachorických korelací mezi dvojicemi testových položek (obrázek č. 5) je druhý konstrukt (faktor) sycen především testovými položkami umístěnými na konci testu (ID33 až ID36). Současně však velmi vysoká hodnota

korelace mezi oběma konstrukty (faktory) opodstatňuje řešení na bázi unidimenzionálního testu.

4.3 Vyhodnocení a reporting výsledků testů

Poslední dvě modelové případové studie se vztahují k vyhodnocení a reportingu výsledků testů, a to s důrazem na širší souvislosti problematiky.

4.3.1 Rozdíly ve výsledcích žáků – vliv rozdílů uvnitř školy a rozdílů mezi školami

Při vyhodnocení ověřovacích testů je jednou z tradičních také otázka, do jaké míry jsou rozdíly ve výsledcích žáků způsobeny: (a) rozdíly mezi školami; a (b) rozdíly uvnitř škol. Dochází ke koncentraci žáků s vyšší úrovní statistické gramotnosti v některých školách, nebo jsou žáci vzhledem ke své úrovni statistické gramotnosti rozděleni mezi školami rovnoměrně? Tato modelová případová studie se věnuje právě této otázce, přičemž pro hodnocení využívá jednak ukazatel podílu správně zodpovězených testových položek testu statistické gramotnosti žáků, jednak ukazatel úrovně statistické gramotnosti žáků měřené na bodové škále z modelové případové studie kapitoly 4.2.2.

- ***Řešení a interpretace***

Řešení modelové případové studie je založeno na výpočtu tzv. koeficientu vnitrotřídní korelace (ICC), a to prostřednictvím odhadů hierarchických regresních modelů, které modelují vztah mezi výsledkem žáků v testu statistické gramotnosti a vysvětlujícími proměnnými na dvou úrovních: (a) na úrovni žáka; a (b) na úrovni školy. Postup řešení modelové případové studie následně zahrnuje dva kroky:

- V prvním kroku postupu jsou odhadovány hierarchické regresní modely, kdy výsledek žáků v testu informační gramotnosti, který je měřený na jedné ze dvou uvedených škál, je modelován ve vztahu k jediné proměnné odpovídající identifikátoru školy.
- Ve druhém kroku postupu jsou z odhadů hierarchického regresního modelu podle prvního kroku vypočteny podíly na celkovém rozptylu výsledků žáků v testu informační gramotnosti, které odpovídají: (a) rozdílům mezi školami (rozptyl pro proměnnou identifikátoru školy, také ICC); a (b) rozdílům mezi žáky, tj. uvnitř škol (rozptyl reziduí).

Hodnota ICC vyšla v případě ukazatele podílu správně zodpovězených testových položek testu statistické gramotnosti žáků 9,9 % a v případě bodové škály 9,4 %. Tyto hodnoty jsou poměrně nízké a indikují spíše menší rozdíly mezi školami vzhledem ke koncentraci žáků s vyšší úrovní statistické gramotnosti v nich.

4.3.2 Hodnocení faktorů ovlivňujících výsledek žáků

Tradiční součástí vyhodnocení ověřovacích testů je také posouzení statistické významnosti vztahů mezi výsledky žáků v testu na jedné straně a vysvětlujícími proměnnými charakterizujícími žáka, třídu či školu na straně druhé. Právě otázkou posouzení podoby vztahů mezi dvěma faktory charakterizujícími žáky a jejich výsledkem v testu statistické gramotnosti se zabývá tato modelová případová studie. Konkrétně se přitom jedná o následující faktory:

(a) pohlaví žáka s rozlišením kategorií dívek a chlapců, s kategorií dívek jako kategorií referenční;

(b) zařazení žáka do kategorie žáků s vyšším socioekonomickým statusem rodinného zázemí.

Vysvětlovanou proměnnou je výsledek žáků v testu statistické gramotnosti, a to na dvou škálách: (a) podílu správně zodpovězených testových položek testu statistické gramotnosti; a (b) bodové škále se středem 500 bodů a směrodatnou odchylkou 100 bodů odpovídající modelové případové studii v podkapitole 4.2.

- ***Řešení***

Řešení modelové případové studie je založeno na odhadech hierarchických regresních modelů, které modelují vztah mezi výsledkem žáků v testu statistické gramotnosti (jedna te tří zvolených škál měření výsledků žáků v testu statistické gramotnosti) na jedné straně a vysvětlujícími proměnnými na straně druhé, přičemž zohledněn je také vliv školy na druhé úrovni hierarchie. Volba hierarchických regresních modelů je motivována skutečností, že odhad nehierarchických regresních modelů je nepřesný, pokud v datech existuje závislost výsledků žáků na škole, kterou navštěvují. Narušen je tak předpoklad nezávislosti dat nehierarchických regresních modelů.

Tabulka č. 37 zachycuje odhady koeficientů hierarchických regresních modelů, hodnoty směrodatné chyby odhadovaných koeficientů a statistickou významnost hodnot koeficientů. Vedle toho jsou reportovány hodnoty standardizovaných koeficientů (viz také obrázek č. 15 pro zachycení standardizovaných koeficientů a intervalu spolehlivosti k nim). Uveďme, že model je odhadován metodou maximální věrohodnosti.

Tabulka č. 37: Odhady hierarchických regresních modelů

Vysvětlující proměnná	Podíl správných odpovědí		Bodová škála	
	β_i (se)	*stand* β_i (se)	β_i (se)	*stand* β_i (se)
Dívka	0,43 (0,64)	0,011 (0,016)	1,2 (3,3)	0,006 (0,016)
Socioekonomický status	22,96*** (0,84)	0,456*** (0,017)	112,4*** (4,3)	0,442*** (0,017)
AIC	25 723		35 486	

*** statisticky významná hodnota na 0,001 hladině významnosti; ** statisticky významná hodnota na 0,01 hladině významnosti; * statisticky významná hodnota na 0,05 hladině významnosti

Zdroj: vlastní zpracování s využitím lme4 package (Bates et al., 2020), lmerTest package (Kuznetsova et al., 2020) a sjstats package (Lüdecke, 2020)

Obrázek č. 15: Standardizované koeficienty vysvětlovaných proměnných hierarchických regresních modelů (95% interval spolehlivosti)

Zdroj: vlastní zpracování s využitím lme4 package (Bates et al., 2020) a sjstats package (Lüdecke, 2020)

- ***Interpretace***

Z informace o statistické významnosti hodnot koeficientů vysvětlujících proměnných a o znaménku těchto hodnot vyplývá, že:

- neexistují statisticky významné rozdíly ve výsledcích dívek a chlapců v testu statistické gramotnosti, a to bez ohledu na zvolenou škálu;
- žáci pocházející z rodin vyššího socioekonomického statusu dosáhli statisticky významně lepších výsledků v testu statistické gramotnosti než žáci z takového rodinného prostředí nepocházející.

Tuto interpretaci potvrzuje také grafické znázornění standardizovaných koeficientů s 95% intervalem spolehlivosti (obrázek č. 15). Zároveň však upozorněme na určité odlišnosti výsledků v závislosti na zvolené škále měření statistické gramotnosti žáků. V některých případech může mít tato skutečnost dopad na interpretaci výsledků.

5. Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání

V návaznosti na představená teoretická východiska (kapitola 3) a zpracované modelové případové studie (kapitola 4) jsou zjištěné poznatky syntetizovány do podoby metodiky vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání (dále i „metodika“). Záměr formulovat takovou metodiku vychází z ne plně využívaných příležitostí spojených s širokou nabídkou teoreticko-metodických přístupů k vyhodnocení ověřovacích testů. Cílem metodiky proto je poskytnout uživatelům komplexní podpůrný nástroj s návodnými metodickými postupy vyhodnocení ověřovacích testů v počátečním vzdělávání. Metodika se přitom zaměřuje na testy, které jsou utvářeny dichotomickými testovými položkami. Podstata vyhodnocení testů, které obsahují polytomické testové položky, je analogická, nicméně konkrétní postupy je nezbytné přizpůsobit specifikům těchto testových položek.

5.1 Podstata metodiky – obecný a specifický rámec metodiky

Podstata metodiky vychází z jejího podpůrného charakteru, kdy uživatelům poskytuje návodné postupy pro naplňování záměrů přípravy a vyhodnocení testů. Za tímto účelem metodika vymezuje tzv. modelové situace, k nímž je definován jednak obecný rámec společný pro všechny modelové situace a jednak specifický rámec, který je pro každou modelovou situaci jedinečný. Obecný rámec metodiky určuje společný obsah pro rozvedení (specifikaci) všech modelových situací, a to v podobě následujících dílčích elementů:

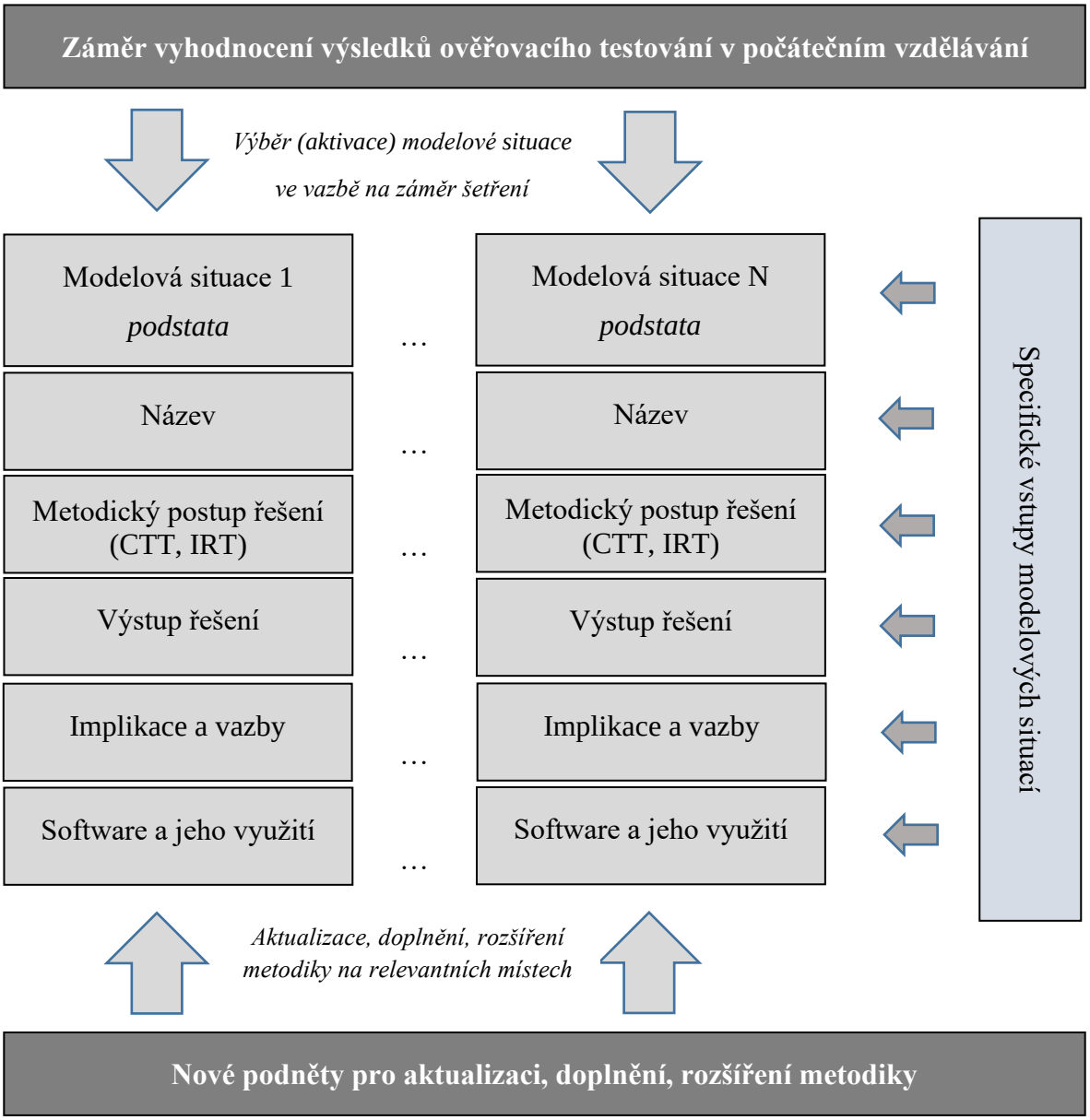
- název modelové situace;
- vysvětlení podstaty modelové situace;
- charakteristika metodického postupu řešení modelové situace, který vychází z teoreticko-metodických postupů CTT nebo IRT;
- představení výstupu řešení modelové situace;
- zasazení řešení modelové situace do širšího situačního kontextu (implikace a vazby k dalším modelovým situacím);
- uvedení relevantního software a ilustrace jeho využití při řešení modelové situace.

Specifický rámec metodiky je pak založen na rozvedení (specifikaci) dílčích elementů obecného rámce pro každou modelovou situaci zvlášť.

Podstata metodiky dále sleduje dva horizontální principy: (a) princip výběru modelové situace ve vazbě na záměr vyhodnocení; a (b) princip otevřenosti metodiky vůči novým podnětům. První z těchto principů je spojený s aktivací řešení modelových situací podle záměru uživatele metodiky. V souladu s tímto principem je uživatelem aktivován odpovídající počet modelových situací, který může zahrnovat jednu i všechny z nich, přičemž při výběru posuzuje uživatel především soulad svého záměru s podstatou modelových situací. Princip otevřenosti metodiky pak respektuje potřebu její flexibility ve vazbě jak na vývoj vědeckého poznání, tak na utváření nových potřeb souvisejících s ověřovacím testováním v počátečním vzdělávání. V souladu

s principem otevřenosti je možné metodiku aktualizovat a doplňovat v reakci na nové podněty. Obrázek č. 16 schematicky zachycuje celkovou podstatu metodiky.

Obrázek č. 16: **Schéma podstaty metodiky**



Doplňme, že podstata metodika respektuje význam předpokladů pro správnou aplikaci metod vycházejících z CTT a především z IRT (blíže viz kapitola 3).

Navrhované metodické postupy pro řešení modelových situací vyžadují využití softwarové podpory. V tomto ohledu je metodika postavena na tzv. balíčcích (knihovnách, *packages*) programovacího jazyka R (dále i „R balíček“), přičemž tento postup je motivován: (a) výbornými vlastnostmi programovacího jazyka R pro provádění pokročilých statistických analýz dat; a (b) snadnou dostupností R balíčků v rámci tzv. svobodné licence. Tyto motivy umožňují prakticky každému zájemci využití metodiky bez ohledu na jeho finanční či jiná

omezení. Konečně pro ilustraci vzorových příkladů řešení modelových situací využívá metodika datové soubory, na nichž byly představeny modelové situace v kapitole 4. Takto vzniká úzká provázanost metodiky s touto knihou.

5.2 Specifický rámec metodiky

Specifický rámec metodiky je rozdělen na dvě části. První část vysvětluje podstatu modelových situací, druhá část pak modelové situace rozvádí v rámci ostatních dílčích elementů obecného rámce metodiky. Takové členění specifického rámce metodiky je motivováno záměrem usnadnit jejímu uživateli nalezení vhodné modelové situace k využití, a to díky přehlednější orientaci ve vysvětlení jejich podstaty. Teprve po tomto kroku uživatel metodiky přechází k vlastnímu návodnému postupu řešení modelové situace. V dalších podkapitolách je blíže ilustrováno rozvedení (specifikace) pěti vybraných modelových situací, přičemž celkově metodika obsahuje osmnáct modelových situací.

5.2.1 Spolehlivost (škály) ověřovacího testu

<table>
<tr><th colspan="2">Název modelové situace</th></tr>
<tr><td colspan="2">Spolehlivost (škály) ověřovacího testu</td></tr>
<tr><th colspan="2">Podstata modelové situace</th></tr>
<tr><td colspan="2">Podstata modelové situace reaguje na záměr uživatele metodiky zjistit, jaká je spolehlivost (škály) ověřovacího testu. Spolehlivost (škály) ověřovacího testu může být negativně ovlivněna řadou vlivů, jako jsou: (a) nekontrolované podmínky testování (např. prostorové a časové podmínky testování); (b) náhodná fluktuace výsledků žáka spojená s jeho osobními náladami a charakteristikami; nebo (c) chyba spojená s kvalitou slovního vyjádření testových položek. Zájmem uživatele metodiky je dosáhnout vysoké spolehlivosti ověřovacího testu, přičemž tento zájem se zvyšuje v závislosti na důležitosti, tj. praktických dopadech, ověřovacího testu.</td></tr>
<tr><th colspan="2">Metodický postup řešení modelové situace</th></tr>
<tr><td colspan="2">Metodický postup řešení modelové situace, která hodnotí spolehlivost (škály) ověřovacího testu na základě odpovědí žáků na testové položky, se skládá z následujících kroků uživatele metodiky:
Krok 1: Uživatel metodiky vybírá, zda bude spolehlivost ověřovacího testu hodnotit s využitím odhadu modelu vycházejícího z IRT.</td></tr>
<tr><td>Ne (přístup CTT)
Krok 2: Uživatel metodiky počítá hodnoty šesti ukazatelů spolehlivosti testu (λ_1 až λ_6), kdy λ_3 je Cronbachovo alfa.
Krok 3: Uživatel metodiky počítá dva ukazatele spolehlivosti testu ω_h a ω_t, kde ukazatel ω_h se vztahuje jen k hlavnímu konstruktu a ukazatel ω_t zohledňuje také vliv skupinových konstruktů dílčích testových položek.</td><td>Ano (přístup IRT)
Krok 2: Uživatel metodiky odhaduje zvolený model v souladu s metodickým postupem modelové situace „Volba a využití škály pro stanovení výsledků žáků v ověřovacím testu“.
Krok 3: Uživatel metodiky odvozuje a vykresluje informační křivku testu</td></tr>
</table>

Výstup řešení modelové situace

Výstupem řešení modelové situace je osm hodnot ukazatelů spolehlivosti λ_1 až λ_6, ω_h a ω_t v intervalu od 0 do 1. Vyšší hodnoty ukazatelů spolehlivosti λ_1 až λ_6 a ω_h ukazují na vyšší spolehlivost testu. Nižší hodnoty ukazatele spolehlivosti ω_h a vysoké hodnoty ukazatele ω_t ukazují na slabší vliv hlavního konstruktu a silný vliv skupinových konstruktů dílčích testových položek (vedlejší konstrukty ověřovacího testu). Taková situace narušuje důvěryhodnost ukazatelů spolehlivosti λ_1 až λ_6.

Výstupem řešení modelové situace je vykreslení informační křivky testu, která na ose *x* zachycuje úroveň zvládnutí hodnoceného konstruktu a na ose *y* korespondující množství poskytnuté informace (spolehlivost) testu. Vyšší množství poskytnuté informace znamená vyšší spolehlivost testu na dané úrovni zvládnutí hodnoceného konstruktu žáky.

Širší situační kontext

Spolehlivost (škály) ověřovacího testu je logicky vztažena ke kvalitě testových položek, kdy vynechání nekvalitních testových položek typicky vede ke zvýšení spolehlivosti (škály) ověřovacího testu.

Nejnižší uváděná minimální hodnota Cronbachova alfa pro akceptovatelnou spolehlivost ověřovacího testu je 0,7, v případě testů s významnými dopady je však typicky vyžadována spolehlivost vyšší.

Srovnání hodnot ukazatelů spolehlivosti ω_h a ω_t je vhodným metodickým postupem pro posouzení unidimenzionality testu, a to ve vazbě na sílu vlivu hlavního konstruktu. Následně je možné pro detekci narušení předpokladu unidimenzionality využít další metodické postupy.

Informační křivka testu je vhodným nástrojem pro posouzení vhodnosti ověřovacího testu k záměru testování, neboť ukazuje množství poskytnuté informace na jednotlivých úrovních zvládnutí hodnoceného konstruktu žáky. Tam, kde je množství poskytnuté informace nedostatečné, lze přidat testovou položku vhodných charakteristik (parametry obtížnosti a diskriminace).

Software a ilustrace jeho využití pro řešení modelové situace

R balíček *psych* (viz Revelle, 2020)

R balíček *ltm* (viz Rizopoulos, 2018)

Přístup CTT pro datový rámec TACR_DATA

```
guttman(TACR_DATA)
```

 # Funkce guttman (R balíček *psych*) *umožňuje výpočet hodnot šesti Guttmanových ukazatelů* λ_1 *až* λ_6.

```
omega (TACR_DATA)
```

 # Funkce omega (R balíček *psych*) *umožňuje výpočet hodnot* ω_h *a* ω_t.

Přístup IRT pro datový rámec TACR_DATA a odhad 2PL modelu

```
PL2model <- ltm(TACR_DATA ~ z1)
```

 # Funkce ltm (R balíček *ltm*) *vede k odhadu parametrů 2PL modelu, které jsou uloženy v objektu PL2model.*

```
plot(PL2model, type = "IIC", items = 0, xlab = "úroveň statistické gramotnosti", ylab = "informace", main = "")
```

 # Funkce plot (R balíček *ltm*) *vykresluje informační křivku testu statistické gramotnosti pro odhadovaný 2PL model z předchozího kroku metodického postupu.*

 # Atribut type s hodnotou IIC *sděluje, že má být vykreslena informační křivka.*

 # Atribut items s hodnotou 0 *sděluje, že má být vykreslena informační křivka celého testu, nejen jednotlivých testových položek. Pro vykreslení vybraných testových položek se zadává číslo jejich pořadí (např. items = c(1, 3, 7, 8)).*

 # Atribut xlab *uvádí označení osy x, v tomto případě úroveň statistické gramotnosti; atribut* ylab *uvádí označení osy y, v tomto případě informace; atribut* main *uvádí název celého grafu, v tomto případě nebude zobrazen žádný název.*

5.2.2 Kvalita testových položek a identifikace nekvalitních testových položek

Název modelové situace
Kvalita testových položek a identifikace nekvalitních testových položek
Podstata modelové situace
Podstata modelové situace reaguje na záměr uživatele metodiky komplexně posoudit kvalitu testových položek ověřovacího testu. Tento záměr je především motivován zájmem uživatele metodiky nalézt nekvalitní testové položky, tj. testové položky s nežádoucími charakteristikami, které se především týkají: (a) jejich obtížnosti; (b) jejich schopnosti rozlišit mezi žáky s dobrými a slabými výsledky v ověřovacím testu; (c) kvality nabízených nesprávných odpovědí (distraktorů) na ně; (d) změny spolehlivosti ověřovacího testu při jejich vynechání; a (e) jejich (ne)spravedlnosti k různým skupinám žáků. V případě modelů vycházejících z IRT je dále posuzována dobrá shoda skutečných a modelem predikovaných odpovědí žáků na testovou položku. V případě identifikovaných nekvalitních testových položek uživatel metodiky rozhoduje o způsobu jejich vyhodnocení. Modelová situace pak také poskytuje vstupní informace o testových položkách při rozhodování o jejich využití při tvorbě nového ověřovacího testu.
Metodický postup řešení modelové situace
Metodický postup řešení modelové situace je založen na výpočtu hodnot ukazatelů (parametrů), které charakterizují kvalitu testových položek: (a) obtížnost testových položek; (b) schopnost testových položek rozlišit mezi žáky s dobrými a slabými výsledky v ověřovacím testu; (c) kvalita nabízených nesprávných odpovědí (distraktorů) na testové položky; (d) změna spolehlivosti ověřovacího testu při vynechání testových položek; a (e) spravedlnost testových položek k různým skupinám žáků. V případě aplikace přístupů vycházejících z IRT je rovněž kvalita testové položky hodnocena prostřednictvím statistik (indexů) dobré shody skutečných a modelem predikovaných odpovědí žáků na testovou položku. Při výpočtu hodnot ukazatelů pro charakteristiky kvality testových položek (a) až (e) jsou aplikovány metodické postupy pěti souvisejících modelových situací: (a) „Zvládnutí testové položky žáky a stanovení obtížnosti testové položky“; (b) „Schopnost testové položky rozlišit mezi žáky podle jejich výsledků v ověřovacím testu, identifikace matoucí správné odpovědi a potenciálně chybný klíč hodnocení“; (c) „Kvalita nabízených nesprávných odpovědí testové položky“; (d) „Změna spolehlivosti ověřovacího testu při vynechání testové položky“; a (e) „Hodnocení spravedlnosti testových položek k charakteristikám různých skupin žáků“. Takto při řešení modelové situace sleduje uživatel metodiky následující kroky postupu: Krok 1: Uživatel metodiky vybírá, zda bude kvalitu testových položek posuzovat s využitím odhadu modelu vycházejícího z IRT.

Ne (přístup CTT)	*Ano (přístup IRT)*
Krok 2: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Zvládnutí testové položky žáky a stanovení obtížnosti testové položky“ pro přístup CTT.	Krok 2: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Zvládnutí testové položky žáky a stanovení obtížnosti testové položky“ pro přístup IRT.
Krok 3: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Schopnost testové položky rozlišit mezi žáky podle jejich výsledků v ověřovacím testu, identifikace matoucí správné odpovědi a potenciálně chybný klíč hodnocení“ pro přístup CTT.	Krok 3: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Schopnost testové položky rozlišit mezi žáky podle jejich výsledků v ověřovacím testu, identifikace matoucí správné odpovědi a potenciálně chybný klíč hodnocení“ pro přístup IRT.

Krok 4: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Kvalita nabízených nesprávných odpovědí testové položky“ pro přístup CTT. Krok 5: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Změna spolehlivosti ověřovacího testu při vynechání testové položky“ pro přístup CTT. Krok 6: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Hodnocení spravedlnosti testových položek k charakteristikám různých skupin žáků“ pro přístup CTT.	Krok 4: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Kvalita nabízených nesprávných odpovědí testové položky“ pro přístup IRT. Krok 5: Uživatel metodiky aplikuje metodický postup řešení modelové situace „Hodnocení spravedlnosti testových položek k charakteristikám různých skupin žáků“ pro přístup IRT. Krok 6: Uživatel metodiky posuzuje úroveň dobré shody skutečných a zvoleným modelem predikovaných odpovědí žáků na testovou položku výpočtem odhadů tří statistik: (a) Yenova Q_1; (b) S-χ^2 statistiky; a (c) PV-Q1 statistiky. K těmto statistikám uživatel metodiky dále stanovuje jejich statistickou významnost a hodnotu RMSEA.
Výstup řešení modelové situace	
Výstupem řešení modelové situace je charakteristika testových položek prostřednictvím: (a) hodnot ukazatele obtížnosti testových položek vyjádřeného jako podíl žáků, kteří zodpověděli danou testovou položku správně; (b) hodnot ukazatele diskriminace testových položek vyjádřeného jako hodnota bodově biseriální korelace či upravené bodově biseriální korelace; (c) identifikace nabízených nesprávných odpovědí (distraktorů) na testové položky, které vykazují nepříznivé vlastnosti; (d) hodnoty změny ukazatelů spolehlivosti testu při vynechání testové položky; (e) zařazení testových položek do kategorií A, B a C vzhledem k úrovni DIF.	Výstupem řešení modelové situace je charakteristika testových položek prostřednictvím: (a) hodnot ukazatele obtížnosti testových položek vyjádřeného jako úroveň zvládnutí hodnoceného konstruktu, na které přibližně 50 % testovaných žáků odpovídá testovou položku správně; (b) hodnot ukazatele diskriminace testových položek vyjadřujícího rychlost změny pravděpodobnosti správné odpovědi žáka se změnou jeho úrovně zvládnutí hodnoceného konstruktu; (c) identifikace nabízených nesprávných odpovědí (distraktorů) na testové položky, které vykazují nepříznivé vlastnosti; (d) zařazení testových položek do kategorií A, B a C vzhledem k úrovni DIF; (e) statistické významnosti a hodnoty RMSEA statistik Yenova Q_1, S-χ^2 a PV-Q1.
Širší situační kontext	
Řešení modelové situace poskytuje komplexní pohled na kvalitu testových položek ověřovacího testu a umožňuje identifikovat testové položky, jejichž kvalitu je nutné hodnotit jako horší. V tomto ohledu platí: • V případě využití přístupů založených na CTT uživatel metodiky věnuje pozornost kvalitě testových položek charakteristických: (a) hodnotou ukazatele obtížnosti vyšší než 0,90 a nižší než 0,10 (0,20); (b) hodnotou ukazatele diskriminace nižší než 0,20, speciálně pak zápornou hodnotou ukazatele diskriminace; (c) přítomností nabízených nesprávných odpovědí (distraktorů) s nepříznivými vlastnostmi; (d) zvýšením spolehlivosti testu při jejich vynechání; a (e) zařazením do C kategorie vzhledem k úrovni DIF. • V případě využití přístupů založených na IRT uživatel metodiky věnuje pozornost kvalitě testových položek charakteristických: (a) hodnotou ukazatele obtížnosti vyšší než +2 (+3) a nižší než -2 (-3); (b) hodnotou ukazatele diskriminace nižší než 0,35; (c) přítomností nabízených nesprávných odpovědí (distraktorů) s nepříznivými vlastnostmi; (d) zařazením do C kategorie vzhledem k úrovni DIF; a (e) hodnotou RMSEA statistik Yenova Q1, S-χ2 a PV-Q1 vyšší než 0,05.	

Uživatel metodiky využívá uvedené informace při vyhodnocení ověřovacího testu, kdy především rozhoduje o způsobu vyhodnocení nekvalitních testových položek, a dále pak při utváření nových testů, kdy mu informace o kvalitě testových položek umožňují rozhodovat o nejvhodnější podobě ověřovacího testu. Takto může být například záměr cílit obtížnost ověřovacího testu na stanovenou úroveň zvládnutí hodnoceného konstruktu reflektován zařazením vyššího počtu testových položek s odpovídající obtížností.

Analogicky k vykreslení informační křivky testu statistické gramotnosti lze vykreslit i informační křivku testových položek (viz modelová situace „Spolehlivost (škály) ověřovacího testu“).

Software a ilustrace jeho využití pro řešení modelové situace

R balíček CTT (viz Willse, 2018)	R balíček ltm (viz Rizopoulos, 2018)
R balíček psych (viz Revelle, 2020)	R balíček difR (viz Magis, Beland a Raiche, 2020)
R balíček difR (viz Magis, Beland a Raiche, 2020)	R balíček mirt (Chalmers, 2020)

Řešení modelové situace je shodné s řešením souvisejících modelových situací. Jedinou výjimkou je hodnocení úrovně dobré shody skutečných a zvoleným modelem predikovaných odpovědí žáků na testovou položku.

Přístup IRT pro hodnocení úrovně dobré shody skutečných a zvoleným modelem predikovaných odpovědí žáků na testovou položku pro datový rámec TACR_DATA

```
PL2model <- mirt(TACR_DATA, 1)
```

 # Funkce mirt (R balíček *mirt*) *vede k odhadu parametrů 2PL modelu, které jsou uloženy v objektu PL2model.*

 # Atribut 1 *vyjadřuje, že má být odhadován unidimenzionální 2PL model.*

```
itemfit(PL2model, fit_stats = "X2")
```

 # Funkce itemfit (R balíček *mirt*) *s atributem* fits_stats = "X2" *vede k zobrazení hodnot statistiky Yenova Q1, a to včetně p-hodnoty této statistiky a hodnoty RMSEA.*

```
itemfit(PL2model, fit_stats = "S_X2")
```

 # Funkce itemfit (R balíček *mirt*) *s atributem* fits_stats = " S_X2" *vede k zobrazení hodnot statistiky S-χ^2, a to včetně p-hodnoty této statistiky a hodnoty RMSEA.*

```
itemfit(PL2model, fit_stats = "PV_Q1")
```

 # Funkce itemfit (R balíček *mirt*) *s atributem* fits_stats = "PV_Q1" *vede k zobrazení hodnot statistiky PV-Q1, a to včetně p-hodnoty této statistiky a hodnoty RMSEA.*

5.2.3 Volba a využití škály pro stanovení výsledků žáků v ověřovacím testu

Název modelové situace

Volba a využití škály pro stanovení výsledků žáků v ověřovacím testu

Podstata modelové situace

Podstata modelové situace reaguje na základní záměr ověřovacího testování – získat informaci o úrovni zvládnutí hodnoceného konstruktu žáky, kteří se ověřovacího testování účastní. Záměrem uživatele metodiky v modelové situaci proto je: (a) zvolit škálu, na které bude úroveň zvládnutí hodnoceného konstruktu žáky hodnocena; a (b) vyhodnotit úroveň zvládnutí hodnoceného konstrukty žáky na zvolené škále (využití škály). Uživateli metodiky se při volbě škály nabízí řada možností, které mají své specifické předpoklady využití.

Metodický postup řešení modelové situace

Metodický postup řešení modelové situace vychází z možnosti využití různých škál pro posouzení úrovně zvládnutí hodnoceného konstruktu žáky, tj. pro stanovení výsledků žáků v ověřovacím testu. Postup řešení modelové situace se skládá z následujících kroků uživatele metodiky:

Krok 1: Uživatel metodiky vybírá, zda bude hodnotit výsledky žáků na škále s využitím odhadu modelu vycházejícího z IRT.

Ne (přístup CTT)

Krok 2: Uživatel metodiky volí podobu škály:

Krok 2a): Škála odpovídající dosaženému skóre žáka v testu

Krok 3a): Uživatel metodiky přiřadí každému žákovi jeho dosažené skóre v testu, tj. počet správně zodpovězených testových položek.

Krok 2b): Škála odpovídající procentuální úspěšnosti žáka v testu

Krok 3b): Uživatel metodiky přiřadí každému žákovi podíl správných odpovědí, kterých dosáhl v řešení testu s vyjádřením v %.

Společně pro oba metodické přístupy

Krok 4: Uživatel metodiky může transformovat škálu kroků 2a) a 3a), respektive 2b) a 3b) na:

(a) alternativní bodovou škálu s průměrnou hodnotou *n* bodů a směrodatnou odchylkou *m* bodů;

(b) škálu percentilového pořadí.

Pozn.: V případě přístupu CTT lze, podobně jako při využití přístupu IRT, zohlednit poznatky z posouzení narušení předpokladů pro vyhodnocení ověřovacího testu, především předpokladu unidimenzionality. Při narušení tohoto předpokladu lze zvolit vhodné opatření.

Ano (přístup IRT)

Krok 2: Uživatel metodiky ověřuje předpoklady pro vyhodnocení ověřovacího testu s využitím modelů IRT, především pak:

(a) předpoklad minimální požadované velikosti výběrového souboru žáků;

(b) předpoklad unidimenzionality testu hodnocený metodickým postupem modelové situace „Unidimenzionalita ověřovacího testu a počet konstruktů v něm obsažených“;

(c) předpoklad lokální nezávislosti testových položek hodnocený metodickým postupem modelové situace „Lokální nezávislost testových položek“.

Krok 3a): Uživatel metodiky rozhodl o splnění předpokladů uvedených v kroku 2 nebo přijímá opatření pro naplnění předpokladů pro odhad modelů v souladu s předpoklady kroku 2.

Krok 4a): Uživatel metodiky odhaduje parametry: (a) 1PL modelu; (b) 2PL modelu; (c) 3PL modelu.

Krok 5a): Uživatel metodiky vybírá nejvhodnější model v souladu s metodickým postupem modelové situace „Výběr nejvhodnějšího modelu vycházejícího z IRT a hodnocení úrovně dobré shody dat modelu a ověřovacího testu“.

Krok 6a): Uživatel metodiky extrahuje hodnoty úrovně zvládnutí hodnoceného konstruktu žáky (Θ) z odhadů parametrů modelu vybraného v kroku 5a).

Krok 3b): Uživatel metodiky rozhodl o nesplnění předpokladů uvedených v kroku 2 a odhaduje multidimenzionální model vycházející z IRT.

Krok 4b) Uživatel metodiky odhaduje parametry multidimenzionálního modelu vycházejícího z paradigmatu IRT, a to pro počet faktorů stanovených v souladu s metodickým postupem modelové situace „Unidimenzionalita ověřovacího testu a počet konstruktů v něm obsažených“.

Krok 5b) Uživatel metodiky extrahuje hodnoty úrovně zvládnutí hodnocených konstruktů žáky (Θ_i) z parametrů multidimenzionálního modelu odhadovaných v kroku 4b).

	Společně pro metodické přístupy (a) a (b) Krok 7a) / 6b): Uživatel metodiky může transformovat škály kroků 6a) a 5b) na: (a) alternativní bodovou škálu s průměrnou hodnotou *n* bodů a směrodatnou odchylkou *m* bodů; (b) škálu percentilového pořadí. *Krok 3c): Uživatel metodiky rozhodl o nesplnění předpokladů uvedených v kroku 2 a hodnotí výsledky na škále s využitím přístupu CTT.*
Výstup řešení modelové situace	
Výstupem řešení modelové situace jsou hodnoty odpovídající úrovni zvládnutí hodnoceného konstruktu žáky: (a) skóre; (b) procentuální úspěšnost; (c) počet bodů na alternativní bodové škále; (d) percentilové umístění.	Výstupem řešení modelové situace jsou hodnoty odpovídající úrovni zvládnutí hodnoceného konstruktu žáky (Θ), s případným vyjádřením v podobě: (a) počtu bodů na alternativní bodové škále; a (b) percentilového umístění.
Širší situační kontext	
Při volbě škály pro stanovení výsledků žáků v ověřovacím testu je žádoucí posoudit naplnění předpokladů správnosti odhadů s danou škálou souvisejících (např. předpoklad unidimenzionality). Při volbě nevhodnější škály vycházející z odhadů modelů založených na IRT lze pro rozhodnutí využít postup hodnocení dobré shody empirických dat a dat modelových. Využití škál vycházejících z odhadů multidimenzionálních modelů založených na IRT má úzkou vazbu na hodnocení optimálního počtu konstruktů (dimenzí, faktorů) v ověřovacím testu obsažených. Volba škály pro stanovení výsledků žáků v ověřovacím testu, tj. úrovně zvládnutí hodnoceného konstruktu, může být spojena s odlišnými závěry, a to především při využití odhadů komplexnějších modelů, které například zohledňují schopnost testových položek diferencovat mezi žáky podle úrovně zvládnutí hodnoceného konstruktu (např. 2PL a 3PL model). Takto žák dosahující lepšího výsledku na škále procentuální úspěšnosti řešení testových položek testu může dosáhnout horšího výsledku na škále vycházející z odhadu 2PL modelu. Tato skutečnost se může projevit také v dalších modelových situacích, v nichž se pracuje s výsledky žáků, jako je například záměr identifikovat faktory, které ovlivňují úrovně zvládnutí hodnoceného konstruktu žáky. Moderní přístupy k reportingu výsledků žáků v ověřovacím testování se snaží spíše vyhýbat škálám založeným na skóre žáků či procentuální úspěšnosti.	
Software a ilustrace jeho využití pro řešení modelové situace	
R balíček *CTT* (viz Willse, 2018)	R balíček *ltm* (viz Rizopoulos, 2018) R balíček *CTT* (viz Willse, 2018) R balíček *mirt* (viz Chalmers, 2020)
Výsledky žáků v ověřovacím testu na škálách vycházejících z přístupů CTT lze snadno stanovit s využitím standardního software. Z tohoto důvodů se ilustrace řešení modelové situace zaměřuje na přístupy IRT. **Přístup IRT pro datový rámec TACR_DATA – 1PL model** PL1model <- rasch(TACR_DATA) *# Funkce* rasch (R balíček *ltm*) *vede k odhadu parametrů 1PL modelu, které jsou uloženy v objektu PL1model.*	

```
PL1model_vzor_skore <- factor.scores(PL1model, method = "EAP")

  # Funkce factor.scores (R balíček ltm) vede k odhadu úrovně zvládnutí hodnoceného konstruktu (Θ) ve vazbě na vzor
  odpovědí na testové položky, přičemž využita jsou data z objektu PL1model odhadovaného v předchozím kroku
  metodického postupu. Výsledky jsou uloženy v objektu PL1model_vzor_skore.

  # Funkce method umožňuje volit metodu odhadu úrovně zvládnutí hodnoceného konstruktu (Θ), v tomto případě se
  jedná o odhad založený na střední hodnotě aposteriorního rozdělení (EAP). Pro přístup založený na Bayesovském
  modálním odhadu (MAP) je volena možnost method = "EB".

PL1theta <- PL1model_vzor_skore$score.dat

  # Příkaz vede k „vytažení“ hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové
  položky a jejich uložení v objektu PL1theta.

write.csv(PL1theta, "score_PL1_EAP.csv")

  # Příkaz vede k uložení hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové
  položky a k jejich uložení v csv souboru score_PL1_EAP.
```

Data souboru score_PL1_EAP.csv, která spojují vzory odpovědí s příslušnou úrovní zvládnutí hodnoceného konstruktu (*Θ*), jsou využita pro přiřazení hodnoty zvládnutí hodnoceného konstruktu každému žákovi.

Přístup IRT pro datový rámec TACR_DATA – 2PL model

```
PL2model <- ltm(TACR_DATA ~ z1)

  # Funkce ltm (R balíček ltm) vede k odhadu parametrů 2PL modelu, které jsou uloženy v objektu PL2model.

  # Parametr z1 odpovídá úrovni zvládnutí hodnoceného konstruktu.

PL2model_vzor_skore <- factor.scores(PL2model, method = "EAP")

  # Funkce factor.scores (R balíček ltm) vede k odhadu úrovně zvládnutí hodnoceného konstruktu (Θ) ve vazbě na vzor
  odpovědí na testové položky, přičemž využita jsou data z objektu PL2model odhadovaného v předchozím kroku
  metodického postupu. Výsledky jsou uloženy v objektu PL2model_vzor_skore.

  # Atribut method umožňuje volit metodu odhadu úrovně zvládnutí hodnoceného konstruktu (Θ), v tomto případě se
  jedná o odhad založený na střední hodnotě aposteriorního rozdělení (EAP). Pro přístup založený na Bayesovském
  modálním odhadu (MAP) je volena možnost method = "EB".

PL2theta <- PL2model_vzor_skore$score.dat

  # Příkaz vede k „vytažení“ hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové
  položky a jejich uložení v objektu PL2theta.

write.csv(PL2theta, "score_PL2_EAP.csv")

  # Příkaz vede k uložení hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové
  položky a k jejich uložení v csv souboru score_PL2_EAP.
```

Data souboru score_PL2_EAP.csv, která spojují vzory odpovědí s příslušnou úrovní zvládnutí hodnoceného konstruktu (*Θ*), jsou využita pro přiřazení hodnoty zvládnutí hodnoceného konstruktu každému žákovi.

Přístup IRT pro datový rámec TACR_DATA – 3PL model

```
PL3model <- tpm(TACR_DATA)

  # Funkce tpm (R balíček ltm) vede k odhadu parametrů 3PL modelu, které jsou uloženy v objektu PL3model.

PL3model_vzor_skore <- factor.scores(PL3model, method = "EAP")
```

 # Funkce factor.scores (R balíček *ltm*) *vede k odhadu úrovně zvládnutí hodnoceného konstruktu (Θ) ve vazbě na vzor odpovědí na testové položky, přičemž využita jsou data z objektu PL3model odhadovaného v předchozím kroku metodického postupu. Výsledky jsou uloženy v objektu PL3model_vzor_skore.*

 # Atribut method *umožňuje volit metodu odhadu úrovně zvládnutí hodnoceného konstruktu (Θ), v tomto případě se jedná o odhad založený na střední hodnotě aposteriorního rozdělení (EAP). Pro přístup založený na Bayesovském modálním odhadu (MAP) je volena možnost method = "EB".*

```
PL3theta <- PL3model_vzor_skore$score.dat
```

 # Příkaz vede k „vytažení" hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové položky a jejich uložení v objektu PL3theta.

```
write.csv(PL3theta, "score_PL3_EAP.csv")
```

 # Příkaz vede k uložení hodnot úrovně zvládnutí hodnoceného konstruktu (Θ) pro všechny vzory odpovědí na testové položky a k jejich uložení v csv souboru score_PL3_EAP.

Data souboru score_PL3_EAP.csv, která spojují vzory odpovědí s příslušnou úrovní zvládnutí hodnoceného konstruktu (*Θ*), jsou využita pro přiřazení hodnoty zvládnutí hodnoceného konstruktu každému žákovi.

Přístup IRT pro datový rámec TACR_DATA – multidimenzionální model

```
MULTIDIMmodel <- mirt(TACR_DATA, 2, itemtype = "2PL")
```

 # Funkce mirt (R balíček *mirt*) *vede k odhadu multidimenzionálního modelu vycházejícího z paradigmatu IRT, přičemž číslovka uvádí, kolika dimenzionální model má být odhadován, tj. zde 2 dimenze (konstrukty, faktory).*

 # Atribut itemtype *umožňuje volit odhadovaný model založený na IRT, v tomto případě se jedná o 2PL model.*

```
summary(MULTIDIMmodel)
```

 # Obecná funkce summary *vede k zobrazení faktorových zátěží testových položek k dimenzím (konstruktům, faktorům) multidimenzionálního modelu.*

```
theta <- fscores(MULTIDIMmodel, method = "EAP")
```

 # Funkce fscores (R balíček *mirt*) *vede k odhadům úrovně zvládnutí hodnocených konstruktů žáky (Θ), přičemž využita jsou data z objektu MULTIDIMmodel odhadovaného v předchozím kroku metodického postupu. Výsledky jsou uloženy v objektu theta.*

 # Atribut method *umožňuje volit metodu odhadu úrovně zvládnutí hodnoceného konstruktu (Θ), v tomto případě se jedná o odhad založený na střední hodnotě aposteriorního rozdělení (EAP). Pro přístup založený na Bayesovském modálním odhadu (MAP) je volena možnost method = "MAP", pro přístup založený na maximální věrohodnosti je volena možnost method = "ML".*

```
write.csv(theta, "score_MULDITIM_EAP.csv")
```

 # Příkaz vede k uložení hodnot úrovně zvládnutí hodnocených konstruktů žáky (Θ) v csv souboru score_MULTIDIM_EAP.

Přístup CTT/IRT pro datový rámec TACR_DATA – alternativní bodová škála

```
MULTIDIMmodel <- mirt(TACR_DATA, 1, itemtype = "2PL")
```

```
theta <- fscores(MULTIDIMmodel, method = "EAP")
```

 # Tyto dva příkazy replikují odhad multidimenzionálního modelu s tím, že v tomto případě je voleno řešení s jednou dimenzí (konstruktem, faktorem). Výstupem příkazů jsou hodnoty úrovně zvládnutí hodnocených konstruktů žáky (Θ).

```
theta_transform <- score.transform(theta, 500, 100)
```

 # Funkce score.transform (R balíček *CTT*) *vede k transformaci hodnot úrovně zvládnutí hodnoceného konstruktu žáky (Θ) uložených v objektu theta na alternativní hodovou škálu s průměrem 500 a směrodatnou odchylkou 100. Hodnoty úrovně zvládnutí hodnoceného konstruktu žáky (Θ) na alternativní bodové škále jsou uloženy v objektu theta_transform. Součástí tohoto objektu je také hodnota percentilového pořadí pro každého žáka.*

5.2.4 Unidimenzionalita ověřovacího testu a počet konstruktů v něm obsažených

Název modelové situace
Unidimenzionalita ověřovacího testu a počet konstruktů v něm obsažených
Podstata modelové situace
Podstata modelové situace vychází ze skutečnosti, že unidimenzionalita ověřovacího testu je jedním z hlavních předpokladů odhadů řady statistik a indexů spojených s vyhodnocením ověřovacích testů, především pak tradičních modelů vycházejících z IRT. Další konstrukty (dimenze, faktory) se mohou v ověřovacím testu vyskytovat například v situacích, kdy: (a) test není tvořen jedním hlavním konstruktem, ale dvěma odlišnými konstrukty, u nichž teorie předpokládala spojení v podobě hlavního konstruktu. (b) testové položky uchopují doplňující konstrukty související s konstruktem hlavním (např. čtenářská gramotnost žáků daná delším uvozujícím textem k matematickým úlohám); (c) žáci odpovídají na daný konstrukt odlišně vlivem svých psychologických procesů, jako je například motivace žáků; (d) je v datech přítomen faktor příslušnosti ke skupině žáků, tj. testové položky nejsou spravedlivé vůči různým skupinám žáků. Záměrem uživatele metodiky je proto posoudit, zda ověřovací test měří jen jeden hlavní konstrukt (dimenzi, faktor) a pokud nikoliv, stanovit počet konstruktů v testu obsažených a současně sílu konstruktu hlavního.
Metodický postup řešení modelové situace
Záměrem řešení modelové situace je vyhodnotit naplnění předpokladu unidimenzionality (výskyt jen jednoho hlavního hodnoceného konstruktu) ověřovacího testu a v případě narušení tohoto předpokladu stanovit optimální počet konstruktů v testu obsažených a současně také sílu konstruktu hlavního. Významnou motivací uživatele metodiky ke sledování tohoto postupu je ta skutečnost, že silné narušení předpokladu unidimenzionality ověřovacího testu zpochybňuje kvalitu dalších odhadů, které jsou na předpokladu unidimenzionality ověřovacího testu založeny (např. odhad tradičních modelů vycházejících z IRT). Pro řešení modelové situace může uživatel metodiky využít řadu různých metodických přístupů, které jsou zachyceny v následujících krocích: Krok 1: Uživatel metodiky vybírá metodický přístup pro hodnocení unidimenzionality ověřovacího testu.
Krok 1a): Uživatel metodiky počítá s ohledem na dichotomický charakter testových položek hodnoty tetrachorických korelací mezi nimi.
Krok 2a): Uživatel metodiky posuzuje vztahy v matici tetrachorických korelací, přičemž se zaměřuje na identifikaci a interpretaci vztahů testových položek s vysokými hodnotami tetrachorických korelací.

Krok 1b): Uživatel metodiky počítá hodnoty vlastních čísel (eigenvalues) faktorů (konstruktů, dimenzí), přičemž za tímto účelem vybírá vhodnou metodu faktorové analýzy. Doporučena je preference metody hlavních os.

Krok 2b): Uživatel metodiky volí optimální počet faktorů (konstruktů, dimenzí) prostřednictvím Kaiserova kritéria, kdy doporučený počet faktorů (konstruktů, dimenzí) odpovídá počtu faktorů s vlastním číslem vyšším než jedna.

Krok 3b): Uživatel metodiky volí optimální počet faktorů (konstruktů, dimenzí) na základě významného zlomu v sutinovém grafu (*scree plot*), kdy doporučený počet faktorů (konstruktů, dimenzí) je umístěn nad významným zlomem v grafu.

Krok 4b): Uživatel metodiky volí optimální počet faktorů (konstruktů, dimenzí) s využitím metody paralelní analýzy, která srovnává skutečné hodnoty vlastních čísel a simulované hodnoty vlastních čísel. Doporučený počet faktorů (konstruktů, dimenzí) je dán počtem faktorů, jejichž skutečná hodnota vlastního čísla je vyšší než simulovaná hodnota vlastního čísla.

Krok 1c): Uživatel metodiky počítá hodnoty VSS kritéria (metoda velmi jednoduché struktury) pro jím stanovený maximální počet faktorů a s rozlišením VSS komplexity 1 a více komplexní VSS komplexity 2.

Krok 2c): Uživatel metodiky volí optimální počet faktorů (konstruktů, dimenzí) podle nejvyšší hodnoty VSS kritéria komplexity 1 či 2.

Krok 1d): Uživatel metodiky počítá průměrné hodnoty korelací mimo hlavní diagonálu (MAP), přičemž v každém kroku odstraňuje vliv nejvýznamnějšího konstruktu (komponenty). Postup je opakován v počtu kroků, které odpovídají počtu testových položek (tzv. Velicerův MAP test).

Krok 2d): Uživatel metodiky volí doporučený počet faktorů (konstruktů, dimenzí) podle nejnižší hodnoty MAP.

Krok 1e): Uživatel metodiky hodnotí sílu narušení předpokladu unidimenzionality testu, tj. úroveň multidimenzionality testu, prostřednictvím výpočtu DETECT indexu.

Krok 2e): Uživatel metodiky posuzuje sílu narušení předpokladu unidimenzionality testu na následující škále:

- Hodnoty nižší než 0,1 (0,2) naznačují unidimenzionalitu testu.
- Hodnoty v intervalu 0,1 až 0,5 (0,2 až 0,4) naznačují slabou multidimenzionalitu testu.
- Hodnoty v intervalu 0,5 až 1,0 (0,4 až 1,0) naznačují středně silnou multidimenzionalitu testu.
- Hodnoty vyšší než 1,0 naznačují silnou multidimenzionalitu testu.

Krok 1f): Uživatel metodiky počítá ukazatele spolehlivosti testu ω_h a ω_t, kde ukazatel ω_h se vztahuje jen k hlavnímu faktoru (konstruktu, dimenzi) a ukazatel ω_t zohledňuje také vliv skupinových faktorů (konstruktů, dimenzí) dílčích testových položek.

Krok 2f): Pro další interpretaci výsledků viz modelová situace „Spolehlivost (škály) ověřovacího testu“

Výstup řešení modelové situace

Výstupem metodického přístupu a) jsou hodnoty tetrachorických korelací pro všechny dvojice testových položek. Shluky vysokých hodnot utváří podezření o narušení předpokladu unidimenzionality ověřovacího testu.

Výstupem metodického přístupu b) jsou hodnoty optimálního počtu konstruktů (faktorů, dimenzí) jednotlivých metodických přístupů (Kaiserovo kritérium, sutinový graf, paralelní analýza). Naplnění předpokladu unidimenzionality ověřovacího testu je zde spojeno s optimálním počtem jednoho konstruktu (faktoru, dimenze).

Výstupem metodického přístupu c) jsou hodnoty optimálního počtu konstruktů (faktorů, dimenzí) jednotlivých metodických přístupů (VSS komplexita 1 a VSS komplexita 2). Naplnění předpokladu unidimenzionality ověřovacího testu je zde spojeno s optimálním počtem jednoho konstruktu (faktoru, dimenze).

Výstupem metodického přístupu d) je hodnota optimálního počtu konstruktů (faktorů, dimenzí) podle hodnot MAP (Velicerův MAP test). Naplnění předpokladu unidimenzionality ověřovacího testu je zde spojeno s optimálním počtem jednoho konstruktu (faktoru, dimenze).

Výstupem metodického přístupu e) je hodnota DETECT indexu. Pro naplnění předpokladu unidimenzionality ověřovacího testu jsou žádoucí hodnoty DETECT indexu, které nenaznačují silnou multidimenzionalitu testu.

Výstupem metodického přístupu f) jsou hodnoty ukazatelů spolehlivosti ověřovacího testu ω_h a ω_t. Vysoké hodnoty ukazatele ω_t ukazují na možný vliv skupinových konstruktů (faktorů, dimenzí) dílčích testových položek narušujících předpoklad unidimenzionality ověřovacího testu.

Širší situační kontext

Hodnocení naplnění předpokladu unidimenzionality ověřovacího testu poskytuje přínosné informace pro poznání konstruktů, které ověřovací test měří, a to také konstrukty dané: (a) stejným uvozujícím textem testových položek; (b) nespravedlivým chováním testových položek vůči určité skupině žáků (DIF analýza); či (c) umístěním testových položek na konci ověřovacího testu. Existence takto utvářených konstruktů může být ukázána také hodnocením lokální nezávislosti testových položek

Hodnocení naplnění předpokladu unidimenzionality ověřovacího testu může poskytnout novou informaci o konstruktech v testu obsažených, ale také potvrdit ex-ante očekávání, která vzešla například z expertního posouzení. Alternativně lze aplikovat přístup založený na klastrové analýze testových položek ověřovacího testu.

Hodnocení unidimenzionality ověřovacího testu je významným nástrojem pro rozhodnutí, jakou škálu pro měření úrovně zvládnutí hodnoceného konstruktu zvolit, například zda preferovat přístupy, které vycházejí z CTT před přístupy, které vycházejí z IRT, případně zda dát přednost odhadu multidimenzionálních modelů.

Při narušení předpokladu unidimenzionality ověřovacího testu je potřeba s opatrností interpretovat řadu odhadů, především pak odhad hodnot ukazatelů spolehlivosti (např. Cronbachovo alfa).

V praxi se zjevně s dokonalými unidimenzionálními testy nesetkáváme, což je žádoucí vzít do úvahy při rozhodování o počtu konstruktů v ověřovacím testu obsažených. Vhodným nástrojem, který zohledňuje tuto myšlenku, je DETECT index.

Z praktického hlediska se jako vhodné jeví diskutovat závěry různých metodických přístupů k hodnocení unidimenzionality ověřovacího testu a ke stanovení optimálního počtu konstruktů v něm obsažených. Takto je například paralelní analýza citlivá k nadhodnocení optimálního počtu faktorů pro velké výběrové soubory žáků.

Software a ilustrace jeho využití pro řešení modelové situace

R balíček *psych* (viz Revelle, 2020); R balíček *sirt* (viz Robitzsch, 2020)

Tetrachorické korelace pro datový rámec TACR_DATA

```
tetrakorelace <- mixed.cor(TACR_DATA)$rho
```

 # Funkce mixed.cor (R balíček *psych*) *vede k výpočtu korelací, jejichž typ je stanoven na základě povahy vstupních dat, přičemž korelace jsou „vytaženy" a uloženy v objektu tetrakorelace.*

```
cor.plot(tetrakorelace)
```

 # Funkce cor.plot (R balíček *psych*) *vede k vykreslení grafu korelací dvojic testových položek a k jejich uložení v objektu tetrakorelace.*

Metodické postupy založené na výpočtu vlastních číslech pro datový rámec TACR_DATA

```
TACR_DATA_paralel <- fa.parallel(TACR_DATA, fm ="ml", fa ="fa", cor = "tet")
```

 # Funkce fa.parallel (R balíček *psych*) *vede jednak k vykreslení sutinového grafu (scree plot), jednak k výpočtu vlastních čísel faktorů faktorové analýzy datového rámce TACR_DATA a jednak k zobrazení optimálního počtu konstruktů (faktorů, dimenzí) paralelní analýzou. Výsledky jsou uloženy v objektu TACR_DATA_paralel.*

 # Atributy fm *a* fa *umožňují výběr podoby faktorové analýzy, zde je volena metoda maximální věrohodnosti (fm = "ML") a metoda hlavních osob (fa = "ML").*

 # Atribut cor *umožňuje vybrat typ korelací, které jsou využity ve faktorové analýze, zde tetrachorické korelace.*

TACR_DATA_paralel$fa.values

 # Příkaz vede k „vytažení" a zobrazení hodnot vlastních čísel (eigenvalues) z objektu TACR_DATA_paralel.

Tímto způsobem získáváme všechny potřebné informace pro řešení modelové situace prostřednictvím Kaiserova kritéria, sutinového grafu a metody paralelní analýzy.

VSS kritéria a Velicerův MAP test pro datový rámec TACR_DATA

vss(TACR_DATA, n=38, fm="ml", cor="tet")

 # Funkce vss (R balíček *psych*) *vede k odhadu hodnot kritérií VSS (VSS komplexity 1 a VSS komplexity 2) a MAP pro počet faktorů 1 až 38 (atribut* n *= 38). Zároveň je pro metody velmi jednoduché struktury (VSS) a Velicerova MAP testu stanoven optimální počet konstruktů (faktorů, dimenzí).*

 # Atribut fm *umožňuje výběr podoby faktorové analýzy, zde je volena metoda maximální věrohodnosti.*

 # Atribut cor *umožňuje vybrat typ korelací, které jsou využity ve faktorové analýze, zde tetrachorické korelace.*

Tímto způsobem získáváme všechny potřebné informace pro řešení modelové situace prostřednictvím metody velmi jednoduché struktury (VSS) a Velicerova MAP testu.

DETECT index pro datové rámce TACR_DATA a DATA_TACR_HIERCH

Skore <- TACR_DATA_HIEARCH$skore

 # Příkaz vede k „vytažení" hodnot skóre z datového rámce TACR_DATA_HIEARCH a uložení v objektu Skore.

expl.detect(TACR_DATA, Skore, nclusters=38)

 # Funkce expl.detect (R balíček *sirt*) *vede k výpočtu a zobrazení hodnoty DETECT indexu pro počet faktorů/klastrů 1 až 38 (atribut* nclusters *= 38).*

5.2.5 Volba vhodného modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu

Název modelové situace
Volba vhodného modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu
Podstata modelové situace
Podstata modelové situace vychází ze záměru uživatele metodiky vybrat vhodný model, který vychází z IRT, pro vyhodnocení ověřovacího testu. Postup spojený s výběrem takového modelu posuzuje, jak dobře model charakterizuje empirická data ověřovacího testu, respektive jak dobře mohly být odpovědi žáků na testové položky ověřovacího testu vygenerovány zvažovaným modelem. Žádoucí je v tomto ohledu vysoká úroveň shody dat ověřovacího testu a modelem generovaných dat. Modelovou situaci lze rovněž rozšířit na výběr nejvhodnějšího modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu, tj. modelu, který vykazuje nejvyšší úroveň shody empirických a modelem generovaných dat. Pokud mezi dvěma modely není v tomto ohledu identifikován významný rozdíl, je preferován jednodušší model.

Metodický postup řešení modelové situace

Modelová situace vychází ze záměru uživatele metodiky vyhodnotit ověřovací test s využitím vhodného modelu vycházejícího z IRT. Za tímto účelem je posuzována úroveň shody dat ověřovacího testu s daty, která jsou generována zvažovaným modelem, přičemž žádoucí je dobrá shoda empirických a modelových dat. Rozšířením modelové situace je volba nejvhodnějšího modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu. Preferován je model vykazující nejvyšší úroveň dobré shody empirických a modelových dat. Pokud mezi dvěma modely nejsou v tomto ohledu zaznamenány významné rozdíly, je dána přednost jednoduššímu z nich. Postup řešení modelové situace se následně skládá z následujících kroků uživatele metodiky pro dva výše uvedené záměry:

Záměr 1: Posouzení vhodnosti modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu

Krok 1: Uživatel metodiky odhaduje model vycházející z IRT v souladu s metodickým postupem modelové situace „Volba a využití škály pro stanovení výsledků žáků v ověřovacím testu".

Krok 2: Uživatel metodiky počítá úroveň dobré shody skutečných a modelem generovaných odpovědí žáků na testové položky ověřovacího testu prostřednictvím: (a) indexu RMSEA; a (b) indexu SRMSR.

Krok 3: Uživatel metodiky srovnává vypočtené hodnoty: (a) indexu $RMSEA_2$ a (b) indexu SRMSR s kritickou hodnotou 0,05. Vyšší hodnoty zamítají nulovou hypotézu o dobré shodě empirických a modelových dat, a nepotvrzují tak vhodnost posuzovaného modelu pro vyhodnocení ověřovacího testu.

Záměr 2: Volba nejvhodnějšího modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu

Krok 1: Uživatel metodiky odhaduje zvažované modely vycházející z IRT v souladu s metodickým postupem modelové situace „Volba a využití škály pro stanovení výsledků žáků v ověřovacím testu".

Krok 2: Uživatel metodiky počítá hodnoty indexů dobré shody pro modely odhadované v kroku 1, a to hodnoty:

(a) testu poměru věrohodností zahnízděných modelů (LRT);

(b) Akaikeho informačního kritéria (AIC);

(c) Bayesova informačního kritéria (BIC).

Krok 3: Uživatel metodiky vybírá jako nejvhodnější model: (a) se statisticky významnou změnou LRT; (b) s nejnižší hodnotou AIC; a (c) s nejnižší hodnotou BIC.

Výstup řešení modelové situace

Výstupem prvního záměru jsou hodnoty indexů RMSEA a RMSR, které umožňují posoudit vhodnost posuzovaného modelu vycházejícího z IRT pro vyhodnocení ověřovacího testu.

Výstupem druhého záměru jsou hodnoty tří indexů dobré shody: (a) LRT; (b) AIC; a (c) BIC, které umožňují vybrat nejvhodnější model vycházející z IRT pro vyhodnocení ověřovacího testu.

Širší situační kontext

Modelová situace úzce navazuje na odhad modelů vycházejících z IRT, neboť pomáhá při výběru nejvhodnějšího z nich. Klíčové principy jsou v tomto ohledu spojeny se zájmem o co nejlepší shodu vybraného modelu s odpověďmi žáků na testové položky ověřovacího testu, ale také o co nejjednodušší model vycházející z IRT.

Software a ilustrace jeho využití pro řešení modelové situace

R balíček *mirt* (viz Chalmers, 2020); R balíček *ltm* (viz Rizopoulos, 2018)

Postup řešení záměru 1 pro datový rámec TACR_DATA

```
PL2model <- mirt(TACR_DATA, 1)
```

 # Funkce mirt (R balíček *mirt*) *vede k odhadu parametrů 2PL modelu, které jsou uloženy v objektu PL2model.*

 # Atribut 1 *vyjadřuje, že má být odhadován unidimenzionální 2PL model.*

M2(MirtPL2model)

 # Funkce M2 (R balíček *mirt*) *vede k odhadu a zobrazení hodnot indexů RMSEA a SRMSR.*

Postup řešení záměru 2 pro datový rámec TACR_DATA

PL1model <- rasch(TACR_DATA)

 # Funkce rasch (R balíček *ltm*) *vede k odhadu parametrů 1PL modelu, které jsou uloženy v objektu PL1model.*

PL2model <- ltm(TACR_DATA ~ z1)

 # Funkce ltm (R balíček *ltm*) *vede k odhadu parametrů 2PL modelu, které jsou uloženy v objektu PL2model.*

 # Parametr z1 odpovídá úrovni zvládnutí hodnoceného konstruktu.

PL3model <- tpm(TACR_DATA)

 # Funkce tpm (R balíček *ltm*) *vede k odhadu parametrů 3PL modelu, které jsou uloženy v objektu PL3model.*

anova(PL1model, PL2model)

 # Funkce anova *vede k zobrazení hodnot LRT, AIC a BIC a rovněž statistické významnosti LRT pro objekty PL1model a PL2model.*

anova(PL2model, PL3model)

 # Funkce anova *vede k zobrazení hodnot LRT, AIC a BIC a rovněž statistické významnosti LRT pro objekty PL2model a PL3model.*

6. Závěr

Hlavním záměrem této knihy bylo na základě syntézy teoreticko-metodických východisek problematiky vyhodnocení ověřovacího testování v počátečním vzdělávání a průmětu těchto východisek v řešení modelových případových studií představit metodiku vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání, která je hlavním výstupem projektu „Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání a její aplikace v modelových případových studiích".

Teoreticko-metodická východiska knihy jsou zasazena do představení postupů vyhodnocení ověřovacích testů, které vycházejí jednak z CTT, jednak z IRT (viz tabulka č. 38 pro zachycení silných a slabých stránek obou přístupů). Zjištění teoreticko-metodické rešerše byla využita pro zpracování celkem dvaceti modelových případových studií, které jsou častou součástí vyhodnocení ověřovacích testů, přičemž osm z nich bylo řešeno přístupy vycházejícími z CTT a zbývajících dvanáct pak přístupy vycházejícími z IRT. Vlastní metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání pak je průmětem hlavních zjištění jednak z rešerše teoreticko-metodických východisek vyhodnocení ověřovacích testů, jednak ze zpracování vlastních modelových případových studií.

Tabulka č. 38: Silné a slabé stránky přístupů k vyhodnocení ověřovacích testů vycházejících z CTT a IRT

	CTT	IRT
Silné stránky	- Dobrá znalost a jednoduchost využívaných konceptů - Méně striktní předpoklady vyhodnocení ověřovacích testů - Možnost vyhodnocení ověřovacích testů s využitím běžného software	- Zohlednění specifických charakteristik testových položek, vysoký potenciál tvorby testů na míru - Při splnění předpokladů nezávislost výsledků na výběrovém souboru testovaných osob
Slabé stránky	- Omezené zohlednění specifických charakteristik testových položek - Závislost výsledků na výběrovém souboru testových osob	- Horší srozumitelnost a vyšší komplikovanost využívaných konceptů - Náročné předpoklady kladené na odhad modelů - Potřeba dobré shody empirických a modelových dat

Zdroj: vlastní zpracování na základě DeVellis (2006), De Champlain (2010), Thorpe a Favia (2012), Wang, Ma a Chen (2010), Toland (2014), Krishnan (2013), DeMars (2010), Edelen a Reeve (2007)

Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání je koncepčně založena na řešení tzv. modelových situací. Pro modelové situace je formulován společný tzv. obecný rámec, který je následně pro každou z nich rozveden v navazujícím rámci specifickém. Součástí obecného rámce modelové situace je mimo jiné představení její podstaty, která slouží uživateli metodiky k výběru těch modelových situací, které jsou vhodné pro naplnění jeho záměru vyhodnocení ověřovacího testu. Specifický rámec modelové situace pak obsahuje návodné postupy řešení, a to včetně ilustrace využití vhodného software na příkladech, které

jsou součástí čtvrté kapitoly této knihy. V tomto ohledu je metodika postavena na balíčcích programovacího jazyka R, který je charakteristický jednak výbornými možnostmi pro provádění pokročilých statistických analýz dat, jednak snadnou dostupností v rámci tzv. svobodné licence.

Z praktického hlediska je metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání využitelná na různých úrovních hodnocení: (a) úroveň žáka (např. utváření vzdělávacích plánů ve vazbě na identifikaci jeho silných a slabých stránek); (b) úroveň třídy či školy (např. výběr vhodných vzdělávacích strategií ve vazbě na identifikaci silných a slabých stránek žáků; využití informací pro tvorbu strategických dokumentů školy); a (c) úroveň systému či území (např. tvorba zpráv o stavu vzdělávacího systému v hodnocené oblasti). Ze své podstaty je metodika metodikou podpůrnou, neboť neposkytuje svým uživatelům jediný *one-size-fits-all* návod, který by měl být sledován za každé situace. Metodika naopak poskytuje pro vybrané modelové situace postupy, které její uživatel volí a sleduje s ohledem na své vlastní záměry a předpoklady hodnocení. Rozhodnutí o podobě využití metodiky tak přísluší jejímu uživateli.

7. Literatura a zdroje informací

ALBANO, A. D. (2018). *Package 'equate'*. [online]. Available from: <https://https://cran.r-project.org/web/packages/equate/equate.pdf>.

BALLOU, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy*, 4(4), 351-383.

BARTHOLOMEW, D. J., TZAMOURANI, P. (1999). The goodness of fit of latent trait models in attitude measurement. *Sociological Methods & Research*, 27(4), 525-546.

BATES, B. et al. (2020). *Package 'lme4'*. [online]. Available from: <https://cran.r-project.org/web/packages/lme4/lme4.pdf>.

BATTAUZ, M. (2018). *Package 'equateIRT'.* [online]. Available from: <https://https://cran.r-project.org/web/packages/equate/equateIRT.pdf>.

BECKER, G. (1992). Human capital and the economy. *Proceedings of the American Philosophical Society*, 136(1), 85–92.

BETEBENNER, D. (2009). Norm-and criterion-referenced student growth. *Educational Measurement: Issues and Practice*, 28(4), 42-51.

BETEBENNER, D. W., LINN, R. L. (2010). *Growth in Student Achievement: Issues of Measurement, Longitudinal Data Analysis, and Accountability*. Princeton: Educational Testing Service.

BINKLEY, M. et al. (2013). Defining twenty-first century skills. In *Assessment and Teaching of 21st Century Skills*. Berlin: Springer, 17-66.

BONIFAY, W. E. et al. (2015). When are multidimensional data unidimensional enough for structural equation modeling? An evaluation of the DETECT multidimensionality index. *Structural Equation Modeling: A Multidisciplinary Journal*, 22(4), 504-516.

BRAUN, H. I. (2005). *Using Student Progress to Evaluate Teachers: A Primer on Value-Added Models*. Princeton: Educational Testing Service.

BRENNAN, R. L. (2011). Generalizability theory and classical test theory. *Applied Measurement in Education*, 24(1), 1-21.

BRIGGS, D. C. (2013). Measuring growth with vertical scales. *Journal of Educational Measurement*, 50(2), 204-226.

BRIGGS, D. C., DOMINGUE, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics*, 38(6), 551-576.

BURGOS, J. G. (2010). Bayesian methods in psychological research: the case of IRT. *International Journal of Psychological Research*, 3(1), 163-175.

COOK, L. L., EIGNOR, D. R. (1991). IRT equating methods. *Educational Measurement: Issues and Practice*, 10(3), 37-45.

DE CHAMPLAIN, A. F. (2010). A primer on classical test theory and item response theory for assessments in medical education. *Medical Education*, 44(1), 109-117.

DE CHAMPLAIN, A. F. (2015). Best-fit model of exploratory and confirmatory factor analysis of the 2010 Medical Council of Canada Qualifying Examination Part I clinical decision-making cases. *Journal of Educational Evaluation for Health Professions*, 12(11), 1-7.

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

DeVELLIS. R. F. (2006). Classical test theory. *Medical Care*, 44(11/S3), 50-59.

DEDE, C. (2010). Comparing frameworks for 21st century skills. In *21st Century Skills*: *Rethinking How Students Learn*. Bloomington: Solution Tree Press, 51-75.

DONOGHUE, J. R., HOMBO, C. (2001). The distribution of an item-fit measure for polytomous items. In *Annual Meeting of the NCME*. Seattle: National Council on Measurement in Education.

DORANS, N. J., HOLLAND, P. W. (1992). *DIF Detection and Description: Mantel-Haenszel and Standardization.* New Jersey: Princeton.

DORANS, N. J., MOSES, T. P., EIGNOR, D. R. (2010). Principles and practices of test score equating. *ETS Research Report Series*, 2010, 2, 1-41.

EDELEN, M. O., REEVE, B. B. (2007). Applying item response theory (IRT) modeling to questionnaire development, evaluation, and refinement. *Quality of Life Research*, 16(1), 5-18.

ERSTAD, O., VOOGT, J. (2018). The twenty-first century curriculum: issues and challenges. In *Second Handbook of Information Technology in Primary and Secondary Education*. Berlin: Springer, 19-36.

FDE (2017). *Florida Standards Assessments 2016-2017. Annual Technical Report*. Tallahassee: Florida Department of Education.

FINCH, W. H., BOLIN, J. E., KELLEY, K. (2014). *Multilevel Modeling Using R*. Boca Raton: CRC Press.

FINCH, H., HABING. B. (2007). Performance of DIMTEST and NOHARM based statistics for testing unidimensionality. *Applied Psychological Measurement*, 31(4), 292–307.

FRYČ, J. et al. (2020). *Strategie vzdělávací politiky České republiky do roku 2030+*. Praha: Ministerstvo školství, mládeže a tělovýchovy.

GILLIES, D. (2017). Human capital theory in education. In *Encyclopedia of Educational Philosophy and Theory*. Berlin: Springer, 1053-1057.

GRAHAM, J. M. (2006). Congeneric and (essentially) tau-equivalent estimates of score reliability. *Educational and Psychological Measurement*, 66(6), 930–944.

HAMBLETON, R. K., JONES, R. W. (1993). Comparison of classical test theory and item response theory and their applications to test development. *Educational Measurement: Issues and Practice*, 12(3), 38-47.

HARWELL, M. R., BAKER, F. B., ZWARTS, M. (1988). Item parameter estimation via Marginal Maximum Likelihood and an EM algorithm: a didactic. *Journal of Educational Statistics*, 13(3), 243-271.

CHALMERS, R. P. (2012). Mirt: a multidimensional item response theory package for the R environment. *Journal of Statistical Software*, 48(6), 1-29.

CHALMERS, R. P., NG, V. (2017). Plausible-value imputation statistics for detecting item misfit. *Applied Psychological Measurement*, 41(5), 372-387.

CHALMERS, P. (2020). *Package 'mirt'*. [online]. Available from: <https://cran.r-project.org/web/packages/mirt/mirt.pdf>.

CHEN, J., DE LA TORRE, J., ZHANG, Z. (2013). Relative and absolute fit evaluation in cognitive diagnosis modeling. *Journal of Educational Measurement*, 50(2), 123-140.

KARAMI, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, 11(2), 59-76.

KERELUIK, K. et al. (2013). What knowledge is of most worth: teacher knowledge for 21st century learning. *Journal of Digital Learning in Teacher Education*, 29(4), 127-140.

KILMEN, S., DEMIRTASLI, N. (2012). Comparison of test equating methods based on item response theory according to the sample size and ability distribution. *Procedia-Social and Behavioral Sciences*, 46, 130-134.

KOEDEL, C., MIHALY, K., ROCKOFF, J. E. (2015). Value-added modeling: a review. *Economics of Education Review*, 47, 180-195.

KRISHNAN, V. (2013). *The Early Child Development Instrument (EDI): an Item Analysis Using Classical Test Theory (CTT) on Alberta's Data*. Edmonton: University of Alberta.

KUZNETSOVA, A. et al. (2020). *Package 'lmerTest'*. [online]. Available from: <https://cran.r-project.org/web/packages/lmerTest/lmerTest.pdf>.

LAMBERT, M. C. et al. (2018). The impact of English language learner status on screening for emotional and behavioral disorders: A differential item functioning (DIF) study. *Psychology in the Schools*, 55(3), 229-239.

LAMPRIANOU, I. (2007). *An Investigation into the Test Equating Methods Used During 2006, and the Potential for Strengthening Their Validity and Reliability*. Coventry: Qualifications and Curriculum Authority.

LIVINGSTON, S. A. (2014). *Equating Test Scores (Without IRT)*. New Jersey: Educational Testing Service.

LÜDECKE, D. (2020). *Package 'sjstats'*. [online]. Available from: <https://cran.r-project.org/web/packages/sjstats/sjstats.pdf>.

MAGIS, D., BELAND, S., RAICHE, G. (2020). *Package 'difR'*. [online]. Available from: <https://cran.r-project.org/web/packages/difR/difR.pdf>.

MAYDEU-OLIVARES, A. (2015). Evaluating fit in IRT models. In *Handbook of Item Response Theory Modeling: Applications to Typical Performance Assessment*. New York: Routledge, 111-127.

MAYDEU-OLIVARES, A., 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.

MAYDEU-OLIVARES, A., GARCÍA-FORERO, C. (2010). Goodness-of-fit testing. In *International Encyclopedia of Education*. Amsterdam: Elsevier, 190-196.

MICHAELIDES, M. P. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Practical Assessment Research & Evaluation*, 13(7), 1-16.

NANDAKUMAR, R. (1994). Assessing dimensionality of a set of item responses – comparison of different approaches. *Journal of Educational Measurement*, 31(1), 17-35.

NARAYANAN, P., SWAMINATHAN, H. (1994). Performance of the Mantel-Haenszel and simultaneous item bias procedures for detecting differential item functioning. *Applied Psychological Measurement*, 18(4), 315-328.

NAUMENKO, O. (2014). *Comparison of Various Polytomous Item Response Theory Modeling Approaches for Task-Based Simulation CPA Exam Data*. Greensboro: The University of North Carolina.

O'DWYER, L. M., PARKER, C. E. (2014). *A Primer for Analyzing Nested Data: Multilevel Modeling in SPSS Using an Example from a REL Study*. Connecticut: Regional Educational Laboratory Northeast & Islands (ED); National Center for Education Evaluation and Regional Assistance (ED); Education Development Center, Inc. (EDC)

O'MALLEY, K. et al. (2011). Making sense of the metrics: student growth, value-added models, and teacher effectiveness. *Bulletin*, 19, 1-4.

ORLANDO, M., THISSEN, D. (2000). Likelihood-based item-fit indices for dichotomous item response theory models. *Applied Psychological Measurement*, 24(1), 50–64.

ÖZDEMIR, B. (2015). A comparison of IRT-based methods for examining differential item functioning in TIMSS 2011 mathematics subtest. *Procedia-Social and Behavioral Sciences*, 174, 2075-2083.

REISE, S. (1990). A comparison of item- and person-fit methods of assessing model-data fit in IRT. *Applied Psychological Measurement*, 14(2), 127–137.

REVELLE, W. (2012). *An Introduction to Psychometric Theory with Applications in R*. Evanston: Northwestern University.

REVELLE, W. (2019). *Using R and the psych Package to Find ω*. Evanston: Northwestern University.

REVELLE, W. (2020). *Package 'psych'*. [online]. Available from: <https://cran.r-project.org/web/packages/psych/psych.pdf>.

RIZOPOULOS, D. (2018). *Package 'ltm'*. [online]. Available from: <https://cran.r-project.org/web/packages/ltm/ltm.pdf>.

ROBITZSCH, A. (2020). *Package 'sirt'*. [online]. Available from: <https://cran.r-project.org/web/packages/sirt/sirt.pdf>.

RUPP, A. P. (2005). Maximum likelihood item response theory estimation. In *Encyclopedia of Statistics in Behavioral Science*. Chichester: John Wiley, 1-7.

RYAN, J., BROCKMANN, F. (2009). *A Practitioner's Introduction to Equating with Primers on Classical Test Theory and Item Response Theory*. Washington: Council of Chief State School Officers.

SANSIVIERI, V., WIBERG, M., MATTEUCCI, M. (2018). A review of test equating methods with a special focus on IRT-based approaches. *Statistica*, 77(4), 329-352.

SCHAFER, W. D. (2006). Growth scales as an alternative to vertical scales. *Practical Assessment Research & Evaluation*, 11(4), 1-6.

STONE, C. A. (2000). Monte Carlo based null distribution for an alternative goodness-of-fit test statistic in IRT models. *Journal of Educational Measurement*, 37(1), 58-75.

STONE, C. A., ZHANG, B. (2003). Assessing goodness of fit of item response theory models: a comparison of traditional and alternative procedures. *Journal of Educational Measurement*, 40(4), 331-352.

TABACHNICK, B. G., FIDELL, L. S. (2007). *Using Multivariate Statistics*. Boston: Pearson Education.

TENDEIRO, J. N., MEIJER, R. R., NIESSEN, A. S. M. (2016). PerFit: an R package for person-fit analysis in IRT. *Journal of Statistical Software*, 74(5), 1-27.

TENDEIRO, J. N. (2018). *Package 'PerFit'*. [online]. Available from: <https://cran.r-project.org/web/packages/PerFit/index.html>.

THOMPSON, N. A. (2009). *Ability Estimation with Item Response Theory*. St. Paul: Assessment Systems Corporation.

THOMPSON, K. L. (2015). *Measuring Student Growth in K–12 Schools Using Item Response Theory within Structural Equation Models.* Hattiesburg: The University of Southern Mississippi.

THOMPSON, N. A. (2016). *Introduction to Classical Test Theory with CITAS*. Minnetonka: Assessment Systems Corporation.

THORPE, G. L., FAVIA, A. (2012). *Data Analysis Using Item Response Theory Methodology: an Introduction to Selected Programs and Applications*. Orone: The University of Maine.

TOLAND, M. D. (2014). Practical guide to conducting an item response theory analysis. *The Journal of Early Adolescence*, 34(1), 120-151.

TRAUB, R. E. (1997). Classical test theory in historical perspective. *Educational Measurement: Issues and Practice*, 16(4), 8-14.

UDOFIA, N. A., UKO, M. P. (2016). Vertical scaling in standards-based educational assessment and accountability in educational systems. *Journal of Research & Method in Education*, 6(4), 65-75.

VAN DER LINDEN, W. J. (2010). Item response theory. In *International Encyclopedia of Education*. Amsterdam: Elsevier, 81-88.

VENABLES, W. N. et al. (2020). *An Introduction to R*. R Core Team. [online]. Available from: <https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf>.

VOOGT, J., ROBLIN, N. P. (2012). A comparative analysis of international frameworks for 21st century competences: Implications for national curriculum policies. *Journal of Curriculum Studies*, 2012, 44(3), 299-321.

WANG, H., MA, C., CHEN, N. (2010). A brief review on item response theory models-based parameter estimation methods. In *The 5th International Conference on Computer Science & Education*. Hefei: National Research Council of Computer Education in Colleges & Universities, 19-22.

WIBERG, M. (2007). *Measuring and Detecting Differential Item Functioning in Criterion-Referenced Licensing Test*. Umeå: Umeå University.

WILLSE, J. T. (2018). *Package 'CTT'*. [online]. Available from: <https://cran.r-project.org/web/packages/CTT/index.html>.

WIRTH, R. J., EDWARDS, M. C. (2007). Item factor analysis: current approaches and future directions. *Psychological Methods*, 12(1), 58-79.

YAVUZ, S., et al. (2018). The comparison of differential item functioning predicted through experts and statistical techniques. *Cypriot Journal of Educational Sciences*, 13(2), 375-384.

ZIEGLER, M., HAGEMANN, D. (2015). Testing the unidimensionality of items. *European Journal of Psychological Assessment*, 31(4), 231-237.

ZHANG, B. (2008). Application of unidimensional item response models to tests with items sensitive to secondary dimensions. *The Journal of Experimental Education*, 77(2), 147-166.

ZHANG. J. (2013). A procedure for dimensionality analyses of response data from various test designs. *Psychometrika*, 78(1), 37-58.

8. Seznam obrázků

9. Seznam tabulek

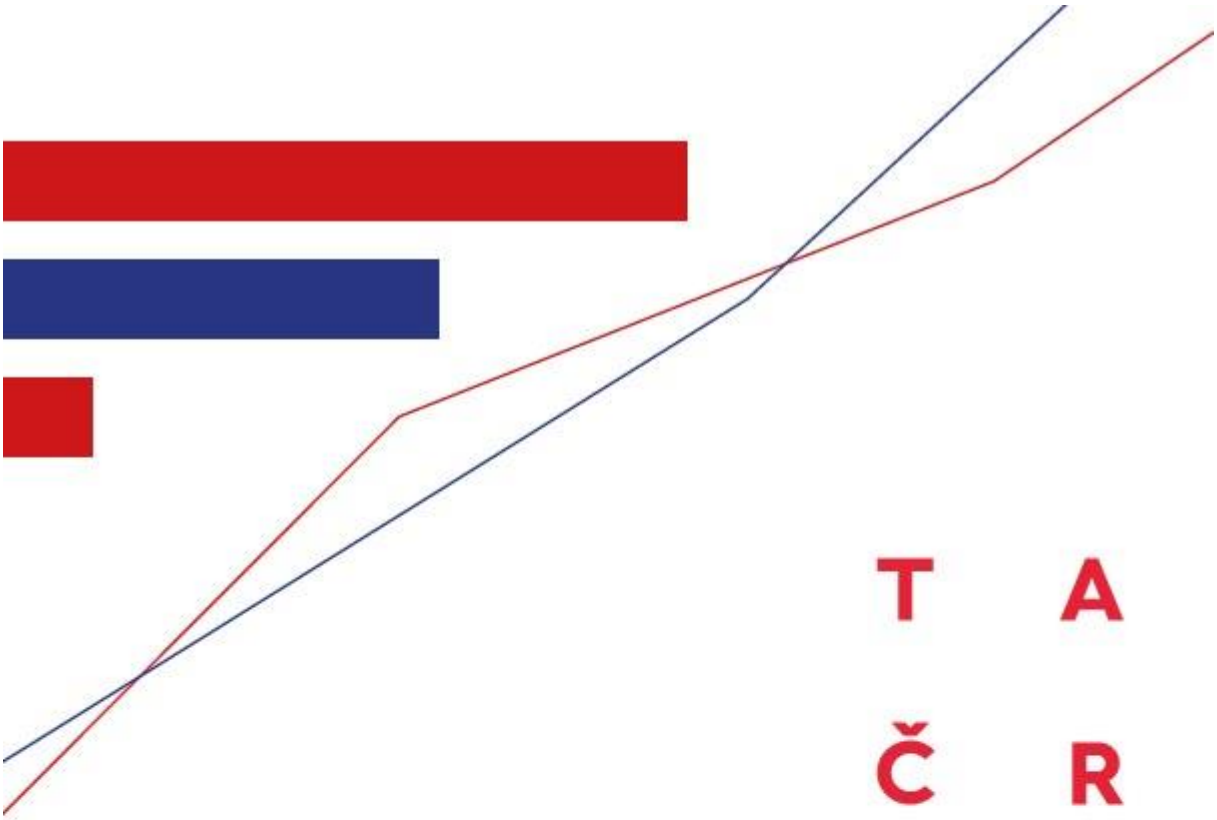
10. SUMMARY

This book is an output of the TACR (Technology Agency of the Czech Republic) funded research project, which investigated solutions to common tasks related to test-based assessment and evaluation in primary and secondary education. The intent of this book, which is consistent with the project objectives, is: (a) to summarize theoretical and methodological perspectives, providing insight into test-based assessment and evaluation; (b) to illustrate how these theoretical and methodological perspectives are used in solving particular case studies; and (c) to introduce how the two elements are synthesized in a methodology that contains practical and usable instructions how to perform tasks common to test-based assessment and evaluation (case studies).

In the first part of the book, classical test theory and item response theory are explained. These are two main theoretical approaches applied in test-based assessment and evaluation, and this is why our methodology emanates just from them. Reflecting this, we carried out an extensive literature review on classical test theory (CTT) and item response theory (IRT) in order to: (a) identify the tasks (case studies) common to test-based assessment and evaluation; and (b) give step-by-step instructions how to perform these tasks (case studies). In sum, twenty case studies were identified and elaborated (the second part of the book), with eight of them being rooted in CTT and the remaining twelve in IRT approaches. In the third part of the book, the methodology is introduced.

The methodology follows a case-study approach conceptually consisted of two parts. A general framework common to all case studies is built firstly. The following components of each case study are described under this framework: (i) title; (ii) brief description; (iii) step-by-step instructions how to perform the case study by applying either CTT or IRT approaches; (iv) output presentation; (v) broader situational context; and (vi) relevant software and its illustrative use. These components are specified separately for each case study. In sum, eighteen case studies are described in the second part of the methodology. Five of them are presented in this book, particularly: (a) test (scale) reliability; (b) item quality and detection of low-quality items; (c) test taker's proficiency level – test scale; (d) test unidimensionality and the number of inherent dimensions (domains); and (e) the most appropriate IRT model for test assessment and evaluation.

Three principles of the methodology are noteworthy. Firstly, it is the user who decides how the methodology is used. A brief description of each case study is provided to assist the users in their search for appropriate case studies. Secondly, the methodology is open to changes, including adaptation of the component items and addition of new case studies. Thirdly, several R packages are used to solve the case studies. The choice of the R environment was motivated by its flexibility for advanced statistical analysis and easy accessibility (General Public License).



Tato kniha byla zpracována jako výstup řešení projektu Technologické agentury České republiky číslo TL01000385 s názvem „Metodika vyhodnocení výsledků ověřovacího testování v počátečním vzdělávání a její aplikace v modelových případových studiích“, a to v rámci programu TL – Program na podporu aplikovaného společenskovědního a humanitního výzkumu, experimentálního vývoje a inovací ÉTA. Řešitelé projektu děkují Technologické agentuře České republiky za finanční podporu při řešení projektu.

V roce 2021 vydala:
NEWTON Academy; 5. května 1640/65; 140 21 Praha

ISBN 978-80-87325-39-1