Srovnatelnost testových verzí slovenské maturitní zkoušky z anglického jazyka

Martina Hulešová

Masarykova univerzita

MUNI
PRESS

Srovnatelnost testových verzí slovenské maturitní zkoušky z anglického jazyka

Martina Hulešová

Masarykova univerzita, Brno 2021

KATALOGIZACE V KNIZE – NÁRODNÍ KNIHOVNA ČR

Hulešová, Martina
Srovnatelnost testových verzí slovenské maturitní zkoušky z anglického jazyka / Martina Hulešová. -- 1. vydání. -- Brno : Masarykova univerzita, 2021. -- 1 online zdroj. -- (Cizí jazyky a jejich didaktiky : teorie, empirie, praxe ; svazek 10)
Anglické resumé
Obsahuje bibliografii a bibliografické odkazy

ISBN 978-80-210-9950-0 (online ; pdf)

* 37.012 * 37.091.26/.27 * 811.111 * 373.5.091.274/.275 * (437.6) * (048.8)
– pedagogický výzkum
– didaktické testy
– angličtina
– maturitní zkoušky -- Slovensko
– monografie

811.111 - Angličtina [11]

Edice: Cizí jazyky a jejich didaktiky: teorie, empirie, praxe
Svazek 10

Recenzenti: doc. PhDr. Lucie Betáková, M.A., Ph.D.
prof. PaedDr. Silvia Pokrivčáková, Ph.D.

ISBN 978-80-210-9950-0
ISBN 978-80-210-9949-4 (brož. vazba)
https://doi.org/10.5817/CZ.MUNI.M210-9950-2021

Obsah

Poděkování

Základem této knihy je disertační práce. Ta by však nevznikla bez mnoha osob, jimž chci touto cestou poděkovat. Děkuji tedy doc. PhDr. Renatě Povolné, Ph.D., za její trpělivý přístup a cenná doporučení, doc. PhDr. Michaele Píšové, M. A. Ph.D., za podnětné diskuze a doporučení, doc. Mgr. Františku Tůmovi, Ph.D., za metodologické rady a zpětnou vazbu, a prof. PhDr. Věře Janíkové, Ph.D. a doc. Mgr. Světlaně Hanušové, Ph.D., za množství inspirace, skvělou zpětnou vazbu a neumdlévající podporu. V neposlední řadě děkuji kolegům z Institutu výzkumu školního vzdělávání a z Fakulty sociálních studií za cenné diskuse a rady, které vždy přišly v pravý okamžik.

Úvod

Testování, hodnocení a interpretace výsledků testování jsou dnes již běžnou součástí lidských činností. Pro jakýkoli kontext, kde se uplatňuje testování (průmyslovou výrobu, testování nových přístrojů a softwarů, vývoj nových farmaceutických přípravků a testování léků, psychologické a klinické testování a měření i pro vzdělávání a běžné školní testy, maturitní zkoušku, přijímací zkoušky apod.) platí stejné principy: smyslem testování je získání užitečných, přesných a spolehlivých informací o sledované charakteristice nebo populaci; interpretace výsledků se vždy musí vztahovat k účelu, kvůli kterému byl testovací nástroj nebo postup vytvářen; závěry jsou využívány pro rozhodovací procesy, a čím významnější jsou důsledky rozhodnutí činěných na základě výsledků testů, tím důležitější je naplňovat standardy oboru a jejich naplňování prokazovat.

V kontextu vzdělávacího systému existují testy, resp. zkoušky tzv. nízké a vysoké důležitosti (*low-stakes* a *high-stakes tests*), podle toho, do jaké míry jejich výsledky trvale ovlivňují život testovaných. Mezi zkoušky vysoké důležitosti patří např. přijímací zkoušky na střední a vysoké školy, maturitní zkouška, státní závěrečná zkouška, atestační a certifikační zkoušky opravňující k výkonu povolání např. lékaře, pilota apod., kde je naprosto zásadní, aby zkouška prokazatelně naplňovala principy a standardy obecně testologické i standardy daného oboru a informovala o nich odpovídajícím způsobem všechny důležité uživatele výsledků zkoušek (*stakeholders*).

Srovnatelnost testových verzí, respektive ekvivalence skórů ve zkouškách vysoké důležitosti je jednou z klíčových podmínek pro smysluplnou a spravedlivou interpretaci výsledků a pro jejich spravedlivé využití při rozhodování o testovaných na všech úrovních: individuální, institucionální i systémové. Je také klíčovým aspektem validity, chápeme-li ji jako spravedlivou a smysluplnou interpretaci testových skórů (Messick, 1995). Zejména u zkoušek vysoké důležitosti, jejichž výsledek ovlivňuje závažným způsobem budoucnost testovaných, jsou

poskytovatelé testu povinni přinejmenším z pohledu obecně přijímaných Standardů pro pedagogické a psychologické testování (AERA, APA, & NCME, 2014) a případně jiných kodexů či principů dobré praxe prokázat kvalitu testového nástroje včetně srovnatelnosti testových verzí a dokumentovat a zpřístupnit informace o procesech a opatřeních, jejichž pomocí bylo srovnatelnosti testových verzí dosaženo.

Obor jazykového testování operuje s několika úhly pohledu na kvalitu nástrojů používaných pro testování. Organizace sdružující jazykové testery obvykle mívají standardy profese, které jsou postulovány v kodexech dobré praxe a v etických kodexech. Dále existují všeobecně akceptované standardy pro pedagogické a psychologické testování, jež vymezují profesně odpovědné jednání a postupy. Využívány jsou teoretické modely popisující předpoklady a postupy pro validaci testových nástrojů. Žádný z výše zmíněných instrumentů či postupů není závazný či vynutitelný, přesto je jejich dodržování v oboru jazykového testování považováno za eticky odpovědné chování a za dobrou praxi a instituce nebo poskytovatelé testu obvykle považují za prestižní, pokud zmíněné principy naplňují a pokud zároveň mohou jejich naplňování prokázat např. ve výzkumných zprávách nebo v odborných časopisech.

V mnoha evropských zemích proběhla v posledních několika desetiletích kurikulární reforma, která mimo jiné zavedla i více či méně centralizovanou zkoušku ověřující výsledky vzdělávání. Mezi tyto země patří také Slovenská republika, která do systému vzdělávání zavedla maturitní zkoušku ukončující středoškolské vzdělání. Slovenská maturitní zkouška je zkouškou částečně centralizovanou – má část obsahující zkoušky spadající do kompetence škol a také centralizovanou externí část, kterou má v gesci stát. Externí část je standardizovaná: zkoušky vznikají na základě stejných specifikací, jsou administrovány a vyhodnocovány za stejných podmínek pro všechny testované, výsledky jsou interpretovány stejným způsobem.

Jde o zkoušku vysoké důležitosti, která je dokladem o úspěšném dokončení středoškolského vzdělání na čtyřletých maturitních vzdělávacích oborech. Výsledky maturitní zkoušky zároveň slouží jako klíčová informace pro přijímací řízení na vysoké školy. Proto je nanejvýš důležité, aby výsledky takto zásadní zkoušky byly spolehlivé, umožňovaly porovnání v rámci jedné maturitní populace i napříč maturitními populacemi – ročníky. Znamená to také, že testové verze by měly být verzemi srovnatelnými (podle některých zdrojů ekvivalentní nebo paralelní):

musí jimi být měřen stejný konstrukt stejným způsobem při každém řádném, opravném nebo náhradním zkušebním termínu a výsledky, tj. skóry z různých testových verzí, musí být porovnatelné a musí být možné je interpretovat stejným způsobem.

Právě otázka, zda, jak a za jakých podmínek jsou testové verze porovnatelné, jaké procesy jsou nebo naopak nejsou a měly nebo mohly by být implementovány v průběhu vývoje zkoušek nebo ex-post, tedy po realizaci testování, byla předmětem našeho zkoumání. V této publikaci prezentujeme výzkum, který vycházel z reálného kontextu maturitní zkoušky ve Slovenské republice. V teoretické části jsme zjišťovali, jaké metody a postupy byly popsány v odborné literatuře (Část I, kapitoly 1–3), vybrané metody jsme v empirické části práce aplikovali na testové verze slovenské maturitní zkoušky realizované v letech 2012–2015 (Část II, kapitoly 4 a 5) a na základě empirických výsledků jsme se pokusili identifikovat ty metody nebo postupy, jež by byly aplikovatelné v kontextu slovenské maturitní zkoušky bez nutnosti upravovat zásadním způsobem stávající systém vývoje testových verzí nebo legislativní prostředí. V aplikační části III v závěrečné kapitole 6 kriticky reflektujeme omezení, které náš výzkum měl, a míru zobecnitelnosti závěrů (Část III, kapitola 6).

Domníváme se, že výsledky a doporučení, k nimž tento výzkum dospěl, by mohly napomoci na cestě za zvyšováním kvality při zjišťování výsledků vzdělávání na systémové úrovni a zároveň na individuální i institucionální úrovni přispějí ke spravedlivosti testování i k větší zjevné validitě zkoušek, tedy k tomu, jak by měla být maturitní zkouška jako zkouška vysoké důležitosti vnímána jak odbornou veřejností, tak především těmi uživateli výsledků zkoušek, na jejichž životy mají tyto zkoušky bezprostřední a zásadní vliv.

I. TEORETICKÁ ČÁST

1 Kontext výzkumného projektu

Výzkumný projekt byl motivován snahou odpovědět na otázky, zda a do jaké míry je otázka srovnatelnosti testových verzí a validity interpretací výsledků testových verzí zadávaných v různých zkušebních termínech řešena obecně a specificky v systému slovenské maturitní zkoušky. Původním záměrem bylo zkoumat tuto otázku na české maturitní zkoušce z anglického jazyka, avšak Centrum pro zjišťování výsledků vzdělávání zajišťující maturitní zkoušku v České republice odmítlo poskytnout data. Z tohoto důvodu jsme se obrátili na instituci zajišťující maturitní zkoušku ve Slovenské republice – NÚCEM, která data pro tento výzkum poskytla.

Výzkum se tedy zaměřil na způsoby řešení otázky srovnatelnosti na příkladu slovenské maturitní zkoušky z anglického jazyka na úrovni B1, konkrétně na testech receptivních dovedností použitých v jarních termínech let 2012–2015, a zjišťuje, jaké postupy by mohly být uplatněny, aby bylo možné srovnatelné testové verze vytvářet a jejich srovnatelnost prokazovat.

Srovnatelnost jako samozřejmost

Přestože je srovnatelnost testových verzí a ekvivalence skórů jednou z významných výzkumných otázek po celou dobu existence moderního testování (von Davier, 2011; Holland, 2007), ve středoevropském prostředí, konkrétně v českém a slovenském standardizovaném testování patří zatím k tématům spíše okrajovým (Anýžová, 2013), jež jsou řešena jen zřídka nebo nejsou systematicky řešena a dokumentována vůbec. Z veřejně přístupných dokumentů dostupných na stránkách poskytovatele slovenské maturitní zkoušky (Národný ústav certifikovaných meraní – NÚCEM)[1] o výsledcích maturitních zkoušek lze vyvodit,

1 www.nucem.sk – analytické zprávy vydávané každoročně po ukončení zkušebního období

že je srovnatelnost testových verzí předpokládána nebo přinejmenším není zpochybňována. Téma srovnatelnosti, ekvivalence, paralelnosti apod. testových verzí není v dostupných oficiálních dokumentech zmíněno. Pokud je nám známo, neexistuje nebo přinejmenším nebyla publikována ani žádná výzkumná studie nebo jiný text, jež by informovaly o tom, jak a zda vůbec je otázka srovnatelnosti testových verzí slovenské maturitní zkoušky řešena.

Příliš obecná specifikace testů

Veřejně dostupné informace o účelu, konstruktu a obsahu testů z anglického jazyka jsou velmi obecné, a tedy i otevřené široké interpretaci. NÚCEM explicitně deklaruje, že maturitní testy z anglického jazyka ověřují dosažení jazykové způsobilosti na úrovni B1 podle Společného evropského referenčního rámce (SERRJ, 2001, dále SERRJ). Chybí podrobné zdůvodnění tohoto tvrzení, odkaz na studie nebo zprávy z projektů, které by se zabývaly procesem přiřazení (*linking*, *alignment*) zkoušek k úrovni SERRJ apod. Jen velmi obecně je definován konstrukt všech tří subtestů (Poslech s porozuměním, Gramatika a Čtení s porozuměním[2]), a to velmi stručně a obecně formulovanými deskriptory, bez konkrétního odkazu na referenční škály SERRJ a v poněkud odlišném znění, než jsou deskriptory referenční úrovně B1. Z dostupných informací není zřejmé, zda existují podrobně zpracované specifikace testů, které by sloužily tvůrcům úloh a sestavovatelům testu. Studie o přiřazení maturitních testů k SERRJ prokazující tvrzení, že testy ověřují úroveň B1 SERRJ, nejsou dostupné, není tedy známo, jak a zda byly provedeny.

Společný evropský referenční rámec jako neměnný dokument

SERRJ jako takový je velmi užitečný dokument, pokud je s ním nakládáno v souladu s účelem, ke kterému byl vytvořen, tedy jako popisný rámec, jež má sloužit jako konzultační materiál při plánování cílů a obsahů jazykových programů, při definování a popisu obsahů jazykových zkoušek a plánování hodnoticích kritérií, jako podklad pro plánování autonomního učení, obecně tedy jako společná platforma při úvahách o vzdělávání a učení se cizím jazykům (Council of Europe, 2001). V souvislosti s vývojem jazykových zkoušek ale nestačí jen odkázat na úroveň SERRJ,

2 Dále jen Poslech, Gramatika a Čtení.

aby to znamenalo, že zkouška ověřuje danou úroveň, takové tvrzení je třeba prokázat. A zároveň, protože SERRJ není jazykově či kontextově specifický, je třeba jeho použití „lokalizovat" a modifikovat pro účely daného kontextu, a tyto úpravy zdokumentovat a zdůvodnit (*localization* – Bachman, 2005, Weir, 2005). Pokud je k SERRJ odkazováno např. ve specifikacích zkoušky nebo při definici účelu a konstruktu zkoušky, je nutné dokumentovat změny, odchylky a doplnění, případně provést studii, která by popsala a potvrdila vztah zkoušek nebo testů k SERRJ. Podrobněji se omezením SERRJ obecně, i konkrétně omezením SERRJ při vytváření srovnatelných testových verzí věnují např. Alderson a kol. (2006), Bachman a kol. (1995) nebo Weir (2005). Weir dokonce říká, že „pro nekritické použití v jazykovém testování není SERRJ ve stávající podobě dostatečně komplexní, koherentní ani transparentní"[3] (Weir, 2005, s. 281) a charakterizuje SERRJ jako dokument, který zdůrazňuje spíše jazykové funkce a opomíjí psycholingvistické aspekty související recepce a produkce řeči (Weir, 2005). Dále kriticky hodnotí čtyři oblasti SERRJ, které podle něj nejsou dostatečně zpracovány a při vývoji testů a zkoušek je třeba jejich dopracování a doplnění. Řečové činnosti obsažené v deskriptorech jsou podle něj jen zřídka vztaženy k očekávané kvalitě výkonu (což souvisí s aspektem validity skórů); formulace deskriptorů nejsou vždy konzistentní a transparentní napříč referenčními úrovněmi (což souvisí s aspektem validity obsahové); v referenčních škálách není stejnoměrně a koherentně popsán kontext těchto činností (což se týká validity kontextové); a téměř vůbec nejsou zmiňovány kognitivní procesy, které se při realizaci úkolů zapojují (validita konstruktová). Všechny tyto parametry by měly být ve specifikacích zpracovány.

Validační proces a spravedlivost

NÚCEM je povinen publikovat všechny testy v jejich úplnosti ihned po ukončení testování, což při případném výzkumu souvisejícím s validací maturitní zkoušky komplikuje jeho realizaci a omezuje nebo dokonce znemožňuje použití některých metod. Po každém termínu testování jsou NÚCEMem zveřejňovány zprávy se statistickými výstupy převážně deskriptivní povahy o úspěšnosti dané populace. Testy, úlohy a položky však nejsou analyzovány do hloubky, případné

3 „In its present form the CEFR is not suficiently comprehensive, coherent or transparent for uncritical use in language testing" (s. 281)

problematické položky jsou sice popsány, chybí však analýza příčin problémů. Nejsou dostupné studie dokumentující přiřazení k SERRJ, ani validační studie, včetně studií dokládajících postupy zajišťující srovnatelnost skórů testových verzí.

Z výše uvedeného vyplývají některé potenciální problémy. Pokud nejsou validita a ekvivalence skórů testových verzí podloženy empirickým výzkumem, pak mohou nastat tyto situace:

- Testové skóry z různých testových verzí mají meziročně různý význam (jinou interpretaci ve smyslu měřeného rysu), a nominálně shodné výsledky z různých zkušebních termínů proto nemohou být považovány za ekvivalentní;
- Rozhodnutí činěná na základě shodných nominálních hodnot testových skórů z různých testových verzí nemusí být meziročně konzistentní a mohou vést k závěrům s různými důsledky, a to i pro testované se stejnou úrovní měřeného rysu;
- Testovaní, kteří budou konat zkoušku v různých zkušebních termínech, mohou mít nerovné podmínky pro prokázání úrovně jazykové způsobilosti, neboť není zaručeno, že jsou testové verze stejně obtížně, nebo že jsou nastavena opatření, která by umožňovala skóry porovnávat;
- Spravedlivost a validita interpretace výsledků proto může být ohrožena.

Jak již bylo zmíněno v úvodu, zkoušky vysoké důležitosti, mezi které patří i maturitní zkouška, by měly naplňovat standardy oboru testování a principy dobré praxe (AERA, APA, & NCME, 2014; kodexy ALTE, EALTA, ILTA[4]), aby bylo možné konstatovat, že interpretace výsledků a jejich využití jsou v souladu s principy validity (Messick, 1995). V návaznosti na tyto etické aspekty použití testů vysoké důležitosti byly stanoveny následující výzkumné cíle a z nich plynoucí výzkumné otázky:

1.1 Výzkumné cíle

1. Zmapovat, jaké metody a postupy pro dosahování srovnatelnosti testových verzí existují a které z nich a z jakého důvodu jsou používány institucemi poskytujícími podobné zkoušky jako NÚCEM (Část I);

4 www.alte.org/resources/Documents/code_practice_en.pdf, www.iltaonline.com/page/CodeofEthics, www.ealta.eu.org/documents/archive/guidelines

2. Zjistit, zda, případně do jaké míry a v jakých aspektech jsou testové verze použité v letech 2012–2015 srovnatelné, a pro toto zkoumání využít některé z metod vybraných na základě studia literatury a dalších kritérií (Část II);
3. Na základě zjištěných výsledků a zkušeností s použitými metodami navrhnout takové postupy a metody, které by mohly být implementovány a v budoucnu využívány poskytovatelem zkoušek (NÚCEMem) při vývoji a sestavování srovnatelných testových verzí a reportování a interpretaci výsledků studentů (Část III).

1.2 Výzkumné otázky

Z výzkumných cílů byly odvozeny tři výzkumné otázky (RQ1–3):

RQ1: Do jaké míry jsou testové verze slovenské maturitní zkoušky z anglického jazyka B1 ekvivalentní z hlediska obsahu, konstruktu a psychometrických vlastností, jaké povahy jsou případně zjištěné odlišnosti a jak zásadní jsou pro interpretaci výsledků?

RQ2: Jsou metody zjišťování srovnatelnosti testových verzí použité v tomto výzkumu dostatečně průkazné, spolehlivé a praktické, aby mohly být používány i v kontextu slovenské maturitní zkoušky?

RQ3: Jaké metody a postupy by mohly být zavedeny do procesu vývoje testových verzí slovenské maturitní zkoušky z anglického jazyka v rámci stávající legislativy, aby bylo dosahováno ekvivalence skórů a srovnatelnosti používaných testových verzí?

2 Kontext slovenské maturitní zkoušky

O reformě maturitní zkoušky na Slovensku[5] a o její nové podobě se začalo diskutovat v 90. letech 20. století. V letech 2000–2004 proběhlo několik celostátních monitorovacích šetření na populaci maturantů (tzv. MONITOR) s cílem poskytnout zpětnou vazbu a ověřit vývoj testů pro externí část maturitní zkoušky. V roce 2004 se konala generální zkouška podle nové koncepce maturitní zkoušky, a to z vybraných předmětů (anglický jazyk, německý jazyk a matematika) na vzorku přibližně 50 tisíc žáků SŠ. V roce 2005 byla zavedena povinná centrálně zadávaná tzv. externí část maturitní zkoušky (dále EČ MZ) z AJ, NJ a matematiky a proběhla generální zkouška pro zadání EČ z dalších cizích jazyků (FJ, RJ, ŠJ), v té době ještě ve třech úrovních obtížnosti. V roce 2007 proběhla generální zkouška EČ pro tzv. vyučovací jazyky (slovenský, maďarský a ukrajinský jazyk). V roce 2008 byl novým školským zákonem[6] zřízen Národný ústav certifikovaných meraní (NÚCEM), který byl pověřen realizovat novou maturitní zkoušku. Od roku 2009 na Slovensku probíhá EČ MZ také on-line formou, EČ MZ z anglického jazyka je tedy nabízena v tradiční formě papír-tužka, nebo on-line, a to ve dvou úrovních obtížnosti (B1 jako povinná pro SOŠ a B2 jako povinná pro gymnázia a volitelná pro SOŠ). O účelu EČ MZ se na stránkách NÚCEMu dočteme následující:

> Maturitná skúška je objektívnym meradlom vedomostí, zručností a všeobecných kompetencií absolventa strednej školy. Maturitné vysvedčenie – doklad o ukončení štúdia na strednej škole – má mimoriadny význam pre študentov, vypovedá o ich schopnosti pokračovať v štúdiu a uplatniť sa v budúcom povolaní.

5 zdroj: http://www.nucem.sk/sk/maturita, https://www.scio.cz/download/tkkv/2015/tkkv-jurenkova.pptx

6 ZÁKON 245/2008 Z. z. o výchove a vzdelávaní (školský zákon) a o zmene a doplnení niektorých zákonov

> Cieľom externej časti a písomnej formy internej časti maturitnej skúšky je overiť a zhodnotiť tie vedomosti a zručnosti maturantov, ktoré nie je možné overiť v dostatočnej miere v ústnej forme internej časti. Testujú sa zručnosti a kľúčové kompetencie ako počúvanie s porozumením (v cudzích jazykoch), čítanie s porozumením, gramatika v kontexte, schopnosť prezentovať vlastný prejav písomnou formou (v cudzích aj vyučovacích jazykoch).

V kompetenci škol zůstává tzv. interní část maturitní zkoušky. Interní část maturitní zkoušky tvoří písemná forma a ústní forma. Na středních odborných školách se maturuje také z teoretické a praktické části odborné složky[7]. V případě cizích jazyků jde o dovednost psaní, která probíhá na škole. NÚCEM zajišťuje centrální zadání písemné práce a kritéria a pokyny pro hodnocení. Práce však hodnotí učitelé dané školy. Ústní zkouška není centrálně nijak řízena.

Tyto změny souvisely s celkovou kurikulární reformou na Slovensku. Koncepce (Koncepcia vyučovania cudzích jazykov na základných a stredných školách) byla přijata v roce 2007. V této koncepci se mimo jiné objevuje explicitní odkazování na SERRJ[8] z hlediska obsahu, struktury i požadované výstupní úrovně na jednotlivých stupních vzdělávání a typech škol. Toto propojení s externím, celoevropsky používaným dokumentem bylo též důvodem k využití SERRJ v jednom z kroků tohoto výzkumu (viz oddíl 3.2).

Reforma státem garantované maturitní zkoušky je součástí kurikulární reformy a musí tedy reflektovat změny v celém vzdělávacím systému. Na jedné straně je zřejmé, že existují omezení, mj. legislativní, finanční, logistická a personální, která významným způsobem ovlivňují možnosti poskytovatelů zkoušek, zejména v případě státních institucí. NÚCEM není v tomto žádnou výjimkou. Na straně druhé je třeba konstatovat, že jde o zkoušku vysoké důležitosti (*high-stakes exam*), jejíž výsledky významně ovlivňují budoucí život testovaných, a měla by tedy být zkouškou standardizovanou ve všech aspektech, od procesů přípravy obsahu, logistiky a vyhodnocení zkoušek, po dodržování standardů souvisejících s dobrou praxí a spravedlivým přístupem k testovaným, tedy standardů obecně přijímaných v pedagogickém a psychologickém testování (např. AERA, APA, & NCME, 1999 a 2014).

7 https://www.iedu.sk/zivotne_situacie/maturita/Stranky/default.aspx

8 A Common European Framework of Reference for Language Learning, Teaching, Assesment. Council of Europe. (2001) Dostupné z: http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf.

Mezi základní požadavky na kvalitu v oboru jazykového testování patří existence validačních procesů, mimo jiné též prokazování validity způsobu, jakým jsou interpretovány výsledky. V případě testových verzí zkoušky se validace týká i podání důkazu o tom, že výsledky různých verzí téhož testu lze interpretovat podle stejného principu a že je lze porovnávat. Zmíněné požadavky popisované v profesních dokumentech[9] nejsou sice právně závazným předpisem, popisují však dobrou praxi a etické principy oboru jazykového testování a cíle, o jejichž dosažení by měli poskytovatelé zkoušek usilovat.

V období od zahájení příprav celé koncepce nové slovenské maturity[10] do jejího zavedení a též i do současné doby došlo k mnohým změnám. V případě cizích jazyků se diskutovalo např. o tom, v kolika úrovních obtížnosti bude zkouška nabízena, jaké úrovně to budou a které z nabízených úrovní budou povinné pro odborné školy a které pro gymnázia. V těchto aspektech také docházelo k nejvýznamnějším změnám. I z toho důvodu se náš výzkum zabývá verzemi z let 2012–2015, kdy byla podoba EČ MZ AJ B1 relativně stabilní.

Pozitivním aspektem slovenské EČ MZ z cizího jazyka je zveřejňování podrobných zpráv a komentovaných statistických analýz z každého testování. Ve srovnání např. s českým CZVV poskytuje NÚCEM po ukončení testovacího cyklu podrobné a jasné zprávy o výsledcích realizovaných zkoušek. Přesto však nejsou některé aspekty související s nastavením a udržením standardů obsahové a procesní kvality explicitně řešeny. Přestože NÚCEM nikde neprokazuje, že jím poskytované zkoušky jsou ekvivalentní v čase, s výsledky je tak nakládáno, a to jak na úrovni celé maturitní populace, tak i na úrovni jednotlivců – ve smyslu stejné intepretace výsledků. Problémem (nejen) slovenské maturitní zkoušky je to, že neexistuje dostatek veřejně dostupných informací, z nichž by bylo možné usuzovat na to, jak byla zkouška vyvíjena, co pro naplňování standardů kvality poskytovatel zkoušek (NÚCEM) dělal a dělá, jakým způsobem probíhá validace a zda existuje výzkum, který by se prokázáním srovnatelnosti testových verzí slovenské maturitní zkoušky zabýval. Další problematickou oblastí EČ MZ je nedostatečně

9 Např. ILTA Code of Ethics a Guidelines for Practice, EALTA Guidelines for Good Practice in Language Testing and Assessment; ALTE Code of Practice a ALTE Minimum Standards.

10 Jejím účelem mělo být stát se standardizovanou prestižní zkouškou kombinující ověřování naplnění středoškolského kurikula a zároveň uspokojení požadavků vysokých škol na zkoušku, jež by mohla být akceptována jako součást přijímacího řízení.

propracovaná a velmi obecná definice konstruktu. Specifikace zkoušek lze sice nalézt na stránkách ŠPÚ[11], jsou však velmi obecné, obsahují pouze informaci o o označení úrovně maturitní zkoušky z anglického jazyka (v tomto případě úroveň B1), čímž pouze odkazují k SERRJ, a dále několik obecných deskriptorů vymezujících obecně dovednosti očekávané u maturantů a měřené maturitním testem. Proto nebylo možné specifikace využít jako jeden z nástrojů výzkumu.

Mezi zásadní otázky, které by měla každá instituce zajišťující zkoušky vysoké důležitosti zkoumat a zodpovědět, patří dvě vzájemně propojené otázky, které jsou i tématem tohoto výzkumu:

> Je zajišťována srovnatelnost testových verzí, respektive stabilita měření napříč různými verzemi téhož testu (*measurement invariance*)?
>
> Je zajištěna interpretační a konsekvenční validita skórů (včetně stanovení hraničního skóru) napříč různými verzemi téže zkoušky?

V případě EČ MZ, konkrétně testů z anglického jazyka na úrovni B1, není známo, zda je v procesu vývoje EČ MZ zabudován mechanismus, který by monitoroval procesy umožňující validní interpretaci výsledků, včetně tvrzení o srovnatelnosti obsahové a psychometrické. Není tedy možné prokázat ani popřít, že zkoušky používané v různých termínech jsou tzv. ekvivalentními verzemi téhož testu. Vzhledem k neexistenci veřejně dostupných informací je pravděpodobné, že validace zkoušky v tom smyslu, jak o ní hovoří např. Standardy (AERA, APA, & NCME, 2014), prozatím komplexně prováděna není.

Výše uvedené problémy se sice týkají jen malého výseku validačních procesů, přesto jejich vyřešení představuje důležitý krok na cestě k smysluplné interpretaci testových skórů a rozhodnutí činěných na jejich základě. Pokud totiž nelze možné prokázat, že verze téhož testu jsou ekvivalentními verzemi téhož testu, pak a) není možné interpretovat stejně výsledné skóry; b) není možné tvrdit, že verze ověřují stejný konstrukt; c) rozhodnutí činěná na základě skórů nemusí být konsistentní; d) není možné zaručit spravedlivý přístup k testovaným; e) spravedlivost výsledků testování může být zpochybněná.

11 https://www.nucem.sk/sk/merania/narodne-merania/maturita/roky/2019-2020?componentId=1571

3 Teoretická východiska výzkumu

V této kapitole se věnujeme ukotvení terminologie pro další práci, zdůvodnění významu srovnatelnosti testových verzí jako aspektu validity a sestavení přehledu metod používaných při prokazování srovnatelnosti testových verzí. Oddíl 3.1 je věnován pojmům konstrukt, validita, validace, spravedlivost, srovnatelnost a ekvivalence a diskusi o nich v souvislosti se srovnatelností testových verzí, oddíl 3.2 představuje základní směry výzkumu srovnatelnosti testových verzí.

3.1 Vymezení klíčových pojmů

Obor jazykového testování operuje s řadou významných konceptů, které spolu vzájemně souvisí nebo se ovlivňují: konstrukt (*construct*), validita, reliabilita, dopad (*washback*, *impact*), zobecnitelnost (*generalizability*), konzistentnost (*consistency*), autenticita (*authenticity*), ekvivalence (*equivalence*), srovnatelnost (*comparability*), praktičnost (*practicality*) nebo proveditelnost (*feasibility*). Zejména v poslední době je stále více akcentováno téma etiky (*ethics*), spravedlivosti (*fairness*) a odpovědnosti vůči uživatelům výsledků zkoušek (*accountability*). Objevuje se myšlenkový proud tzv. kritického jazykového testování (*critical language testing*), který zdůrazňuje centrální postavení testovaných (*test-takers*) při posuzování kvalit testovacího nástroje. Diskuse v rámci odborných platforem[12] se pokoušejí vymezit obsah a vztah výše uvedených konceptů, a i když se ne vždy shodují na hierarchii či přesném vymezení pojmů, shodují se na tom, že spolu vzájemně souvisí a hrají klíčovou roli při posuzování kvality, tj. při validaci[13] testů nebo zkoušek. Dále pro účely tohoto výzkumu vymezujeme nejdůležitější pojmy.

12 Odborné časopisy (např. Language Testing, Language Assessment Quarterly), diskusní fóra (l-TESTL, Research Gate), konference mezinárodních institucí ILTA, EALTA, ALTE a další.

13 Podle van der Walta a Steyna (2008) je validita abstraktní pojem a validace je proces, jehož prostřednictvím je tento koncept popisován, hodnocen a interpretován.

3.1.1 Konstrukt

Urbinová (2004, s. 156) vymezuje konstrukt jako „cokoli, co je vytvářeno jako produkt lidské mysli (*mind*), ale není to možné pozorovat přímo. Konstrukt je abstrakce vztahující se ke konceptům, idejím, teoretickým entitám, hypotézám (…).“ Hendl (2009) označuje konstrukt jako „pojem nebo ideu specificky navrženou pro daný výzkum nebo tvorbu modelu“ (s. 27). Je pozorovatelný nepřímo, jako tzv. latentní proměnná (*latent variable, latent trait*), a to na projevech chování různého typu.

Konstrukt jako každá abstrakce je určitým zjednodušením (skutečnosti, ideje, jevu apod.). Právě zjednodušení a abstraktní povaha usnadňuje a umožňuje určitý fenomén uchopit, zkoumat, popsat a interpretovat. Tato abstrakce neexistuje sama o sobě, ale jak již z označení *konstrukt* vyplývá, je konstruována právě prostřednictvím popisů (např. specifikací testu) a operacionalizací (obsahového naplnění). V případě jazykového testování jsou takovými konstrukty např. jazyková způsobilost obecně, úroveň B2 dle SERRJ, řečová činnost recepce, čtení s porozuměním apod. Abychom tyto konstrukty mohli zkoumat, je třeba je teoreticky ukotvit: popsat jejich podstatu, určit, co je indikátorem-projevem existence tohoto konstruktu, jak lze tyto projevy elicitovat a pozorovat (např. pomocí úloh a položek, s využitím konkrétních testovacích technik, stanovením způsobu vyhodnocení a interpretace výsledků). Je tedy nutné nejprve definovat předmět zkoumání (konstrukt komunikační jazykové kompetence nebo některé z jejích složek, např. porozumění mluvené řeči), poté určit, jakým způsobem a pomocí jakého obsahu lze projevy konstruktu nejlépe získat nebo pozorovat (specifikace nástroje, např. testovacích technik, množství a formáty úloh, zacílení položek, formáty očekávaných odpovědí, způsob hodnocení těchto odpovědí) a rozhodnout, jak mají být tyto projevy interpretovány (definice kvality a kvantity sledovaného, stanovení, do jaké míry se pozorovaný projev shoduje s teoretickým vymezením konstruktu apod.). Vzhledem k tomu, že se v jazykovém testování konstrukt obvykle operacionalizuje prostřednictvím úloh či položek a výkon testovaných je sumarizován pomocí skórů, jsou konstrukty testových verzí porovnávány nepřímo, mj. právě přes výsledný skór coby obraz měřeného konstruktu. I to je důvod, proč je srovnatelnost testových verzí, resp. ekvivalence interpretací výsledných skórů klíčovou otázkou jak z pohledu vývoje testových verzí, tak z pohledu validace a spravedlivosti při využívání výsledků zkoušek. Můžeme shrnout, že pokud je stejný

konstrukt měřen a operacionalizován prostřednictvím různých testových verzí téhož testu, je srovnatelnost měřeného konstruktu napříč testovými verzemi základní podmínkou pro interpretaci a porovnávání výsledků (van der Vijver & Poortinga, 2005) a pro úvahy o případném vyrovnávání skórů z různých testových verzí.

3.1.2 Validita a validace

Mezi zásadní požadavky kladené na každé měření patří spolehlivost a smysluplnost závěrů a rozhodnutí činěných na jejich základě a obhajitelnost a spravedlivost rozhodnutí činěných na základě těchto výsledků, což lze shrnout pod zastřešující termín validita.

Podle van der Walta a Steyna (2008) je validita abstraktní pojem a validace je proces, jehož prostřednictvím je tento koncept popisován, hodnocen a interpretován. Validace takto pojímaná se tedy již nevztahuje na test jako nástroj, nýbrž se týká celého procesu vývoje a zejména použití a interpretace testu. V diskusích o validitě tak v popředí stojí otázky užitečnosti testu (*test usefulness*), smysluplnosti (*meaningfulness*) interpretace výsledků zjištěných testem a důsledků, jaké mohou mít rozhodnutí činěná na základě výsledků.

Z historické perspektivy lze shrnout, že až do prosazení se komunikačního přístupu k výuce a hodnocení v cizích jazycích (na konci 70. let a především v 80. letech) existovaly dva hlavní směry v nahlížení na kvality testu (Spolsky, 1995; Weir, 2005). První směr ovlivněný severoamerickou psychometrickou tradicí se orientoval více na reliabilitu testu jako základní kritérium kvality testu a akcentoval přesnost a spolehlivost měření jako zásadní hledisko spravedlivého testování: test, který byl spolehlivý, byl validním testovacím nástrojem. Druhý směr reprezentovaný zejména syndikátem UCLES ve Velké Británii, resp. jejich zkouškami, kladl důraz na obsah a konstrukt testu, tedy na to, co je testy měřeno, a na korespondenci konstruktu s edukační realitou a vzdělávacími cíli (Spolsky, 1995). Právě v souvislosti s nástupem komunikačního přístupu do výuky, a tedy i do testování docházelo ke sbližování obou směrů a v současnosti je zřejmá hodnota obou přístupů, jak lze vidět v současných diskuzích o validačních rámcích a pojetí validity.

Proměny diskusí o validitě probíhaly paralelně se změnou nahlížení na obsah, strukturu a pojetí jazykových testů v průběhu 20. století (Spolsky, 1995) a s tím, jak se od 70. let 20. století postupně prolnul

přístup k testování orientovaný na psychometrické vlastnosti testů s přístupem orientovaným na obsah a konstruktovou validitu a na smysluplnost interpretace výsledků. Přelomový pohled na validitu nejvýrazněji zformuloval Messick (např. 1987, 1989, 1993, 1995), který zpochybnil validitu jako výlučnou vlastnost testu a zdůraznil pojetí validity jako míry shody teoretických východisek s realizací testu, s interpretací a použitím skórů a s důsledky, které s sebou použití skórů nese.

> To validate an interpretive inference is to ascertain the degree to which multiple lines of evidence are consonant with the inference, while establishing that alternative inferences are less well supported. To validate an action inference requires validation not only of score meaning but also of value implications and action outcomes, especially appraisals of the relevance and utility of the test scores for particular applied purposes and of the social consequences of using the scores for applied decision making. Thus the key issues of test validity are the interpretability, relevance, and utility of scores, the import or value implications of scores as a basis for action, and the functional worth of scores in terms of social consequences of their use. (Messick, 1989, s. 5)[14]

Centrální roli v Messickových úvahách o validitě mají konstrukt testů (tj. konstruktová validita) a společenský dopad, který plyne z použití výsledků testu v praxi (tj. konsekvenční validita). Toto nové pojetí validity spolu s akceptací komunikačního přístupu ovlivnilo i další výzkumníky: vznikaly nové validační modely; některé byly pouze teoretické, jiné byly výsledkem snahy vytvořit praktický validační nástroj uplatnitelný v konkrétních kontextech. V souvislosti s Messickovým průlomovým textem jsou zejména 90. léta 20. století obdobím intenzivních diskusí o povaze validity (Chapelle, 1999). Messickovo pojetí bylo zabudováno i do Standardů pro pedagogické a psychologické testování (AERA, APA, & NCME, 1999, 2014).

Weirův socio-kognitivní model (2005) do validačního procesu kromě obsahu pojmů jako kognitivní, kontextový, kriteriální, konsekvenční a skórový aspekt validity integruje také pohled na charakteristiky testovaných jako jeden z důležitých aspektů, který musí být brán v potaz při validačním procesu. Bachman (1990), později Bachman a Palmer

14 „Validací interpretačního úsudku se rozumí zjištění, do jaké míry jsou různé typy důkazů a dokazování v souladu s úsudkem, a zároveň do jaké míry jsou alternativní úsudky méně robustní. Validací akčního úsudku (závěrů) vyžaduje nejen validaci významu skóru, nýbrž také validaci důsledků a rozhodnutí, zejména pak zhodnocení relevance a užitečnosti testových skórů pro konkrétní účel, a zhodnocení společenských důsledků využití testových skórů pro rozhodování. Klíčovými otázkami validity tedy jsou míra interpretovatelnosti, relevance a využitelnost skórů, význam nebo implikace skórů jako podkladů pro rozhodnutí a funkční hodnota skórů ve smyslu společenských důsledků použití skórů.“ (překlad autorky)

(1996, 2010) akcentují smysluplnost a interpretaci skórů vztaženou k účelu, ke kterému byl test vytvořen, a k jeho plánovanému využití, přičemž tato interpretace musí být v souladu s procesy, pomocí nichž byl skór získán. Zdůrazňovány jsou účel a použití testu jako primární kritéria pro validační proces, dále smysluplnost používání hodnoticího nástroje a jeho výsledků, a dopady a důsledky použití testu. Ve starších pracích je validační model autory označován jako model užitečnosti testu (*test usefulness*) (Bachman & Palmer, 1996), později jej transformují do tzv. AUA – *Assessment Use Argument* (Bachman & Palmer, 2010). Ten je praktičtěji orientovaný a slouží jako rámec pro výstavbu argumentace při validaci testu. Důležitá je provázanost jednotlivých aspektů a jejich vzájemná závislost. Autoři zdůrazňují nutnost zabudovat úvahy o validitě do všech procesů tvorby testů již od samého počátku a apelují na to, aby se validace a validační argumentace staly součástí vývoje a použití testů.

Rovněž Standardy (AERA, APA, & NCME 1999, 2014), které integrují různá pojetí validace, představují validační proces jako postupné budování obhajitelné, o důkazy se opírající argumentace při interpretaci výsledků. V českém kontextu na tyto Standardy a jejich pojetí validity odkazují srozumitelným způsobem např. Chvál, Straková a Procházková (2015, s. 183–184).

V současné době tedy existuje mnoho dalších validačních modelů, tedy přístupů k validačnímu procesu, z nichž každý akcentuje jiný aspekt validace. Můžeme zmínit např. Kaneův interpretačně-argumentační model (1990, 2011), který však nezahrnuje argumentaci související s důsledky použití testu; Bachmanův *Assessment Use Argument model* (2010) nebo Weirův socio-kognitivní model, používaný zejména v kontextu zkoušek Cambridge ESOL (Weir, 2005; Khalifa & Weir, 2009; Geranpayeh & Taylor, 2013). Všechny modely jsou komplexní, liší se spíše ve strukturování argumentace, v tom, jaké aspekty validity akcentují a jak praktické jsou při aplikaci v různých kontextech. Mají však společné to, že oproti předchozím chápáním validity jako vlastnosti testu a akcentování kriteriálního pohledu na validitu a validaci (nejčastěji pomocí korelace s jiným hodnoticím nástrojem) dochází ke konceptuálnímu posunu a většina validačních modelů se shoduje v nahlížení na validitu z pohledu užitečnosti testovacího nástroje a smysluplnosti interpretace testových skórů (Bachman, 1990), přičemž jsou primárně akcentovány důsledky, které s sebou nese využití testových skórů v procesu rozhodování o testovaných, a odpovědnost poskytovatele testu vůči uživatelům výsledků zkoušek.

3.1.3 Spravedlivost a validita

K chápání vztahu validity a spravedlivosti existují různé přístupy (pro přehled např. Xi, 2010). Spravedlivost může být považována za jeden z aspektů validity, nebo naopak za kvalitu, jež je podmínkou pro validitu testu.

Xiová (2010, s. 154) chápe spravedlivost jako „srovnatelnou validitu pro všechny identifikovatelné skupiny testovaných". Proti tomuto pojetí se vymezuje Davies (2010), který ve své odpovědi na pojetí Xiové reaguje a zpochybňuje vyčleňování spravedlivosti jako samostatné, či dokonce nadřazené kategorie, a odkazuje ke Standardům (AERA, APA, & NCME, 1999), které spravedlivost definují jako rovný přístup ke všem testovaným, nezatíženost (absence of bias), spravedlivost výstupů testování a rovnost (equality) učebních příležitostí. Davies (2010) tvrdí, že pokud je toto zajištěno a prokázáno, pak proběhly tytéž procesy, jaké probíhají v rámci validace. Existenci samostatné kategorie spravedlivosti zpochybňuje. Podle Daviese není možné očekávat stejnou míru schopností a dovedností u všech testovaných, je však třeba, aby byly zajištěny rovné a spravedlivé podmínky testování.

S výše zmíněným Daviesovým pohledem na spravedlivost souzní např. i pojetí spravedlivosti nastíněné v dokumentaci zkoušek poskytovaných ETS15. Aby byl test spravedlivý, je podle výzkumného týmu ETS třeba zkoumat například to, zda některé jeho části (položky, úlohy) nezvýhodňují některé skupiny testovaných, zda je naplněn princip rovných příležitostí pro všechny testované k prokázání jejich schopností a dovedností, nebo zda byly vhodně modifikovány testy pro testované se speciálními potřebami.

Důležitější než vzájemný vztah a hierarchie pojmů spravedlivost a validita je to, že tyto pojmy reprezentují důležité aspekty, které musí být v průběhu validačního procesu zkoumány a musí být podány důkazy o nejlepší možné praxi při vývoji a použití konkrétního testu. Pak je teprve možné posuzovaný test prohlásit za smysluplný a užitečný hodnoticí nástroj vhodný pro daný účel. Zároveň je každá validace lokální, tzn., že souvisí s podmínkami, v jakých test existuje. Validace by tedy měla pracovat s argumenty, které se týkají použití testu v konkrétních podmínkách a pro konkrétní testované.

15 ETS – Educational Testing Service: https://www.ets.org/gre/institutions/about/fairness/

3.1.4 Srovnatelnost a ekvivalence

Téma srovnatelnosti, resp. ekvivalence[16], a pojmy jako ekvivalentní, paralelní a alternativní testové verze apod. jsou na poli jazykového testování relativně nové. Jedna z prvních klíčových studií, která se zabývala otázkou porovnatelnosti výsledků dvou testů, jež deklarují, že měří stejný konstrukt, je studie Bachmana a kol. (1995). Autoři se v ní zabývali otázkou použití testových skórů z mezinárodních jazykových zkoušek IELTS a TOEFL, respektive tím, do jaké míry mají výsledky obou zkoušek stejný význam a zda mohou být považovány za srovnatelné pro účely přijímacích řízení do studijních programů s anglickým vyučovacím jazykem. Srovnatelnost testových verzí je, a i u zmíněné studie byla jednou z klíčových podmínek pro smysluplnou interpretaci testových skórů, souvisí s konzistencí a spravedlivostí rozhodování a závěrů, které jsou na základě výsledků činěny. Je zřejmé, že poskytovatel nebo tvůrce testu nemůže být zodpovědný za vše, co se týká vlivu testování na testované a další uživatele výsledků zkoušek (Davies, 2008), je však zodpovědný za kvalitu testu, který tvoří nebo poskytuje. Do určité míry může ovlivnit i to, jakým způsobem a k jakému účelu budou skóry využívány a jak mají být interpretovány. Maximální snaha o srovnatelnost testových verzí, resp. ekvivalenci skórů a jejich interpretaci je proto nezbytnou podmínkou pro etické použití jakéhokoli testovacího nástroje. Tvůrce nebo poskytovatel testu by tedy měl udělat maximum pro to, aby byl schopen u testových verzí empiricky prokázat srovnatelnost obsahu, srovnatelnost způsobu měření a ekvivalenci měřeného konstruktu, což jsou vstupní podmínky nutné pro jakékoli následné porovnávání testových skórů. Pokud mají být testové skóry z různých testových verzí interpretovány stejně, pak by měly být používány takové postupy, aby mohlo být dosahováno ekvivalence skórů, a to i v případě různě obtížných testových verzí či různých populací. Pouze tak je možné na základě výsledků a jejich intepretací činit spravedlivá rozhodnutí.

3.1.5 Pohledy na srovnatelnost

Srovnatelnost, resp. ekvivalence testových verzí má několik úrovní, podle toho, jak přísná kritéria (viz oddíl 3.1.5.2) uplatňujeme. Můžeme na ni nahlížet z různých perspektiv. Ve zkouškách vysoké důležitosti, kam patří

16 Pojem ekvivalence bývá v literatuře používán v různých kontextech, s odlišným významem. Zde jej překládáme dvojím způsobem: v souvislosti se skóry a s abstraktními koncepty používáme spojení ekvivalence skórů, konstruktová ekvivalence; pojem srovnatelnost používáme ve vztahu k testovým verzím, kde existuje způsob, jakým dvě nebo více verzí téhož testu porovnávat a případně prohlásit za srovnatelné, tj. velmi podobné nebo identické. Podrobněji se tomuto věnujeme v dalších částech textu.

i maturitní zkouška, by mělo být dosaženo co možná nejkomplexnějšího pohledu na srovnatelnost a co nejvyšší úrovně srovnatelnosti, aby bylo možné porovnávat výsledky testovaných a vyslovit spravedlivé a validní závěry o výsledcích na úrovni jednotlivců i celé populace. Níže se pokusíme vysvětlit různé úrovně srovnatelnosti a používanou terminologii.

Ve Standardech (AERA, APA, & NCME, 1999, 2014) jsou tři níže uvedené pojmy chápány jako obsahově shodné, nejčastěji je používán pojem alternativní verze.

Alternativní, paralelní nebo **ekvivalentní verze** (*alternate, parallel, equivalent forms*) jsou ve Standardech definovány jako dvě nebo více zaměnitelných verzí testu, které měří stejný konstrukt stejným způsobem. V praxi to znamená, že testové verze:

a) mají stejné pokrytí ověřovaného obsahu a pro tento obsah používají stejné testovací techniky, tj. vycházejí ze shodných specifikací (von Davier, 2011, používá termín *nominálně paralelní testové verze* převzatý z Lorda a Novicka, 1968);
b) jsou administrovány za shodných podmínek;
c) ve srovnatelných populacích vykazují srovnatelné psychometrické vlastnosti (shodný průměrný skór a směrodatnou odchylku).

V definici, či spíše v odlišení alternativních (někdy též srovnatelných – *comparable*) a paralelních verzí od verzí ekvivalentních nepanuje shoda. Například starší verze Standardů (1999) je odlišuje na základě rozdílů v hrubých skórech a nutnosti skóry transformovat či vyvažovat[17]; jiní autoři považují převod hrubých skórů za běžný statistický postup (Bachman, 2004; ITC, 2005; Brown & Hudson, 2002; Urbina, 2004) a pojmy paralelní a ekvivalentní jsou v jejich pojetí synonyma. Jako příklad lze uvést definici ekvivalentních testových verzí podle ALTE, která je totožná s definicí paralelních verzí ve Standardech (1999):

> (T)estové verze jsou vytvářeny podle stejných specifikací a měří tutéž kompetenci. Aby testové verze naplňovaly přísná kritéria ekvivalence v rámci klasické teorie testů, musí při administraci stejným osobám vykazovat stejnou průměrnou obtížnost, varianci a kovarianci. (ALTE Multilingual Glossary of Language Testing Terms, 1988, citováno v Khalifa & Weir, 2009, s. 193).[18]

17 Paralelní verze (parallel forms) jsou takové, které „vykazují v každé testované populaci shodný průměrný hrubý skór, směrodatnou odchylku, shodnou strukturu chyby měření a korelaci s jinými měřeními“ (AERA, APA, & NCME, 1999, s. 173–174) (překlad autorky).

18 „(Test forms) are based on the same specifications and measure the same competence. To meet the strict requirements of equivalence under classical test theory, different forms of a test must have the same mean difficulty, variance, and co-variance, when administered to the same persons“ (překlad autorky).

Paralelnosti nebo ekvivalence (srovnatelnosti) v tomto pojetí je v praxi téměř nemožné dosáhnout obvyklými procesy vývoje testových verzí (Taylor, 2004, cit. podle Khalifa & Weir, 2009, s. 193; MLTDE, 2011, s. 82), tzn. bez transformace[19] skórů, využití metod vyrovnávání, převodu hrubých skórů na společnou škálu (*scaling*), provazování (*linking*), nebo bez rutinního využívání teorie odpovědi na položku (*Item Response Theory – IRT*) při vývoji testů. I nová verze Standardů (AERA, APA, & NCME, 2014) chápe pojmy alternativní, paralelní, ekvivalentní za obsahově shodné a předpokládá uplatňování výše zmíněných statistických postupů jako běžnou součást vývoje testových verzí.

3.1.5.1 Obsahový pohled na srovnatelnost

Na srovnatelnost můžeme nahlížet také z perspektivy předmětu a cíle měření. V tomto smyslu lze rozlišit v zásadě čtyři různé pohledy na srovnatelnost. Jde o zkoumání srovnatelnosti verzí téhož testu zadávaného v různých jazykových mutacích pro různé jazykové a/nebo etnické skupiny (*cross-linguistics* nebo *cross-cultural comparability*), o studie porovnávající verze téhož testu zadávaného různým způsobem, např. test tužka-papír a test zadávaný na počítači (*cross-methods comparability*), studie porovnávající testy produkované různými institucemi, orientované na stejnou nebo podobnou skupinu testovaných mající stejný nebo podobný konstrukt a interpretaci testových skórů, např. testy IELTS a testy TOEFL (*cross-institutions comparability*). Poslední typ studií porovnává stabilitu nebo srovnatelnsost verzí téhož testu produkovaného pro různé termíny testování (*cross-versions stability/comparability* nebo *stability/comparability over time*).

První typ, tj. **mezijazykové srovnávání** (*cross-linguistics* nebo *cross-language studies*), popisovaný např. Sirecim a Allaloufem (2003), Joldersmou (2011), Hauptovou a Kochovou (2012) nebo Anýžovou (2013), se zabývá testy, které jsou zadávány v různých jazycích, respektive překládány do různých jazyků. Podle Sireciho a Allaloufa (2003, s. 149) je třeba prokázat, že u všech cílových jazykových skupin a) se ověřovaný konstrukt (construct) převodem do jiného jazyka nezměnil, b) je konstrukt měřen stále stejným způsobem u všech pozorovaných skupin, c) jsou úlohy statisticky i jazykově ekvivalentní[20] a d) podobné skóry napříč verzemi vypovídají o stejné úrovni výkonu (*proficiency*).

19 Pojem transformace, resp. význam, v jakém ho užíváme, vysvětlujeme v oddíle 3.2.3.

20 Autoři pojem *ekvivalentní* nevysvětlují. Z obsahu textu vyvozujeme, že je synonymem k pojmu paralelní.

Autoři se shodují na kombinaci metod subjektivního posuzování a psychometrických postupů a na nutnosti posoudit nejprve ekvivalenci konstruktů. Hauptová a Kochová (2012) se zabývají úrovněmi srovnatelnosti testových verzí (ekvivalencí konstruktu, ekvivalencí jednotek měření, tj. metrickou ekvivalencí a ekvivalencí škálovou), a dále zkreslením výsledků neboli systematickou chybou měření (*bias*). Tvrdí, že je-li zjištěn problém tohoto druhu, pak nemohou být testové verze považovány za ekvivalentní (Haupt & Koch, 2012, s. 66). Rozlišují systematické chyby na úrovni konstruktu, na úrovni metody, jež vzniká v důsledku použitých testovacích technik/metod nebo způsobu administrace, a dále systematické chyby na úrovni položek, ke které dochází, má-li stejná skupina testovaných se stejnými schopnostmi v různých testových verzích jinou pravděpodobnost volby správné odpovědi. Tato systematická chyba se může projevovat buď stejně na všech hladinách skórů, nebo ovlivňuje některé skupiny testových skórů zásadněji než jiné, a je třeba zvážit míru závažnosti tohoto problému z hlediska interpretace výsledků.

Druhý typ studií se zabývá **srovnatelnosti formátů verzí** (*cross--methods studies*). Porovnávají se verze zkoušek zadávané v různých formátech, např. verze papír-tužka versus verze administrovaná na počítači (Bachman a kol. 1995), nebo porovnávají přímé a nepřímé testovací techniky (O´Loughlin, 2001). V těchto studiích jsou využívány metody analýzy struktury obsahu, korelační analýza, ANOVA (analýza rozptylu), konfirmatorní faktorová analýza, metody korpusové lingvistiky porovnávající texty v subtestech z pohledu type/token, délky slov a délky a struktury vět atd. (Choi, Sung, & Boo, 2003).

Třetí typ studií se věnuje **meziinstitucionálnímu srovnávání** (*cross-institutional studies*). Porovnává zkoušky, které mají podobný či shodný deklarovaný konstrukt (např. úroveň jazykové způsobilosti), ale jsou nabízeny různými, často konkurenčními institucemi. Zde je možné uvést jako příklad jednu z prvních takových studií, která porovnává zkoušky TOEFL poskytovatele ETS a zkoušky IELTS poskytovatele Cambridge ESOL (Bachman a kol. 1995), nebo pozdější studii Kunnana a Carra (2017). I zde je využívána např. analýza struktury obsahu, metody korpusové lingvistiky analyzující využité texty, korelační analýza a faktorová analýza.

Poslední oblastí je zkoumání **srovnatelnosti různých verzí téhož testu** (*cross-versions studies, stability studies*). Studie se zabývají porovnáním různých verzí téhož testu připravovaných pro různé testové termíny, tedy srovnatelností nebo stabilitou testových verzí v čase. Jde

např. o vývojové studie, které sledují změny v chování apod. v závislosti na věku, a proto potřebují kontrolovat stabilitu měřeného konstruktu (Widaman, Ferrer, & Conger, 2010). V případě jazykového testování můžeme jako příklad tohoto typu zkoumání uvést studie, které porovnávají populace (skupiny z různých zemí, skupiny konající zkoušku v různých termínech), a potřebují prokázat, že odlišnosti ve výsledcích nejsou způsobeny tím, že by byl měřen jiný konstrukt jiným způsobem, pokud použijí jinou verzi téhož testu. Častěji se vyskytují studie s experimentálním designem, méně studie, které zkoumají stabilitu testových verzí v čase na již realizovaných testech. Příkladem experimentální studie je práce Baghaeie (2010), který zjišťoval, zda neexistence vyrovnávacího mechanismu pro dvě verze testu na čtení s porozuměním vede k nespravedlivým rozhodnutím o úspěšnosti testovaných, nebo studie Weira a Wuové (2006), kteří porovnávají tři verze ústní zkoušky GEPTS-I.

Nutno poznamenat, že v oblasti jazykového testování není literatura k této problematice příliš rozsáhlá, a pokud existují zmínky o tomto tématu, pak spíše v rámci zpráv o validaci zkoušek jako takových, zejména u poskytovatelů zkoušek, kteří disponují bankou úloh s kalibrovanými parametry a rutinně používají metody vyrovnávání testových skórů[21]. Také se však objevují tvrzení o ekvivalenci či paralelnosti testových verzí bez dalších odkazů na podrobnější dokumentaci, nebo jsou tato tvrzení postavena na argumentu obsahové srovnatelnosti – tj. že testy jsou vytvářeny podle stejné specifikace a standardizovaným postupem.

Studie na téma srovnatelnosti testových verzí poskytovaných jednou institucí pro různé zkušební termíny na datech z realizovaného testování, tedy v neexperimentálním designu, není v tomto přehledu zastoupena, neboť se nám totiž nepodařilo najít publikovanou studii tohoto typu. Na základě dosavadní zkušenosti v oboru testování, studia literatury a dotazů v institucích poskytujících národní a mezinárodní jazykové zkoušky (NÚCEM, CZVV, CIEP, Cambridge ESOL, TestDAF, ÚJOP UK apod.) se domníváme, že existují dva přístupy k této problematice. První z nich můžeme zhruba popsat jako přístup založený na promyšleném designu pretestací a stanovení hraničního skóru, využívání teorie odpovědi na položky (IRT) nebo jiných statistických postupů umožňujících vyrovnávání nebo provazování skórů z různých testových verzí a vytváření banky úloh se známými parametry úloh, včetně psychometrických,

21 Např. Research Notes: http://www.cambridgeenglish.org/research-notes/; ETS Research: http://search.ets.org/researcher/query.html?col=all&ql=&qt=validity; http://stepeiken.org/comparability

z nichž se potom sestavují testové verze se srovnatelnými vlastnostmi. Druhý přístup k sestavování testových verzí, který je kriticky zmiňován např. Spolskym (1995), Bachmanem a kol. (1995), Weirem a Wuovou (2006), bychom mohli charakterizovat jako spíše intuitivní přístup, kdy je srovnatelnost testových verzí odvozována od toho, že testové verze jsou vyvíjeny podle stejných testových specifikací obsahu, mají stejnou strukturu a obsahují tytéž testovací techniky. Souhlasíme s tvrzením některých autorů (např. Weir, 2005), že intuitivní přístup je nedostatečný pro to, aby bylo možné tvrdit, že takto vytvářené testové verze jsou srovnatelné a skóry se stejnou nominální hodnotou mohou být považovány za ekvivalentní a zaměnitelné, tj. že je lze interpretovat stejným způsobem.

3.1.5.2 Psychometrický pohled na ekvivalenci

Srovnatelnost (zde je na místě spíše termín ekvivalence, resp. ekvivalence skórů) lze posuzovat také z psychometrického pohledu. Anýžová (2013) spojuje pojem ekvivalence s měřením; koncept ekvivalence měření (*measurement invariance*) pak definuje spíše ve smyslu ekvivalence konstruktové, tedy měření totožných znaků i za různých okolností. Rozlišuje tři úrovně – ekvivalenci teoretického konceptu (konstruktu), ekvivalenci položek, kterými se konstrukt operacionalizuje a měří, a ekvivalenci škál měření (s. 31). Vandenberg a Lance nebo (2000) Bialosiewiczová, Murphyová a Berryová (2013) operují také s termínem ekvivalence (též stabilita) měření a odlišují různé úrovně ekvivalence měření. Bialosiewiczová, Murphyová a Berryová definují ekvivalenci měření jako stav, kdy je „vztah mezi manifestními indikátory proměnných a měřeným konstruktem stejný napříč různými skupinami nebo v čase“[22]. V textu popisují čtyři úrovně ekvivalence měření:

1. **Konfigurální** nebo **konstruktová ekvivalence** (*configural/ construct equivalence*), kde platí, že ve sledovaných testových verzích pozorujeme **stejnou strukturu konstruktu**, tj. stejné položky měří tentýž konstrukt (stejnou vlastnost, latentní rys atd.) napříč jednotlivými testovými verzemi. Typickými metodami zkoumání konfigurální ekvivalence je strukturní modelování (SEM), jehož specifickým případem je konfirmatorní faktorová analýza (CFA). Dále je velmi často využívána

22 When the relationship between manifest indicator variables (scale items, subscales etc.) and the underlying construct are the same across groups or across time.

exploratorní faktorová analýza (EFA). Při využití CFA se konfigurální ekvivalence projeví tak, že v různých testových verzích bude stejná nebo velmi podobná struktura faktorů, tzn. že počet faktorů a vztah jednotlivých faktorů k položkám (indikátorům měřené latentní proměnné – konstruktu) je stejný napříč testovými verzemi. Invariance faktorových zátěží pak naznačuje, že konstrukty různých testových verzí mají pro populace testovaných stejný význam (napříč testovými verzemi), což plyne z toho, že je identický vztah mezi konstruktem a odpověďmi testovaných na položky, jež tento konstrukt nebo subkonstrukt měří (Bialosiewicz, Murphy, & Berry, 2013, s. 8–9). Není-li dosaženo alespoň konfigurální ekvivalence, nemůže být již dosaženo ekvivalence na žádné další úrovni. Na této úrovni ekvivalence ale ještě nelze v různých testových verzích porovnávat vztahy latentních nebo manifestních proměnných s jinými proměnnými, nebo průměrné skóry (Anýžová, 2013).

2. **Metrická ekvivalence** (*metric equivalence, measurement unit equivalence*, ekvivalence měřicí jednotky) je další úrovní ekvivalence, která předpokládá, že testové verze mají stejné škály, tj. jednotky měření a rozsah škály. Úroveň metrické ekvivalence zahrnuje konfigurální ekvivalenci a navíc předpokládá, že hodnoty faktorových zátěží (*factor loadings*) napříč testovými verzemi jsou ekvivalentní. Hodnoty faktorových zátěží vyjadřují, nakolik rozdíly v odpovědích testovaných reflektují rozdíly v úrovních konstruktu měřeného. Při zkoumání metrické ekvivalence se porovnávají fity metrického a konfigurálního modelu a pokud mezi nimi není signifikantní rozdíl, lze považovat faktorové zátěže za ekvivalentní napříč testovými verzemi. Pokud jsou naplněny tyto požadavky, pak je měřený konstrukt v každé z testovaných populací interpretovaný stejně. Je-li dosaženo metrické invariance, lze porovnávat faktorové variance a kovariance (Bialosiewicz, Murphy, & Berry, 2013, s. 8–9). Na této úrovni nelze v různých testových verzích porovnávat průměrné skóry, lze ale porovnávat vztahy latentních nebo manifestních proměnných s jinými proměnnými.
3. **Skalární ekvivalence** (*scalar equivalence*) je další úrovní ekvivalence a její dosažení znamená, že pozorované skóry položek vztažené k příslušnému faktoru mají shodný počátek napříč verzemi a skóry jsou totožně interpretovány (Anýžová, 2013). V takovém případě lze již přímo porovnávat průměrné skóry jedinců napříč testovými verzemi.

4. **Striktní ekvivalence** (*strict equivalence*) má dvě podúrovně. Na první úrovni porovnáváme, zda jsou shodné variance faktorů (*invariance of factor variances*). Dále nás zajímá, do jaké míry jsou shodné chyby indikátorů jednotlivých proměnných (*invariance of individual indicator variable's error*). Naplnění podmínky striktní ekvivalence umožňuje porovnávání reliability položek (Anýžová, 2013; Bialosiewicz, Murphy, & Berry, 2013).

3.2 Metody zkoumání a způsoby dosahování srovnatelnosti testových verzí

Jak jsme uvedli výše, mají-li být porovnávány různé verze téhož testu, měly by být naplněny určité vstupní podmínky. Měl by být měřen podobný nebo stejný konstrukt podobným nebo stejným způsobem, testové verze by měly mít stejný účel a výsledky (skóry) by měly být interpretovány stejným způsobem a vykazovat podobné nebo shodné psychometrické vlastnosti. Základní podmínkou ale je, aby testové verze měly něco společného: např. shodnou definici konstruktu a specifikace; koná je stejná neboekvivalentní skupina testovaných; díky designu administrace se překrývají v některých položkách; porovnávané verze sdílí položky s nějakým jiným, třetím testem (podrobněji v oddílech 3.2.3 a 3.2.4). Můžeme totiž jen těžko porovnávat jevy, objekty nebo skutečnosti, které spolu nijak nesouvisí.

Lze shrnout, že nejprve je třeba zjistit, do jaké míry jsou testové verze porovnatelné (viz oddíl 3.1.5) a poté rozhodnout, jakým způsobem a zda vůbec je možné přikročit k případné transformaci těchto skóru za účelem dosažení jejich ekvivalence, a tedy shodné interpretace. V následujících oddílech představujeme některé z těchto metod a postupů.

3.2.1 Analýza struktury obsahu

Analýza struktury obsahu je procesem využívajícím subjektivní posuzování, tj. lidský úsudek. Při posuzování testu mohou mít posuzovatelé různé úkoly, například a) určit, co položky v testu ověřují, b) do jaké míry se jednotlivé položky vztahují k obsahu vymezenému např. kurikulem, sylabem, specifikací testu, c) jaký je vztah položek, specifikace testu a kurikula, d) jak reprezentativní výběr vzhledem ke kurikulu položky tvoří, e) jak dobře položky reprezentují konstrukt, který je testem měřen. Součástí analýzy struktury obsahu může být také posouzení

charakteristik testových úloh nebo položek (formát, instrukce, kontext, komunikační situace apod.), neboť výkon v testu je ovlivněn mj. interakcí schopností testovaného s charakteristikami testové úlohy (Bachman, 1990; Bachman & Palmer, 1996 a 2010; Khalifa & Weir, 2009).

Uvedené kroky analýzy struktury obsahu s sebou nesou řadu předpokladů, které musí být naplněny. Tyto předpoklady nebo podmínky související s výběrem a charakteristikami posuzovatelů, s úkoly, které mají plnit, a s metodami, jakými jsou vyhodnocovány výsledky. Posuzovatelé by měli být pečlivě vybráni podle kritérií, která by měla korespondovat s účelem analýzy struktury obsahu (např. úroveň znalosti cílového jazyka, obeznámenost s kontextem zkoušky, profesní charakteristiky). Dále je třeba zvážit, jaké školení a v jakém rozsahu je třeba posuzovatelům poskytnout, s jakými materiály budou pracovat, jakou metodou budou obsahovou analýzu provádět, jak se budou sbírat data a jak se budou data vyhodnocovat a interpretovat (o expertním posuzování např. Alderson, 1993; Lumley, 1993; Kirkebøen, 2009; Popham, 1978). Musí být také stanoveno, k čemu všemu a jakým způsobem by se měli posuzovatelé vyjadřovat. Z některých studií např. vyplývá, že požadovat na posuzovatelích odhad obtížnosti testových položek je přinejmenším sporné (Alderson & Lukmani, 1989, Alderson 1990a, 1990b; Norman Verhelst, 2012, osobní korespondence). Podle Aldersona (1990a, 1990b) je pro posuzovatele obtížné shodnout se na tom, co daná položka ověřuje za dovednost. Na druhou stranu může problém přiřazení určité položky k dovednosti nebo znalosti z popisných modelů spočívat i v tom, jak detailně jsou formulovány popisné kategorie, ke kterým mají posuzovatelé položky přiřazovat, a na tom, jak podobné si tyto kategorie jsou, neboť například obsahový překryv může způsobovat posuzovatelům problém rozhodnout se pouze pro jednu kategorii. Obtížnost rozhodování o vztahu položky k popisné kategorii může být způsobena i tím, že jednotlivé popisné kategorie fungují dobře tehdy, je-li popisovaný konstrukt skutečně oddělitelný na diskrétní, vzájemně nezávislé elementy.

Analýzou struktury obsahu se zabývá např. Wuová (2014). Autorka vychází ze socio-kognitivního rámce[23] představeného Weirem (2005, s. 5). Při analýze testu pracuje tento srozumitelný a relativně jednoduchý model mj. s kontextovým a kognitivním aspektem validity, jejichž složky mohou fungovat jako deskriptory při obsahové analýze. Jiný přístup

23 Pro validaci cambridgeských zkoušek Main Suite Cambridge ESOL Examinations.

popisují Bachman, Davidson a Milanovic (1996). Tito autoři vycházejí z Bachmanovy[24] definice komunikační jazykové schopnosti (dále KJS). Komponenty KJS spolu s charakteristikami testovacích technik jsou předmětem posuzování v různých verzích téhož testu. Prezentované výsledky naznačují, že použitá metoda a postup fungují při identifikaci shod a rozdílů v ověřovaném obsahu mezi verzemi testů, a to i přes autory naznačená slabá místa např. ve formulaci komponentů KJS. Nicméně Bachmanova kategorizace KJS, ačkoli je teoreticky zakotvená v modelu komunikační kompetence, není příliš běžná a v prostředí zvyklém na popis komunikační jazykové kompetence podle SERRJ, včetně referenčních úrovní, a na terminologii a kategorizace bližší spíše Weirovu socio-kognitivnímu modelu (2005) by mohla být obtížně aplikovatelná. Jako schůdnější se proto jeví využít pro analýzu struktury obsahu, resp. pro stanovení kategorií popisujících obsah testu spíše model odvozený od Weirova modelu, nebo model SERRJ. Jako velmi jasný a transparentní se ale jeví Bachmanem popsaný způsob provedení obsahové analýzy. Zdá se též jednoduše aplikovatelný jak z hlediska školení posuzovatelů, tak z pohledu následného statistického vyhodnocení a interpretace výsledků.

Zde pod pojmem analýza struktury obsahu rozumíme proces, při kterém byl analyzován jazykový test z pohledu toho, co měří či ověřuje z hlediska obsahu (jazykové znalosti a řečové dovednosti a mikrodovednosti), a jaké kognitivní procesy probíhají při řešení testu. Pro analýzu byly využity popisné modely a kategorie, ke kterým panelisté vztahovali svá posuzování (podrobněji o popisných modelech a analýze struktury obsahu viz oddíly 3.2.1 a 4.1). Výběr posuzovatelů, jejich školení a kvalita popisného modelu jsou klíčovými aspekty při použití této metody (Alderson, 1993, Kirkebøen, 2009, Lumley, 1993, Popham, 1978), proto byli panelisté vybráni podle předem stanovených charakteristik (např. znalost obsahu, kontextu, zkušenost s popisným modelem, reprezentativita apod.). Jejich úkolem bylo přiřadit testové položky jednotlivých testových verzí k popisným kategoriím (testovým specifikacím, popisným modelům apod.). Cílem analýzy struktury obsahu bylo u jednotlivých verzí zjistit pokrytí obsahu položkami a napříč testovými verzemi zjistit, do jaké míry je struktura obsahu podobná.

24 Communicative language ability (Bachman, 1990), později communicative language knowledge (Bachman, 2010)

3.2.2 Zkoumání konstruktové ekvivalence

Ekvivalencí konstruktu nazýváme stav, kdy testové verze prokazatelně měří stejný konstrukt napříč všemi studovanými skupinami. Postup analýzy ekvivalence konstruktů popisují např. Sireci a Allalouf (2003) u různých jazykových verzí téhož testu, který zadali rozsáhlému vzorku testovaných, a také na simulovaných datech. Tvrdí, že i když jde jen o překlad testu do různých jazyků bez jakýchkoli jiných úprav a zásahů, nelze předpokládat, že jsou testy z hlediska konstruktu ekvivalentní, a je třeba jejich ekvivalenci dokázat (2003, s. 4). Takový důkaz je důležitý i z hlediska konstruktové validity, neboť překladem může docházet např. k vnesení konstruktově irelevantních prvků, tedy takových, které přímo s ověřovaným konstruktem nesouvisí, a které přesto mohou ovlivnit výsledek (k tématu též Haupt & Koch, 2012). Stejně je tomu také pro jakékoli další tzv. adaptace téhož testu pro různé skupiny testovaných, např. úpravy pro skupiny se specifickými potřebami (se zrakovým či sluchovým postižením, dyslexie apod.), a rovněž pro vytváření verzí testu odvozením od stejných specifikací, ale s využitím jiného obsahu.

Pro analýzu konstruktu se obvykle využívají kvantitativní metody, jejichž cílem je zjistit nebo potvrdit vnitřní strukturu konstruktu. Na rozdíl od analýzy struktury obsahu nepracují tyto metody se subjektivním posuzováním, nýbrž s daty z testování, v našem případě s odpověďmi testovaných na testové položky. Obvykle jsou využívány postupy strukturního modelování (SEM) a konfirmatorní nebo exploratorní faktorová analýza.

Exploratorní přístupy nevyžadují předem specifikaci modelu neboli hypotézu o struktuře konstruktu; hledají nejjednodušší interpretovatelnou strukturu dat. Příkladem exploratorního přístupu jsou analýza hlavních komponent (PCA) a exploratorní faktorová analýza (EFA). Smyslem exploratorních přístupů je najít společné rysy položek a vztahy mezi položkami a sdružit je do menšího počtu skupin, kde se určitá skupina položek vztahuje k nově vytvořené latentní proměnné = faktoru a je tímto faktorem vysvětlována. Úkolem výzkumníka je z nabízených možností najít takovou strukturu faktorů, která nejlépe vysvětluje vztahy mezi položkami, a smysluplně tyto faktory interpretovat.

Konfirmatorní přístup, který uplatňuje konfirmatorní faktorová analýza (CFA) nebo modelování strukturními rovnicemi (SEM), naopak vyžaduje od výzkumníka nejprve zformulovat hypotézu o struktuře vztahů mezi položkami, resp. proměnnými, specifikovat model s navrženými faktory a vzájemnými vztahy položek (manifestních proměnných).

Následně provedená analýza poté ověří shodu vstupního modelu s reálnými daty. Úkolem výzkumníka je posoudit míru shody modelu s daty a interpretovat akceptovatelnost výstupních hodnot.

3.2.3 Statistické postupy vedoucí k porovnatelnosti skórů

Metody popisované v předchozích oddílech se uplatňují při zkoumání srovnatelnosti obsahové a konstruktové. V tomto oddíle představíme některé z mnoha postupů, jejichž cílem je umožnit porovnání nebo dosažení ekvivalence skórů. Souhrnně je označujeme jako transformace skórů[25]. Transformaci chápeme jako obecný pojem zastřešující záměrnou a teoreticky i empiricky podloženou obvykle statistickou manipulaci se skóry testů, respektive testových verzí. V kontextu tohoto výzkumu je účelem transformace dosažení porovnatelnosti skórů z různých testových verzí, nebo jejich ekvivalence, a tudíž zaměnitelnosti.

Transformaci lze uplatňovat v různých fázích vývoje a sestavování testových verzí, obvykle se tak děje po pretestacích nebo po ostrém testování. Porovnávané testové verze musí mít vždy něco společného (vzorek testovaných, položky, externí kritérium apod.). Lze tedy za určitých předpokladů porovnávat i dva rozdílné testy (nikoli tedy verze téhož testu), pokud je konají stejní testovaní. V takovém případě ale výzkumník musí mít dobré důvody pro takové porovnávání a vhodně interpretovat zjištění a závěry, které z tohoto porovnání vyvodí. Pokud lze smysluplně porovnávat jen určité aspekty, pak je nutné zvážit, zda je možné dosáhnout skutečné ekvivalence skórů a jejich zaměnitelnosti.

Pro aplikaci transformačních postupů je však základní podmínkou prokázání obsahové a konstruktové ekvivalence. Livingston (2004), Dorans, Moses a Eignor (2010), Weir (2005) i další autoři se ale shodují na tom, že ani testové verze, které jsou konstruktově ekvivalentní, v praxi nemohou vykazovat zcela shodné psychometrické charakteristiky. Jedním z hlavních důvodů je fakt, že konstrukt je v každé testové verzi realizován v jiném kontextu – testové verze obsahují různé texty, stimuly, znění položek atd., takže obtížnost, diskriminační schopnost položek a další statistiky se napříč testovými verzemi odlišují. Proto je obvykle třeba přistoupit k transformaci skórů.

25 Tento termín používáme účelově pro tento dizertační projekt, jsme si vědomi, že ve statistice a psychometrice má poněkud jiný význam.

Postupů pro transformaci skórů je poměrně mnoho. Můžeme na ně nahlížet skrze různá kritéria: podle toho, zda využívají pozorovaný skór (*observed score*), nebo pravý skór (*true score*), klasickou teorií testů, nebo teorii odpovědi na položku, zda používají neúplný design a kotvicí položky, nebo nikoli, lineární, nebo nelineární metody vyrovnávání, nebo podle toho, v jaké fázi vývoje a sestavování testových verzí se uplatňují apod.

O postupech transformace založených na využití klasické teorie testů a pozorovaných skórech pojednává velmi podrobně především Livingston (2004). Livingston tyto postupy souhrnně označuje jako vyrovnávání skórů (*score equating*). Vyrovnávání je statistický postup, který upravuje skóry testových verzí tak, aby mohl být považovány za zaměnitelné, přičemž upravuje rozdíl v obtížnosti testových verzí, nikoli v jejich obsahu (Kolen & Brennan, 2014, s. 2–3).

Livingston (2004) v zásadě rozlišuje dva základní postupy: lineární vyrovnávání (*linear equating*) a ekvipercentilové vyrovnávání (*equipercentile equating*). Při vyrovnávání hraje klíčovou roli design, jehož prostřednictvím byla nasbírána data, a to, zda je využito kotvení. Livingston uvádí pět základních typů designu a komentuje podmínky pro jejich využití. Jsou to jednoskupinový design, design ekvivalentních skupin, zkřížený design, design s interním kotvením a design s externím kotvením. Vhodná kombinace postupu vyrovnávání a designu sběru dat je určující pro míru přesnosti a spolehlivosti celého procesu vyrovnávání.

Transformace skórů mohou probíhat i v paradigmatu teorie odpovědi na položku (IRT). Platí ovšem podobné podmínky – pro aplikaci určitého postupu vyrovnávání skórů je potřeba zvážit vhodný design a to, zda bude skór odvozen od pozorovaného skóru, nebo nikoli. Jednou zvýhod využití IRT je širší škála možností provazování testových verzí. Nevýhodou jsou však daleko větší nároky IRT např. na velikost výzkumného souboru, zkušenost s přípravou i interpretací dat, relativní náročnost analýz a požadavky na speciální software. Aby byly být skóry z různých verzí porovnatelné, musí být obvykle transformovány a převedeny na společnou škálu tak, že pro každý původní hrubý skór existuje na společné škále nový transformovaný skór, a původní (referenční) skóry a skóry reportované (transformované) musí mít stále stejnou relativní pozici na škále bez ohledu na to, jaká metoda byla pro vyrovnávání použita.

Livingston (2004) popisuje proces vyrovnávání v zásadě jako sled univerzálních kroků: nejprve je provedena kontrola vlastností skupin testovaných s určitým průměrem a směrodatnou odchylkou skórů (pro skupiny příliš odlišné z hlediska distribuce měřené vlastnosti je problematické vyrovnávat obtížnost testových verzí), následuje volba postupu, výpočet vyrovnávací rovnice, převedení skórů na novou škálu pomocí této rovnice (s. 7). Podle Livingstona (2004) existují určitá omezení vyrovnávacích metod, se kterými je třeba při výběru metody a interpretaci výsledků počítat. Platí, že výstupy vyrovnávání jsou poměrně přesné pro určitou cílovou skupinu, ale již méně pro jednotlivce; čím podobnější jsou si vyrovnávané verze, tím přesnější bude zobecnění z cílové skupiny na jinou. Další komplikací je, že skórová škála s celými čísly může vést při zaokrouhlování ke ztrátě přesnosti nebo že původní a nová škála mohou mít různé jednotky, počátek a maximum, a reportované skóry se tak mohou dostat pod nebo nad hranice nové škály.

Zatím jsme používali termín transformace jako obecné, zastřešující označení pro manipulaci se skóry, a pojem vyrovnávání pro postup, kterým se dosahuje ekvivalence skórů. Ovšem terminologie používaná pro postupy vedoucí k porovnatelnosti či ekvivalenci skórů není zcela jednotná, různí autoři dávají stejným termínům odlišné obsahy, a naopak, stejné věci někdy nazývají různými jmény. Pojem *linking* je např. von Davierovou (2011, s. 2) používán v několika rovinách: jako obecný termín pro označení vztahu (mezi skóry, parametry položek apod.) testových verzí a jako označení pro statisticky méně silnou podobu vyrovnávání (*equating*) a dále jako synonymum pro kalibraci, tj. proces převodu parametrů položek testových verzí na společnou škálu pomocí IRT. Poměrně uceleně diskutují tyto pojmy také Chen, Huang a MacGregor (2009). Autoři připouštějí, že terminologie používaná pro procesy související s provazováním či vyrovnáváním skórů je nejasná a závisí na interpretačním a klasifikačním rámci, ke kterému se autoři nebo uživatelé přiklání, a dále na cíli porovnávání skórů. Oni sami identifikovali několik různých postupů (*equating, linking, concordance, expectation, projection, prediction, calibration, moderation*) a několik klasifikačních rámců: zmiňme např rámec Kolena a Brennana (2004), Feuera a kol. (1999), Hollanda a Doranse (2006), dále Mislevyho (1992) a Linnovu taxonomii (1993). Chen, Huang a MacGregor (2009) ve svém přehledu považují vyrovnávání za speciální příklad provazování, protože při něm musí být naplněny poměrně striktní podmínky a předpoklady.

Chen, Huang a MacGregor (2009), Livingston (2004), Dorans, Moses a Eignor (2010), Kolen a Brennan (2014) a další autoři se však shodují v tom, že za vyrovnané verze lze považovat verze, jejichž skóry byly získány z konstruktově ekvivalentních verzí; vyrovnávací transformace obou verzí jsou vzájemně inverzní a platí symetrická zaměnitelnost skórů z obou porovnávaných verzí; vyrovnávané verze mají podobnou reliabilitu; naplňují podmínku rovnosti (*equity),* tj. testovaným by mělo být zcela lhostejné, kterou z verzí konají, neboť jejich vyrovnávání není ovlivněno tím, jaká podskupina populace byla pro vyrovnávání použita (*group invariance*). Další podmínky, které je třeba naplnit nebo zvážit před vyrovnáváním a které se týkají volby designu, statistického postupu a metody vyrovnávání a přístupu k validaci výsledku vyrovnávání velmi podrobně popisují a diskutují např. Dorans, Moses a Eignor (2010) nebo Kolen a Brennan (2014).

Vybrané postupy pro vyrovnávání skórů

Nejjednodušším způsobem vyrovnávání je podle Livingstona (2004) **vyrovnávání průměrů** *(mean equating)*. Jde o posun skórů na škále tak, aby průměry testových verzí měly na nové škále stejnou pozici vůči ostatním skórům, například tak, že se ke každému skóru nové verze přičte tolik bodů, kolik činí bodový rozdíl mezi průměry porovnávaných verzí. Tento postup je velmi jednoduchý, pokud mají vyrovnávané testové verze stejnou distribuci skórů a srovnatelné psychometrické charakteristiky. Problematické však může být vyrovnávání v případě, že jedna nebo obě testové verze mají odlišnou (nesymetrickou) distribuci skórů na obou stranách průměru, nebo se vzájemně liší distribucí průměrů, případně položky v obou verzích vykazují odlišnou diskriminační schopnost. Pak by nemusela být naplněna podmínka stejné relativní pozice pozorovaných a pravých skórů testovaných. Kvůli tomuto riziku nebývá tento jednoduchý postup doporučován.

Lineární vyrovnávání *(linear equating)* vychází při převodu na novou škálu z relativní pozice skóru, která je dána průměrem a směrodatnou odchylkou. Vztah mezi původními a transformovanými skóry lze popsat lineární rovnicí, která určuje počátek a sklon transformační přímky. Zde existují možná rizika: za prvé, rozsah nové škály se může lišit od rozsahu škály transformované a některé skóry tedy mohou být pod minimem nebo nad maximem rozsahu nové škály; výsledné skóry

nemusejí být celá čísla; za druhé, sklon přímky je závislý na charakteristice skupiny testovaných, na distribuci měřené proměnné ve skupině testovaných (pro některé skupiny může být určitá testová verze snazší než pro skupiny jiné). Tyto nevýhody může odstranit využití ekvipercentilového vyrovnávání.

Ekvipercentilové vyrovnávání *(equipercentile equating)* je nelineární transformací a vychází z odlišné definice relativní pozice skóru, jež není dána průměrem a směrodatnou odchylkou, nýbrž tím, v jakém percentilovém ranku testované skupiny se skór nachází (Livingston, 2004, s. 17). Tato transformace sice zachovává distribuci skórů, ale neděje se lineárně. Pouze v případě, že jsou distribuce skórů v obou porovnávaných skupinách (testových verzích) stejné, pak se ekvipercentilové vyrovnávání v podstatě blíží výsledkům lineárního vyrovnávání. Pokud je ale distribuce skórů v porovnávaných testových verzích (skupinách) odlišná, mělo by být aplikováno vyhlazení (*smoothing*) (Livingston, 2004, s. 20–22). Vždy platí, že proces vyrovnávání je relativně přesný, pokud vyvozujeme závěry o celé skupině testovaných, a daleko méně přesný pro jednotlivce.

3.2.4 Designy využívané pro vyrovnávání

Při vyrovnávání hraje důležitou roli design, jehož prostřednictvím byla nasbírána data, tedy to, jak byly definovány soubory testovaných, jak byly koncipovány testové verze, jejichž skóry mají být porovnávány nebo vyrovnávány, zda bylo využito kotvení a jaké statistické postupy budou využity; zda klasická teorie testů (Classical Test Theory – CTT), nebo teorie odpovědi na položku (Item Response Theory – IRT). Možných designů s různými kombinacemi těchto parametrů je poměrně mnoho, níže proto uvádíme jen některé z nich. Bez ohledu na zvolený design platí, že administrace testových verzí musí probíhat identickým způsobem.

Design bez společných položek

Nejprve uvádíme designy, při nichž se administrují obvykle dvě na sobě nezávislé (nepropojené) testové verze velmi rozsáhlým souborům testovaných, u nichž lze díky velikosti a očekávané heterogenitě dané náhodným výběrem usuzovat na jejich ekvivalenci. Porovnávaných verzí může být i více než dvě, princip zůstává stejný, jen design a postupy jsou složitější. U těchto designů jsou zafixované vlastnosti testovaných (stejní

testovaní konají různé verze) a odlišnosti v dosažených výsledcích lze připisovat charakteristikám testových verzí. Z rozdílů v úspěšnostech (skórech) osob se stejnou úrovní měřeného rysu v různých verzích lze statisticky usuzovat na tzv. rozdíl pravých skórů, a tento rozdíl využít jako vstupní informaci při vyrovnávání.

Při využití **designu náhodných skupin** (*random groups design*) jsou porovnávané testové verze zadávány souboru testovaných podle náhodného klíče, z čehož vzniknou (např. dvě) náhodně rozdělené skupiny. Tento design však vyžaduje rozsáhlý soubor testovaných osob, aby bylo možné dosáhnout náhodného rozdělení a zároveň usuzovat na shodné rozložení sledované charakteristiky. Pokud by platil předpoklad shodného rozložení měřené proměnné v obou skupinách, pak lze výsledky obou testových verzí porovnávat přímo (Jelínek, Květoň, & Vobořil, 2011).

Sběr dat může probíhat také tak, že jsou obě verze, jak referenční, tak vyrovnávaná verze, administrovány stejným respondentům. Pak hovoříme o tzv. **jednoskupinovém designu** (*single group design*), přičemž se vychází z předpokladu, že je možné zobecnit vztah mezi oběma testovými verzemi z této skupiny testovaných na cílovou populaci. Jednoskupinový design je statisticky silný, může však znásobit riziko prozrazení, u testovaných se také může projevit únava nebo efekt učení se z první konané testové verze.

Při použití **zkříženého designu** (*counter-balanced design)* se postupuje tak, že dvě skupiny testovaných konají dvě testové verze, ale každá skupina v opačném pořadí (skupina 1 nejprve koná verzi A a poté verzi B, skupina 2 nejprve verzi B a poté verzi A). Cílem je eliminovat vliv pořadí (*order effect*), ve kterém skupiny testovaných konají obě testové verze, a efekt učení se z první konané verze. Výhodou tohoto designu jsou nižší nároky na počet respondentů než u jednoskupinového designu nebo designu náhodných skupin. I u zkříženého designu platí, žejde o statisticky silný design, ale může představovat bezpečnostní riziko pro použité testové verze, a že se u testovaných může projevit únava; efekt učení se z první konané testové verze je zde ale zkřížením eliminován.

Pokud z nějakých důvodů (časová zátěž, administrativní nebo legislativní důvody apod.) není možné administrovat obě testové verze v jednoskupinovém nebo zkříženém designu nebo designu náhodných skupin, lze také uplatnit design **ekvivalentních skupin** (*equivalent groups design*). Tento design vyžaduje, aby skupiny testovaných byly ekvivalentní, co se týče vnějších (počet, věk, pohlaví atd.) i latentních charakteristik (např. rozložení měřené dovednosti).

Design se společnými položkami

Můžeme také uplatnit design, kdy jsou zadávány skupinám testovaných testové verze tak, aby tyto verze byly propojeny společnými položkami. Jde o tzv. kotvení, též zakotvování (*anchoring*). Skrze výsledky ve společných položkách, jejichž obtížnost předpokládáme jako neměnnou, je možné porovnávat výkony testovaných v těch částech testu, které se pro obě porovnávané skupiny lišily. Pokud byly skupiny testovaných ekvivalentní, pak jsou zafixované vlastnosti kotvicích položek a odlišnosti ve výsledcích lze připisovat charakteristikám testovaných – tedy měřené vlastnosti. Z rozdílu mezi výkony obou skupin testovaných ve společných položkách a ve zbytku testu lze vyvodit rozdíl v obtížnostech testových verzí. Design se společnými položkami lze za určitých podmínek použít i pro neekvivalentní skupiny (*non-equivalent groups anchor test* – NEAT design). NEAT design pracuje s tzv. syntetickou populací odvozenou od obou skupin. Protože je obtížně v tomto designu odlišit rozdíly ve výsledcích dané rozdílnými schopnostmi skupin testovaných a odlišnosti dané rozdílnou obtížností testových verzí, je klíčové věnovat pozornost výběru kotvicích položek. Mezi metody obvykle využívané pro vyrovnávání v designu neekvivalentních skupin s kotvicími položkami patří Tuckerova metoda, vhodnější pro testové verze s rozdílnou obtížností, a Levineova metoda, vhodnější pro testové verze s rozdílnými distribucemi skórů u obou skupin (Kolen & Brennan, 2014).

Rozlišujeme dva základní způsoby kotvení. **Vnitřní kotvení** *(internal anchoring)* využívá tzv. kotvicí položky nebo úlohy, které jsou společné oběma testovým verzím. Klíčové je, aby kotvicích položek byl dostatečný počet a aby tyto položky dobře reprezentovaly celý měřený konstrukt, neboť chceme tvrdit, že rozdíl ve skórech získaných v kotvicích položkách reflektuje rozdíl v dosažené úrovni měřeného konstruktu v porovnávaných testových verzích. Tomuto designu se také někdy říká *linked* nebo *incomplete design* – propojený design. Podle Livingstona (2004) je největší výhodou vnitřního kotvení jednoduchá administrace, na druhé straně jeho největší nevýhodou je složitý proces vývoje testových verzí a to, že kotvicí položky se objevují v administrovaných testových verzích opakovaně, což může znamenat bezpečnostní riziko. Pro **externí kotvení** *(external anchoring)* je třeba, aby skupiny testovaných konaly kromě dvou různých porovnávaných testových verzí ještě jeden stejný test, který zde funguje jako kotvicí. Podobně jako

u vnitřního kotvení, i zde je třeba, aby tento test, resp. položky v něm obsažené, dobře reprezentovaly měřený konstrukt, společný i oběma porovnávaným testovým verzím.

Výběr kotvicích položek jako klíčový aspekt kotvení

Zcela základním předpokladem pro uplatnění kotvení a výběr kotvicích položek je celková kvalita testu, z něhož jsou položky vybírány, dále hraje klíčovou roli reprezentativita vybraných kotvicích položek vzhledem ke konstruktu celého testu (van den Heuvel-Panhuizen a kol., 2009). Kotvicí položky musí být pro testované zcela nové a musí co nejvíce reflektovat obsah, konstrukt a formát vyrovnávaných verzí. Je důležitá také jejich pozice, neboť na obtížnost položek může mít vliv i kontext, ve kterém se nacházejí. Kotvicí položky by měly vykazovat dobrou korelaci s testem. Počet kotvicích položek je dalším zásadním parametrem. Podle Livingstona (2004) by kotvicích položek měla být přibližně jedna pětina z celkového počtu položek v testové verzi a rozsah obtížnosti kotvicích položek by měl pokrýt celou škálu obtížnosti testu (Livingston, 2004), aby bylo možné co nejvíce snížit statistickou chybu při vyrovnávání. Pokud se u kotvicích položek zjistí při testování nějaký problém, nesmí být měněny, je možné pouze některé položky vyřadit. Zjednodušeně můžeme shrnout, že kotvicí položky by měly obsahovými a statistickými vlastnostmi dobře reprezentovat celý test (Kolen & Brennan, 2014).

S tím souvisí problém typický pro jazykové testy: ty obvykle sestávají nikoli z jednotlivých izolovaných položek, nýbrž ze sad – úloh, které se obvykle váží ke společnému stimulu (textu, obrázku apod.) a používají stejnou testovací techniku. V případě jazykových testů musíme tedy uvažovat při výběru kotvicích položek ve smyslu těchto celků. Aby kotvicí úlohy (nikoli pouze položky) reprezentovaly dobře konstrukt, je pravděpodobné, že jich bude muset být více než změněných 20 %.

3.2.5 Využití teorie odpovědi na položku

Výše uvedené postupy vyrovnávání, u kterých např. Livingstonem (2004) popisuje využití klasické teorie testů (CTT) a zpracování pozorovaných nebo pravých skórů[26], je možné využít i v paradigmatu teorie odpovědi na položku (IRT). Základní principy a podmínky pro volbu designu a metody a pro validaci postupu jsou shodné. Při využití CTT se pomocí vyrovnávacích rovnic vytváří v podstatě tabulka pro převod skórů. Tato tabulka obsahuje skóry z referenční i nové, vyrovnávané verze. Vyrovnávané skóry se pomocí lineární nebo ekvipercentilové transformace převedou na reportovanou škálu. Vyrovnávání v rámci IRT pracuje s odhadem úrovně latentního rysu zohledňujícího jak schopnosti testovaného, tak zároveň obtížnost testových položek. V IRT nejsou odhady parametrů testovaných závislé na specifické sadě položek, odhady parametrů položek nezávisí na skupině testovaných, která tyto položky řešila. Na druhé straně ale využití IRT vyžaduje naplnění přísnějších podmínek a větší soubory testovaných. Pokud jsou tyto podmínky pro využití některého z modelů IRT naplněny, samotné vyrovnávání již není komplikované a jde v podstatě o lineární transformaci. U jednoskupinového designu se obě verze tzv. kalibrují společně, výsledky jsou již na společné škále a není potřeba žádného vyrovnávání, pouze převodu na reportovanou škálu (společnou oběma vyrovnávaným verzím). Pro NEAT design lze využít konverzi (conversion), kdy se vyrovnávané testové verze pomocí IRT kalibrují odděleně. Nejprve se vypočítají parametry testových verzí odděleně a poté se vypočte vyrovnávací rovnice pro konverzi skórů. Při souběžné kalibraci (concurrent calibration) se datové soubory obou testových verzí zkombinují a všechny položky a všichni testovaní (resp. jejich parametry jako obtížnost položek a úroveň měřené schopnosti testovaných) se převedou na společnou škálu. Kalibrace s kotvicími položkami (fixed anchor calibration) kombinuje oba předešlé přístupy. Kalibrace obou verzí probíhá odděleně, a poté se kotvicí položky (jejich parametry) z již kalibrované verze využijí při kalibraci verze druhé, přičemž parametry kotvicích položek jsou fixovány. Tím jsou parametry nové verze kalibrovány na škálu, kterou sdílí s původní verzí. Existují i další metody, které podrobně vykládají např. Kolen a Brennan (2014), ale zevrubné pojednání o těchto metodách je mimo rozsah této publikace.

26 Nikoli proto, že by to bylo jediné možné řešení, nýbrž proto, že jeho kniha je původně výukovým materiálem pro kurzy, jejichž cílem bylo záměrně představit metody jiné než založené na IRT, v prostředí společnosti ETS, které klasickou teorii testů využívalo spíše okrajově a účelově.

II. EMPIRICKÁ ČÁST

4 Metody využité ve výzkumném projektu

V teoretické části I jsme představili koncept srovnatelnosti testových verzí teoreticky a přiblížili některé metody a postupy, které jsou využívány při vývoji a sestavování testových verzí s cílem dosáhnout srovnatelnosti. V této empirické části přibližujeme metody, které byly využity v empirické části výzkumu. Jde jen o některé z metod představených v teoretické části I. Jejich výběr byl ovlivněn snahou co nejvíce napodobit stav a podmínky současné slovenské MZ a pracovat s reálnými daty – skutečnými žákovskými odpověďmi z realizované maturitní zkoušky.

Kapitola 4 tedy popisuje aplikaci vybraných metod v empirické části výzkumu. Nejprve se věnuje analýze struktury obsahu s využitím popisných modelů dle SERRJ a kognitivních procesů za pomoci expertního posuzování; dále analýze ekvivalence konstruktu pomocí faktorové analýzy; poté porovnání psychometrických charakteristik položek i subtestů a populací pro všechny čtyři testové verze. Pracovali jsme s testovými verzemi tak, jak byly zadány, to znamená, že jsme nevyužili záměrně žádný specifický design sběru dat a nevyřazovali jsme žádné položky.

4.1 Obsahová ekvivalence: analýza struktury obsahu

Na tomto místě se věnujeme výstupům z expertního posuzování ekvivalence obsahu testových verzí MZ z AJ na úrovni B1 a prezentujeme výsledky analýzy shody posuzovatelů. Zajímá nás, do jaké míry jsou verze testu z anglického jazyka 2012–2015 obsahově ekvivalentní, tzn., zda lze říci, že je struktura obsahu subtestů testových verzí shodná, zda je shoda posuzovatelů na struktuře obsahu dostatečná, do jaké míry lze výsledky posuzování považovat za spolehlivé, a zda lze strukturu obsahu, jak ji identifikovali posuzovatelé, využít i při interpretaci faktorového řešení exploratorní faktorové analýzy (EFA). Dále bylo naším cílem zjistit, do jaké míry je metoda analýzy struktury obsahu praktickou a spolehlivou metodou pro zkoumání obsahové srovnatelnosti testových verzí, zda jsou

popisné modely (model odvozený od SERRJ a model popisující kognitivní procesy předpokládané při řešení položek) užitečnými a vhodnými nástroji pro popis testových verzí. Zajímalo nás také, zda expertní posuzování vztahu položek a deskriptorů popisných modelů je využitelné a vhodné i v kontextu zkoušky vysoké důležitosti, jakou je i maturitní zkouška, a takto ověřit, zda a za jakých podmínek může analýza struktury obsahu vygenerovat výsledky využitelné při zkoumání konstruktové ekvivalence testových verzí pomocí faktorové analýzy.

4.1.1 Popis metody

Obsahová analýza, jak ji definuje např. Krippendorff (2004), je empirická metoda, která využívá explorační přístup s cílem predikovat nebo vyvozovat závěry, přičemž posuzovatelé analyzují a interpretují vstupní data v souladu s předem danou sadou popisných kategorií. Posuzovatelé při obsahové analýze zkoumají data, texty, obrazy apod. a snaží se pochopit a interpretovat jejich smysl, význam, sdělení apod. (Kripendorff, 2004, s. xviii). Na rozdíl od takto popsané metody zkoumání textových produktů není cílem analýzy struktury obsahu, kterou zde popisujeme, analyzovat text jako takový. Postupy v našem výzkumu aplikujeme na poněkud odlišný kontext. Analyzovaným obsahem se zde přeneseně myslí soubor testových úloh, kde je každá úloha tvořena textem (nebo texty) a k nim připojenými položkami. Tyto úlohy ověřují subkonstrukty, jejichž popis nebo definice jsou předem dány. V případě slovenské MZ jde o konstrukt vztahující se deklarativně k externímu standardu, jímž je SERRJ, resp. referenční úroveň B1 v poslechu, čtení a jazykové kompetenci.

Nicméně vzhledem k již diskutované velké míře obecnosti, s jakou je definován konstrukt slovenské MZ z AJ, nebyly k dispozici žádné informace nebo podklady, které by bylo možné využít jako popisný model nebo z nich takový model odvodit, neboť dostupné specifikace zkouše a katalog požadavků ke zkoušce jsou velmi obecné. Proto jsme přikročili k vytvoření vlastních popisných modelů, které dle našeho soudu konkrétněji a komplexněji postihují podstatu úrovně B1 tak, jak je popsána v SERRJ [27], a tuto úroveň doplňují o pohled na kognitivní procesy probíhající při řešení úloh (podrobněji v oddíle 4.1.2).

27 Výzkumný projekt započal a analýza struktury obsahu proběhla v době, kdy ještě nebyla k dispozici doplněná verze SERRJ, tzv. Companion Volume s novými deskriptory. Vycházeli jsme proto pouze z původní verze SERRJ z roku 2001, resp. z českého překladu z roku 2002. Při případných replikacích nebo při zavedení do praxe bychom doporučili pracovat s rozšířenou verzí SERRJ, tedy s tzv. CEFR: Companion Volume.

Vznikly dva typy popisných modelů. První typ popisných modelů obsahuje kategorie (deskriptory) vztažené k teoretickému konstruktu řečových činností poslech a čtení dle SERRJ, odvozené od tzv. Can-Do statements referenčních úrovní), a ke gramatickým kategoriím jazykové kompetence definovaným Purpurou (2004). Rozhodnutí využít pro tvorbu modelů SERRJ bylo podpořeno tím, že dokumentace zkoušek na webových stránkách NÚCEMu a také interpretace výsledků zkoušek jednoznačně odkazuje k SERRJ jako základnímu dokumentu pro definici konstruktu zkoušek z AJ v EČ MZ.

Druhý typ popisných modelů staví na kognitivních procesech, které by měly probíhat při řešení položek v daném subtestu. Pro řečovou činnost čtení vychází popisný model z Weirova socio-kognitivního modelu (Khalifa & Weir, 2009, Weir, 2005), pro poslech byl částečně adaptován Weirův socio-kognitivní model a doplněn o popisy kognitivních procesů při poslechu na základě zejména Fielda (2009) a Beckera (2016). Popisný model pro subtest gramaticko-lexikální byl odvozen z Purpurova modelu jazykové kompetence (2004, 2014a, 2014b, 2017).

Vstupními informacemi pro posuzovatele byly úlohy (texty a na ně navazující položky) v subtestech Poslech, Gramatika a Čtení z let 2012–2015. Všechny realizované testové verze lze najít na stránkách NÚCEMu. Tyto testové verze jsou z pohledu vnější struktury identické. Celkem obsahuje jedna testová verze 60 položek rozdělených do tří subtestů. Subtest Poslech obsahuje tři úlohy, subtest Gramatika dvě úlohy a subtest Čtení tři úlohy. Úkolem posuzovatelů bylo pomocí popisných modelů určit cíle ověřované položkami v testových verzích.

4.1.2 Popisné modely

Tvorbě popisných modelů předcházelo důkladné studium literatury i odborných diskusí. Zamýšleli jsme se zejména nad tím, do jaké míry lze konstrukt čtení, resp. poslechu a jazykové kompetence dělit na diskrétní jednotky nebo tzv. mikrodovednosti, a popsat je např. deskriptory referenčních úrovní SERRJ, nebo zda jde u jazykové kompetence a řečových činností spíše o konstrukt jednorozměrný, jež dále nelze smysluplně dělit (více např. Khalifa & Weir, 2009; Goh & Aryadoust, 2015). Z dostupné literatury vyplynulo, že snahy o popis konstruktu čtení a poslechu pomocí tzv. mikrodovedností a následná analýza kvantitativními

metodami, jako např. pomocí konfirmatorní faktorové analýzy, nevedly k jednoznačnému potvrzení nebo vyvrácení teorie o parcelaci jazyka, resp. řečovýc činností a dovednosti na sadu diskrétních kategorií. Také SERRJ v deskriptorech referenčních úrovních sice pracuje s oddělenými popisy (Can-Do statements), avšak tyto deskriptory nejsou zamýšleny primárně jako oddělené kategorie, naopak, pouze dohromady mohou poskytovat komplexní pohled na kontinuum referenčních úrovní. Protože však bylo nutné vytvořit nástroj, který by umožnil porovnání obsahové struktury a následně porovnání konstruktu, bylo nutné se pro účely této analýzy přiklonit k tomu, že lze teoreticky popsat položky diskrétními kategoriemi a předpokládat, že lze identifikovat u každé položky či skupiny položek převažující obsah.

Khalifová a Weir (2009) také uvádějí, že přístupy k analýze obsahu nebo konstruktu testů pomocí deskriptorů SERRJ neberou v úvahu charakteristiky testovaných ani kognitivní procesy, které musí nebo by měl testovaný aktivovat při řešení úloh. Tyto kognitivní procesy autoři považují za důležitou složku konstruktu a akcentují ji i v socio-kognitivním rámci.

Z výše uvedeného tedy vyplynulo i rozhodnutí doplnit popisný model zaměřený na obsah (odvozené od SERRJ) o druhý model zaměřený na kognitivní procesy. Popisné modely dle SERRJ pro Čtení a Poslech obsahují deskriptory řečových činností převzaté ze škál úrovně B1 SERRJ. Model pro Gramatiku byl odvozen od Purpurova modelu jazykové kompetence (Purpura, 2004, 2014a, 2014b, 2017). Na obrázcích 1a až 1c je vidět ukázka odvozování modelů, vždy primární zdroj vlevo a výsledný model vpravo. Celé popisné modely obsahuje příloha A.

Obrázek 1a

Ukázka odvozování popisných modelů pro Poslech od SERRJ deskriptorů

	SERRJ deskriptory pro poslech s porozuměním
B1+	Dokáže porozumět nekomplikovaným faktografickým informacím týkajícím se věcí každodenního života a zaměstnání, rozpozná jak obecná sdělení, tak specifické podrobnosti za předpokladu, že jde o zřetelnou výslovnost a všeobecně známý přízvuk.
B1	Dokáže porozumět hlavním myšlenkám vysloveným spisovným jazykem o běžných tématech, se kterými se setkává v práci, ve škole, ve volném čase atd., a to včetně krátkých vyprávění.
B1	Obvykle dokáže sledovat hlavní myšlenky delší diskuse, které je svědkem, za předpokladu, že řeč je spisovná a zřetelná.
B1	Dokáže sledovat s porozuměním hlavní linii krátké jednoduše členěné přednášky na známá témata, pokud je přednesena zřetelně a ve spisovném jazyce.
B1+	Dokáže sledovat s porozuměním přednášku ve svém oboru, pokud jde o známé téma a prezentace je velmi jednoduše členěná a jasně uspořádaná.
B1	Chápe jednoduché technické informace, jako je návod k obsluze předmětů každodenní potřeby. Dokáže porozumět podrobným orientačním pokynům.

	Odvozené popisné kategorie Položka ověřuje:
A	zachycení nekomplikované faktografické (konkrétní) informace.
B	porozumění podrobným orientačním pokynům nebo jednoduchým technickým informacím.
C	porozumění hlavním bodům=důležitým informacím textu/nahrávky o známých (běžných) záležitostech.
D	dovednost sledovat s porozuměním delší nahrávku a pochopení hlavní linie textu.
E	pochopení smyslu/hlavní myšlenky textu (určité části textu).
F	dovednost sledovat s porozuměním delší nahrávku a pochopení hlavní myšlenky/hlavních myšlenek textu.

Obrázek 1b

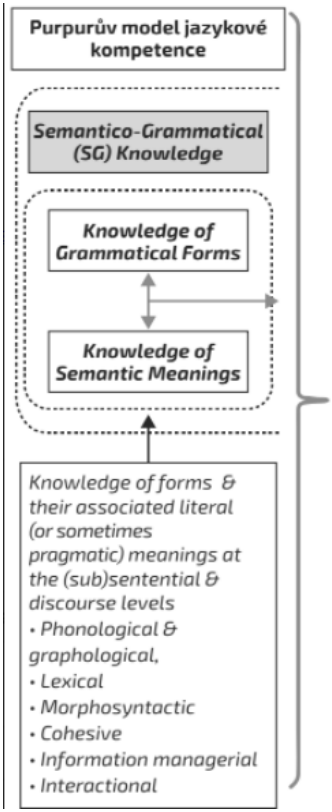
Ukázka odvozování popisných modelů pro Čtení od SERRJ deskriptorů

	SERRJ deskriptory pro poslech s porozuměním
B1+	Dokáže porozumět nekomplikovaným faktografickým informacím týkajícím se věcí každodenního života a zaměstnání, rozpozná jak obecná sdělení, tak specifické podrobnosti za předpokladu, že jde o zřetelnou výslovnost a všeobecně známý přízvuk.
B1	Dokáže porozumět hlavním myšlenkám vysloveným spisovným jazykem o běžných tématech, se kterými se setkává v práci, ve škole, ve volném čase atd., a to včetně krátkých vyprávění.
B1	Obvykle dokáže sledovat hlavní myšlenky delší diskuse, které je svědkem, za předpokladu, že řeč je spisovná a zřetelná.
B1	Dokáže sledovat s porozuměním hlavní linii krátké jednoduše členěné přednášky na známá témata, pokud je přednesena zřetelně a ve spisovném jazyce.
B1+	Dokáže sledovat s porozuměním přednášku ve svém oboru, pokud jde o známé téma a prezentace je velmi jednoduše členěná a jasně uspořádaná.
B1	Chápe jednoduché technické informace, jako je návod k obsluze předmětů každodenní potřeby. Dokáže porozumět podrobným orientačním pokynům.

	Odvozené popisné kategorie Položka ověřuje:
A	zachycení nekomplikované faktografické (konkrétní) informace.
B	porozumění podrobným orientačním pokynům nebo jednoduchým technickým informacím.
C	porozumění hlavním bodům=důležitým informacím textu/nahrávky o známých (běžných) záležitostech.
D	dovednost sledovat s porozuměním delší nahrávku a pochopení hlavní linie textu.
E	pochopení smyslu/hlavní myšlenky textu (určité části textu).
F	dovednost sledovat s porozuměním delší nahrávku a pochopení hlavní myšlenky/hlavních myšlenek textu.

Obrázek 1c

Ukázka odvozování popisných modelů pro Gramatiku od Purpurova modelu jazykové kompetence (levá část schématu převzata z Purpura, 2017)



	Odvozené popisné kategorie Položka ověřuje:
A	znalost a použití pravidel slovotvorby/morfologie.
B	znalost a použití lexika/frazeologie.
C	znalost a použití pravidel morfosyntaxe.
D	znalost a použití pravidel syntaxe (včetně prostředků textové návaznosti na úrovni věty a nadvětných celků).
E	znalost a použití pragmatických významů (vztahy, vyvozování významů, postojů) a jazykových funkcí.

Socio-kognitivní model Khalifové a Weira (2008), který byl východiskem při přemýšlení o koncepci modelů kognitivních procesů, sestává ze tří částí zobrazených jako sloupce (viz obr. 2). Ač byl zpracován primárně autory pro dovednost čtení, domníváme se, a to i na základě literatury, která zmiňuje jeho adaptaci pro další využití v kontextech mimo oblast čtení, že jeho principy jsou přenositelné i do jiných oblastí komunikační kompetence, na dovednost poslech i na jazykovou znalost, resp. kompetenci. Obecně platný je a v různých modelech jazykové kompetence funguje tzv. *goal setter*, který je zodpovědný za aktivaci metakognitivních strategií, které umožní zvolit určitý způsob práce s komunikační úlohou. Volba tohoto způsobu souvisí s charakteristikami úlohy, s jejím cílem (vyhledání informace, porozumění myšlence, volba gramatické formy nebo významu apod.), s kontextem, ve kterém se odehrává, a také s charakteristikami testovaného, včetně jeho úrovně jazykové kompetence (měřeného rysu). Vychází z teoretického předpokladu, že pokud má např. testovaný pracovat s dlouhým textem, měl by k němu přistoupit podle toho, jaký je cíl, úkol, který má splnit: má vyhledávat informace konkrétní povahy, porozumět detailně informacím, které je třeba z textu vyvodit na základě dílčích signálů, nebo porozumět myšlenkám, argumentaci apod. Podle toho aktivuje další procesy (např. volbu typu čtení/poslechu, strategie zpracování textu), v čemž mu právě goal setter napomáhá.

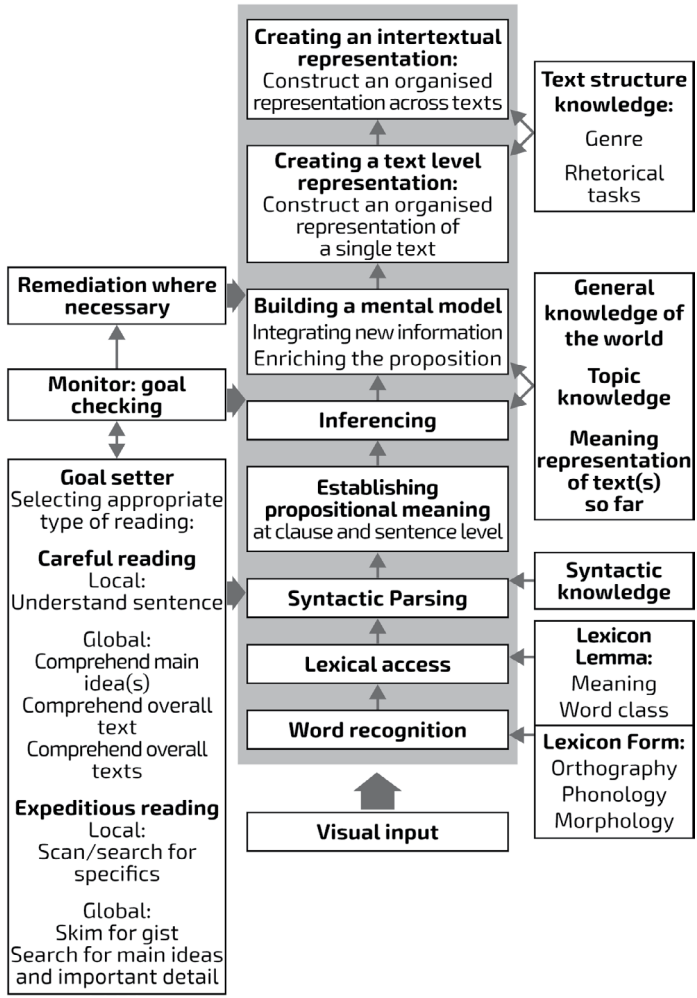
Pro účely posuzování kognitivních procesů spojených s řešením úloh jsme se inspirovali procesy v levé části schématu, které se váží na stanovení účelu čtení, neboť se domníváme, že jsou součástí výuky anglického jazyka, pracuje se s nimi při nastavování kurikula, sylabů apod. a také nejvíce korelují s obsahem deskriptorů SERRJ. Např. pozorné čtení na globální úrovni je předpokladem pro porozumění hlavním myšlenkám textu, volba rychlého čtení na lokální úrovni je strategií pro rychlé vyhledání konkrétní informace apod. Rozhodnutí, jaký je účel čtení (podrobné nebo rychlé čtení na lokální nebo globální úrovni) a jaký typ čtení (*scanning, skimming, search reading, careful reading*) bude zvolen, determinuje následnou aktivaci kognitivních procesů. Plné využití těchto procesů je limitováno úrovní jazykové, pragmatické a strategické kompetence testovaných (Khalifa & Weir, 2009, s. 7). Při vytváření popisného modelu kognitivních procesů pro poslech byla situace poněkud komplikovanější. Nenalezli jsme mnoho jednoznačných a empiricky podložených modelů, nepočítáme-li modely převzaté ze čtení nebo modely pracující s výčty mikrodovedností. Model pro poslech jsme tedy odvozovali od socio-kognitivního modelu pro čtení a zapracovali jsme zejména Fieldovo pojetí poslechu (Field, 2009). Z hlediska zaměření pozornosti a účelu poslechu (povahy činnosti) Field

dělí poslech zhruba na lokální a globální poslech. My jsme ponechali kategorii globálního poslechu, kdy je věnována pozornost s porozuměním textu jako celku, a namísto lokálního jsme pojmenovali příslušnou oblast selektivní poslech, kde předpokládáme, že se posluchač spíše než na celek zaměřuje na jednotlivosti, podrobnosti. Volba těchto procesů by měla být právě ovlivňována zněním položky. Toto dělení podle nás reprezentuje i různou míru pozornosti, kterou je třeba poslouchání věnovat (nižší u selektivního poslechu, vyšší u globálního poslechu), a je zároveň analogické s rychlým a podrobným čtením u modelu čtení. U Gramatiky jsme využili Purpurův model jazykové kompetence (2004), konkrétně jsme převzali kategorie znalost gramatické formy, znalost gramatického významu a znalost pragmatického významu.

Můžeme shrnout, že při zpracování popisných modelů byly uplatněny dva různé pohledy – pohled na obsah testových verzí perspektivou referenčního rámce SERRJ, tedy na to, jaké specifické cíle nebo řečové činnosti ověřují úlohy a položky, a pohled na subtesty z hlediska kognitivních procesů a strategií, o nichž předpokládáme, že probíhají při efektivním řešení testových položek (Council of Europe, 2001; Khalifa & Weir, 2009; Field, 2008; Purpura, 2004). Vycházíme z předpokladu, že oba popisné modely budou do určité míry korelovat, tj. že efektivní řešení položek ověřujících určitou skupinu specifických cílů (podle deskriptorů SERRJ) bude korelovat s určitým kognitivním procesem či strategií.

Obrázek 2

Socio-kognitivní model čtení (převzato z Weir, 2005; Khalifa & Weir, 2009)



4.1.3 Posuzovatelé a postup jejich práce

Analýzy struktury obsahu se zúčastnili čtyři zkušení posuzovatelé[28]. Všichni měli zkušenost s používáním SERRJ ve své praxi, a to při výuce cizího jazyka nebo při práci s hodnocením a testováním, a byli též dobře obeznámeni s teoriemi osvojování cizího jazyka. Jejich úkolem posuzovatelů bylo přiřadit každou testovou položku k příslušnému deskriptoru (kategorii) v obou modelech (příloha A). Nejprve jsme provedli pilotáž celého procesu i nástrojů určených pro sběr dat. Následovala diskuse s posuzovateli o procesu posuzování. Na základě výsledků pilotáže a diskuse byly nástroje poupraveny. Byl také zpracován písemný popis postupů pro posuzovatele, posuzovatelé byli vyškoleni v metodě a interpretaci popisných modelů a provedli jsme zkušební nácvik procesu posuzování a zápisu dat. Po školení byly posuzovatelům zaslány materiály a záznamové listy a byli požádáni, aby individuálně posoudili testové verze, resp. přiřadili každou položku k jednomu z deskriptorů, a to v obou popisných modelech. Zaznamenaná přiřazení zaslali posuzovatelé elektronicky výzkumníkovi.

4.1.4 Testové verze využité pro analýzu struktury obsahu

Posuzovatelé pracovali s testovými verzemi, které byly administrovány hlavní části slovenské maturitní populace v jarním zkušebním termínu v letech 2012–2015.

4.1.5 Data z analýzy struktury obsahu a prvotní rozhodnutí

Data získaná od posuzovatelů měla mít jednotný formát, tzn., že pro každou položku jsme měli od každého hodnotitele v obou modelech získat kód deskriptoru (např. C nebo SG), který položku podle posuzovatele nejlépe charakterizuje. Tabulka 1 je ukázkou struktury hrubých dat získaných od čtyř posuzovatelů H1, H3, H4, H5 pro subtest Poslech 2012 a popisný model s deskriptory SERRJ. První řádek s čísly 1–20 obsahuje čísla položek. Sloupce s písmeny obsahují kódy popisných kategorií (podrobněji viz příloha A – popisné modely) tak, jak byly položkám přiřazeny posuzovateli H1–H5.

28 Původně bylo osloveno pět posuzovatelů, všichni se zúčastnili pilotáže a slíbili pokračovat, nicméně posuzovatel H2 poté nedodal žádná data. Kvůli kontinuitě interpretace jsme už kódy posuzovatelů nepřečíslovávali.

Hrubá data získaná od všech posuzovatelů za každý subtest byla následně sloučena do jednoho celku. Celkem jsme tak získali 36 sad dat.

V datech vidíme šedě zvýrazněnou buňku, označující položku, u které nemáme data. Podobné vynechání se sporadicky objevilo i u jiných subtestů a posuzovatelů. Důvodem mohlo být buď opomenutí přehlédnutí na straně posuzovatele[29], nebo jako je tomu v tomto konkrétním případě, problematická testovací technika[30]. Jde o přiřazovací úlohu v poslechu, kdy jedna z položek (v roce 2012 to byla položka 16) má nulovou odpověď – testovaní měli pouze označit křížkem, že tato položka má zůstat bez přiřazení. Nebylo tedy možné ji zcela posoudit, avšak pro účely našeho výzkumu jsme – pouze v takovémto případě – položce přiřadili stejný deskriptor, jaký byl přiřazen položkám ze stejné části. U jiných úloh, kde se tato technika nevyskytovala, takto nebylo možné uvažovat, ponechali jsme proto položky bez popisu a v analýze jsme je považovali za chybějící hodnoty.

Tabulka 1
Ukázka struktury hrubých dat pro Poslech 2012 získaných od čtyř posuzovatelů

Model SERRJ: vztah položek k deskriptorům SERRJ

ID	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
H1	C	C	C	C	C	C	C	C	C	C	C	C	C	D	D	D	D	D	D	D
H3	D	A	A	D	D	D	D	A	A	D	A	A	D	E	E	E	E	E	E	E
H4	C	C	C	D	C	C	C	C	C	C	C	C	C	D	D		D	D	D	D
H5	C	C	C	C	C	C	C	C	C	C	C	C	C	F	F	F	C=F	C=F	C	C=F

Dalším aspektem, který jsme museli řešit v hrubých datech, byl fakt, že se posuzovatelé v některých případech nedokázali rozhodnout, který deskriptor je pro položku dominantní, a uvedli u některých položek deskriptorů více. Např. v tabulce 1 u modelu SERRJ vidíme, že posuzovatel H5 zařadil položku 17 k deskriptorům *C+F (C=F)*. V některých případech posuzovatelé váhali i mezi třemi deskriptory. Pro analýzy to znamenalo buď multiplikovat každý krok několikrát a provést analýzy oddělené s položkami zařazenými pokaždé do jiné skupiny deskriptorů

29 Z časových důvodů nebylo možné žádat posuzovatele o opravu.

30 Tuto informaci jsme získali v tomto konkrétním případě přímo od posuzovatele, avšak vzhledem k tomu, že byl jediný, kdo to takto pojal, nezasahovali jsme do instrukcí pro ostatní tři posuzovatele.

(nejprve do C, poté do F), což by znamenalo značné množství kombinací pro analýzu dat, nebo rozhodnout místo posuzovatelů o zařazení pouze do jedné z kategorií (do C, nebo do F), nebo uvažovat o sloučení deskriptorů do méně kategorií a získat tak např. tři kategorie místo původních šesti. První řešení považujeme za zásadní manipulaci s daty a za interpretaci dat, která může vést k interpretaci zásadně odlišné od záměru posuzovatelů a může ovlivnit výslednou strukturu obsahu. Proto jsme se přiklonili k tomu, že jsme přistoupili ke sloučení deskriptorů do nadřazených kategorií, a to i na základě důkladné revize a kritického zhodnocení popisných modelů astudia literatury, které se týká práce s podobnými modely (např. Alderson, 1993, Alderson a kol. 2006, Weir, 2005).

Grafický přehled struktury obsahu

Při vypracování přehledu jsme postupovali tak, že při grafickém znázornění struktury obsahu, tedy zastoupení kategorií v subtestech u jednotlivých posuzovatelů a u celé skupiny posuzovatelů, jsme zdvojené deskriptory započítávali jejich vahou. To znamená, že údaj C u položky x jsme započítali jako C=1, údaj C=F jsme započítali jako C=0,5 a F=0,5 z celkové váhy 20 (každý subtest měl 20 položek a tudíž 20 možností přiřazení). I zde jsme pracovali se sloučenými kategoriemi, tzn., že pokud F bylo součástí nadřazené kategorie např. DEF, v grafickém výstupu se objeví kategorie DEF, nikoli F.

Přehled četností shody

Pro výpočet četnosti shody jsme postupovali obdobně. Dva posuzovatelé se mohli shodnout celkem ve dvaceti instancích. Pokud např. oba položce přiřadili C, byla tato shoda započítána hodnotou 1. Pokud jeden přiřadil C, druhý F, byla započítána shoda hodnotou 0. Pokud jeden přiřadil C, druhý C=F, byla shoda započítána hodnotou 0,5. Analogicky byla počítána četnost shody jednoho posuzovatele se zbylými třemi a četnost shody všech čtyř posuzovatelů.

Výpočet koeficientu shody AC1

Ve druhé fázi jsme pro výpočet koeficientu AC1 pracovali vždy s prvním uvedeným deskriptorem a využili jsme nadřazené popisné kategorie, do nichž byly vřazeny původní deskriptory. Pokud tedy např. u Poslechu posuzovatel u položky uvedl D=C, ve vstupních datech pro výpočet AC1 bylo uvedeno D, resp. sloučené DEF.

Problémem, který vyvstal, byla výraznější odlišnost v chování některých posuzovatelů, resp. v datech od nich získaných. Týkalo se to zejména posuzovatele H3, který se u některých subtestů nebo modelů výrazněji odlišoval od ostatních posuzovatelů. Před prováděním analýz jsme se tedy museli rozhodnout, zda data od H3 započíst, nebo zda posuzovatele H3 vyloučit, a pokud vyloučit, tak zda vyloučit všechna jeho posouzení, nebo pouze ta, která vykazují odlišnosti. Po opětovné analýze popisných modelů a důkladném zhodnocení charakteristiky deskriptorů v souvislosti se získanými daty jsme se rozhodli data od posuzovatele H3 v souboru ponechat a zpracovat, neboť jsme neočekávali absolutní shodu a posuzovatel H3 byl v posuzování konzistentní sám se sebou napříč testovými verzemi. Určitou míru odlišností jsme navíc zaznamenali i u ostatních posuzovatelů, a nebylo možné zjistit, zda je odlišné chování způsobeno tím, jak posuzovatelé chápou popisné modely, problematičností vstupních dat, nebo metodou.

Sloučení některých popisných kategorií do nadřazeného celku tedy vyřešilo z větší části problém s položkami, kterým posuzovatelé přiřadili více než jeden deskriptor. Také některé odlišnosti u posuzovatele H3 byly tímto sloučením vyřešeny. Před sloučením byl nejprve proveden důkladný rozbor použitých deskriptorů, porovnání jejich obsahu, struktury (zejména úplnosti a koherence napříč škálami) a míry překryvu, a to jak u deskriptorů odvozených od SERRJ, tak u deskriptorů kognitivních procesů. Ukázalo se, že čím si byly deskriptory podobnější, tím méně byli posuzovatelé schopni je při individiuálním posuzování odlišit, a tím méně se shodovali na přiřazení deskriptoru k položce. Nově vytvořené širší kategorie zahrnují deskriptory, které jsou si blízké obsahem a významem. Příloha A obsahuje popisné modely po sloučení kategorií.

4.1.6 Data z analýzy struktury obsahu

Data z analýzy struktury obsahu byla zpracována několika způsoby. Ve formě grafu je možné pozorovat pohled posuzovatelů na strukturu subtestů v jednotlivých letech 2012–2015, dále to, do jaké míry jsou si testové verze podobné, pokud bychom hodnocení posuzovatelů sloučili do jednoho „superposuzovatele“, a také prezentujeme přehled toho, jak odlišné či podobné jsou testové verze z pohledu jednoho posuzovatele (obr. 4–9). V tabulkovém přehledu nahlížíme na shodu posuzovatelů číselně a uvádíme přehled četností shody posuzovatelů (tab. 1–4). Pohled na strukturu obsahu testových verzí využíváme jako jedno z hledisek při interpretaci faktorů nalezených při faktorové analýze.

4.1.7 Shoda posuzovatelů

Základním indikátorem spolehlivosti rozhodnutí a závěrů, které budou na základě výsledků realizovány, je míra shody posuzovatelů a spolehlivost získaných údajů (*inter-rater agreement*[31] a *inter-rater reliability*). Shodou posuzovatelů máme v našem případě na mysli to, jak se posuzovatelé shodnou mezi sebou na hodnocení (zařazení do kategorie, přisouzení hodnoty apod.), reliabilitou označujeme míru jejich spolehlivosti, a tedy jejich teoretickou zaměnitelnost. Reliabilita je obvykle matematicky definována jako podíl systematické variance a celkové variance posuzování. Reliabilita ve výzkumné praxi znamená, že každý výzkumník má dostatek informací k tomu, aby doložil, že jeho data byla nasbírána takovým způsobem, který eliminuje nebo minimalizuje jejich kontaminaci chybou, a také že data mají stejný význam pro kohokoli, kdo s nimi pracuje (Krippendorff, 2004, s. 211).

Uvědomujeme si, že reliabilita, chápeme-li ji ve shodě s Krippendorffem (2004) jako stabilitu a reprodukovatelnost výsledků a přesnost dat, je především výsledkem interakce mezi posuzovatelem (a jeho charakteristikami), posuzovaným objektem (a jeho charakteristikami), použitým nástrojem (a jeho interpretací posuzovateli) a metodou (a její aplikací), a že není možné všechny tyto proměnné kontrolovat. V každém expertním posuzování je přítomen lidský faktor – subjektivita, jež ovlivňuje jak vstupní data, tak jejich interpretaci. S tímto vědomím interpretujeme vypočtené hodnoty reliability a shody.

Povaha proměnných

V obou popisných modelech tohoto výzkumného projektu jde o posuzování kategorických (nominálních) proměnných s latentními charakteristikami, které je nutno interpretovat a vyvozovat. McCray (2013) pro tento typ proměnných navrhuje termín judgemental variable, tedy proměnné, které „odrážejí sice informovaný, ale přesto subjektivní názor na konkrétní zkoumanou vlastnost“ (překlad autorky[32]). Definice ani interpretace takových proměnných není přímočará a jednoznačná a posuzovatelé je mohou interpretovat odlišně i přes poskytnutý trénink. Takovéto proměnné se liší od jiných (např. od kategorií původ, pohlaví, přítomnost či nepřítomnost řečové vady) tím, že jejich vymezení, definování

31 též např. inter-coder agreement, inter-observer agreement

32 Variables which reflect the subjective, yet informed opinion of a judge about a specific matter under investigation.

a interpretace nemusí být jednoznačné, a jsou ovlivněny subjektivní internalizací významu poskytnutého výzkumníkem. V tomto výzkumu nebylo možné kontrolovat veškeré vlivy, proto bylo třeba se spolehnout na to, že i při individuálním posuzování hodnotili posuzovatelé co nejvíce v souladu s instrukcemi a s nácvikem, drželi se společně prodiskutované interpretace popisných kategorií, a následně jsme zohlednili povahu proměnných při interpretaci výsledků analýz.

Koeficienty shody a faktory, které je ovlivňují

Nejprve tedy odlišíme pojmy spolehlivost a shoda v souvislosti s provedeným expertním posuzováním. V odborné literatuře i v praxi bývají tyto dva pojmy často zaměňovány. Objevují se spojení jako *inter-rater reliability* nebo *inter-rater agreement, intra-rater agreement* nebo *intra-rater reliability* se stejnou intepretací. Pro charakteristiku vztahu shody a spolehlivosti (reliability) lze použít Krippendorffův výrok: „shoda je to, co měříme, reliabilita je to, co bychom chtěli vyvodit" (Krippendorff, 2004, s. 215).

Jak jsme již uvedli, někdy jsou tyto pojmy považovány za synonyma, jinde se autoři snaží o jejich odlišení. Kottner a Steiner (2011) v reakci na text Costa-Santosové a kol. (2011) upřesňují, že shoda souvisí s otázkou, do jaké míry jsou skóry (výsledky, hodnocení) shodné, podobné, nebo naopak odlišné. Na míru shody na zařazení do určité kategorie nemá vliv variabilita mezi posuzovanými subjekty ani distribuce měřeného rysu (s. 701). U nominálních nebo kategorických dat jde o míru shody posuzovatelů na zařazení nebo pojmenování apod. určitého posuzovaného jevu podle daného kódovacího schématu (Lavrakas, 2008, s. 344).

Koeficienty reliability obvykle udávají podíl variability skórů získaných od posuzovatelů vůči celkové variabilitě skórů v posuzovaném vzorku a říkají, jak dobře tyto skóry rozlišují mezi hodnocenými subjekty (kategoriemi). Z toho vyplývá jedna ze zásadních vlastností reliability, a sice její závislost na variabilitě, neboli „kde není variabilita, tam není reliabilita". Pokud se tedy posuzovatelé zcela shodnou, bude shoda absolutní, ale reliabilita bude 0. Dále o tomto hovoříme v souvislosti s tzv. kappa paradoxy a také v oddíle 4.3, kde interpretujeme výsledky analýzy struktury obsahu.

Vzhledem k tomu, že dokonalá reliabilita je v praxi nedosažitelná (Krippendorff, 2004, s. 221), je třeba vyhodnotit, jakou hodnotu reliability

(při použití určitého koeficientu) může v kontextu svého výzkumu akceptovat. Míra akceptovatelnosti se liší zejména podle toho, jaké důsledky může mít využití dat nebo závěrů s nižší mírou spolehlivosti neboli „potřeba přesnosti se zvyšuje s narůstajícím významem důsledků rozhodnutí a interpretací“ (AERA, APA, & NCME, 2014, s. 33, překlad autorky).

Vztah mezi shodou a reliabilitou závisí na četných faktorech. Vysoká míra shody neznamená automaticky vysokou reliabilitu, a naopak. Koeficienty reliability nabývají hodnot 1 (dokonalá reliabilita) až 0 (žádná reliabilita), teoreticky až -1 (záporné hodnoty obvykle značí problém v datech). V některých oborech, např. u klinických studií, kde důsledky posuzování mohou ovlivnit lidské životy, akceptují výzkumníci obvykle pouze data s hodnotami reliability velmi blízkými 1. Záměrem tohoto výzkumu bylo využít výstupy z analýzy struktury obsahu jako pomocnou informaci pro definování modelu pro analýzu konstruktu a případná nižší míra spolehlivosti není tedy zásadní.

Existuje mnoho různých indexů a koeficientů, které lze použít pro analýzu shody a reliability posuzovatelů, v praxi jich jícxh bývá standardně používáno jen několik. Níže uvádíme krátký výčet a charakteristiky těch nejpoužívanějších a zároveň takových, jejichž použití jsme zvažovali. Zdůvodňujeme také rozhodnutí reportovat spolu se zvoleným koeficientem také procentuální shodu[33].

Procentuální shoda *(percent agreement)* je často používaným indexem, využitelným pouze pro nominální (kategorické) proměnné. Udává podíl shodných hodnocení v celkovém počtu možných shod. Výhodou tohoto indexu je jednoduchost výpočtu a snadná a přímočará interpretace. Za jeho největší nevýhodu lze označit fakt, že nezohledňuje shodu, která může být produktem náhody – náhodnou shodu, čímž může nadhodnotit míru shody posuzovatelů. Pravděpodobnost náhodné shody se zvyšuje s klesajícím počtem kategorií, na kterých se posuzovatelé mají shodnout; naopak čím vyšší je počet kategorií, tím obtížnější je dosáhnout vysokého procentuálního podílu shody (Lavrakas, 2008). Dále platí, že procentuální podíl shody souvisí s pravděpodobnou četností kategorií. Pokud jedna z kategorií výrazně převažuje, zvyšuje se procentuální shoda a intuitivně očekáváme, že pravděpodobnost náhodné shody klesá. Ne všechny koeficienty ale pracují s tímto předpokladem a nadhodnocují tak vliv náhodné shody, čímž vykazují nepředvídatelné výsledky (*erratic results* – podle Gwet, 2008).

33 Někdy je také používán termín *přímá shoda*, *procentuální podíl shody*.

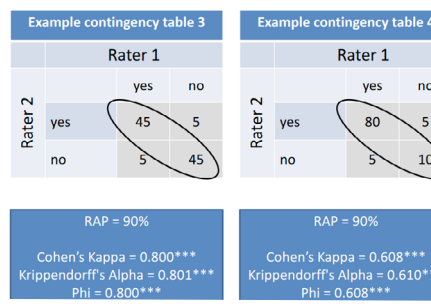
Korekci na náhodnou shodu posuzovatelů zahrnuje **Scottovo p (π)**, neboť bere v úvahu počet kategorií v datech (Scott, 1955). Je však použitelný pouze pro nominální proměnné a dva posuzovatele a předpokládá stejné rozložení kategorií u obou posuzovatelů. **Fleissovo kappa (κ)** je rozšířením Scottova p (π) pro dva a více posuzovatelů, sdílí však výše uvedená omezení.

Cohenovo kappa (κ) sice bylo adaptováno i pro více posuzovatelů posuzujících nominální proměnné a obsahuje korekci pro tzv. bias, tedy nerovnoměrné rozložení kategorií u jednotlivých posuzovatelů (Gwet, 2011), podle některých autorů však stále nejspolehlivěji funguje u dvou posuzovatelů (např. Lavrakas, 2008).

Komplexnější, avšak zároveň i na výpočty náročnější metodou se jeví **Krippendorffovo alpha (α)**, také Kalpha. Zahrnuje korekci na náhodnou shodu, je použitelné pro více posuzovatelů, všechny typy škál, chybějící data, i malé vzorky, výsledky jsou porovnatelné napříč sadami dat (Krippendorff, 2015, osobní komunikace[34]) Jeho nevýhodou je ale to, že korekce na náhodnou shodu je aplikována vždy, a to bez ohledu na to, že převaha některé kategorie může být naprosto přirozenou vlastností zkoumaných dat, nikoli něčím, co by mělo být korigováno.

Převaha některé z kategorií, nebo také převaha často volené kategorie (*high trait prevalence*) je jedním z tzv. kappa paradoxů (Cicchetti & Feinstein, 1990, McCray, 2013, Thompson & Walter, 1988). Projevuje se i v našich datech. McCray (2013) jej ilustruje kontingenčními tabulkami s výstupy binárního hodnocení dvou posuzovatelů a s hodnotami čtyř různých koeficientů shody (obr. 3).

Obrázek 3
Ukázka výstupů binárního hodnocení dvou posuzovatelů – ilustrace kappa paradoxu (převzato z McCray, 2013)



34 http://web.asc.upenn.edu/usr/krippendorff/dogs.html

Jednoznačně vidíme, že posuzovatelé (Rater 1 a Rater 2) se shodli při obou hodnoceních na 90 % jevů (*RAP=90%, Raw Agreement in Percent*), avšak hodnoty koeficientů v tabulce vlevo a v tabulce vpravo jsou různé podle toho, zda byl přibližně stejný výskyt shody v obou kategoriích (obrázek vlevo: ANO i NE 45x), nebo zda převažovala jedna z kategorií (obrázek vpravo: 80x ANO, 10x NE). Podle Gweta (2002) není důvod, proč by v uvedených příkladech měla být nízká hodnota koeficientů shody. Podle něj intuitivně a správně očekáváme v druhém případě (u situace v pravé části obr. 4) nižší hodnotu náhodné shody. Gwet (2002) tvrdí, že v případě vysoké procentuální shody nebo převahy některé z kategorií není opodstatněný předpoklad o nutnosti korigovat všechny párové shody vůči pravděpodobnosti náhodné shody; naopak to podle něj může vést v některých případech i k „nepředvídatelným výsledkům".

Kalpha a další koeficienty typu kappa bývají často kritizovány pro nezohledňování kappa paradoxů a kvůli tomu, že striktně vycházejí z teoretických modelů, nikoli z empirické podoby šetření a dat. Gwet (2011, 2014, 2016) proto ve statistickém balíčku AgreeStats navrhuje koeficienty AC1 a AC2, které berou v úvahu povahu dat a z ní vyplývající podíl dat, která mohou být ovlivněna náhodnou shodou, a redukují tak nadhodnocený vliv pravděpodobnosti náhodné shody. Klein (2018, s. 878) uvádí, že Gwetův AC1 a AC2 pracuje při statistické inferenci se dvěma zdroji rozptylu, a to nejen mezi posuzovanými subjekty, ale také mezi posuzovateli. Tyto koeficienty lze využít při výpočtu shody dvou a více posuzovatelů, dokáží pracovat s chybějícími daty, při korekci na náhodnou shodu zohledňují design posuzování (počty posuzovatelů, posuzovaných subjektů a popisných kategorií) a do výpočtů zahrnují i fenomén nerovnoměrného rozložení kategorií v datech (bias, nebo podle Gweta high-trait prevalence) a nehomogenních marginálních četností. Gwet (2008 a 2011) doporučuje vždy spolu s výsledky Kalpha či jiného koeficientu reportovat také informace o procentuálním podílu shody a případné převaze některé z kategorií. Výhodou balíčku AgreeStat je kromě jednoduchého ovládání (jde v podstatě o Excel) i to, že simultánně počítá i ostatní koeficienty, tedy právě i například procentuální podíl shody. Proto jsme se rozhodli AgreeStat využít a reportujeme shodu posuzovatelů pomocí koeficientu AC1 a procentuální shody (přímé shody).

4.1.8 Zjištění z analýzy struktury obsahu: zastoupení popisných kategorií

Jak jsme již zmínili v předchozích oddílech, distribuce popisných kategorií má na hodnotu indexu přímé shody vliv: pokud jedna z kategorií výrazně dominuje (tzv. *bias* nebo *high trait prevalence* podle Gweta, 2002), pak roste hodnota indexu přímé shody a my intuitivně očekáváme nižší pravděpodobnost náhodné shody. Ne všechny indexy nebo koeficienty pro výpočet shody tento fakt při výpočtech zohledňují. Důsledkem jsou potom nižší nebo zmatečné hodnoty koeficientů shody (Gwet, 2011). Proto prezentujeme výsledky analýzy struktury obsahu z různých úhlů pohledu a pomocí dvou různých koeficientů. Výsledky jsou nejprve zobrazeny v grafické podobě, kde jednotlivé grafy zobrazují procentuální podíl deskriptorů v každém subtestu a ve všech testových verzích (2012–2015). Zaměřujeme se na to, jak každý posuzovatel hodnotil zastoupení deskriptorů v subtestech. Zajímá nás také, jak toto zastoupení viděl tzv. superposuzovatel (prostý součet za všechny posuzovatele), a to před a po sloučení popisných kategorií. Grafy zobracují zastoupení cílů popsaných deskriptory popisných modelů a ověřovaných položkami v testových verzích. Tento podíl může však být u každého posuzovatele reprezentován odlišnými položkami. Grafické zpracování přehledu zastoupení jednotlivých popisných kategorií v subtestech testových verzí tedy odpovídá na otázku, do jaké míry se posuzovatelé shodnou na struktuře subtestů, tj. na zastoupení jednotlivých popisných kategorií, bez ohledu na to, jakými položkami je tento podíl naplněn. Zajímalo nás, zda posuzovatelé nahlížejí na subtesty jako na identické z hlediska struktury, a to z pohledu jednoho hodnotitele a jeho pohledu na všechny čtyři testové verze, tak z pohledu všech čtyř testových verzí posuzovaných všemi čtyřmi posuzovateli. Dále prezentujeme v tabulkách četnost shody mezi dvěma posuzovateli a četnost shody všech čtyř posuzovatelů.

Subtest Poslech 2012–2015

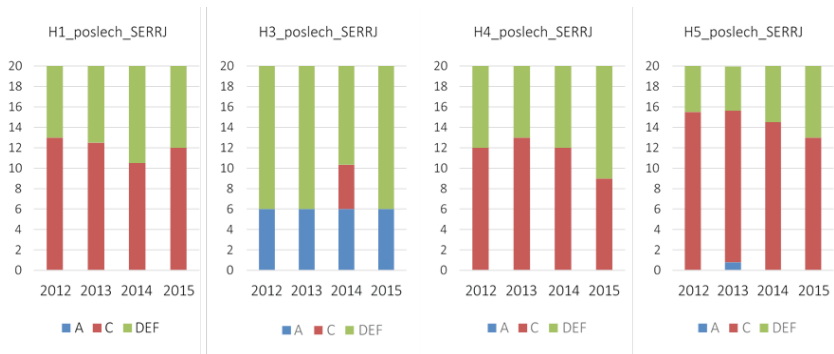
U subtestů Poslechu v modelu podle SERRJ byly pozorovány malé rozdíly týkající se především rozdílného podílu deskriptorů C a DEF u jednotlivých posuzovatelů (horní pás s grafy na obr. 4a). Pokud bychom se na posuzovatele podívali jako na jednu entitu, tj. na jakéhosi „superposuzovatele“, a jednotlivá hodnocení sumarizovali prostým součtem, pak v dolním pásu obrázku 4a vidíme téměř identickou strukturu všech čtyř verzí subtestu Poslech, i když i zde je stále patrná určitá proměnlivost podílu deskriptorů C a DEF.

Obrázek 4a

Struktura obsahu pro sloučené kategorie popisného modelu SERRJ v Poslechu

Obrázek 4b

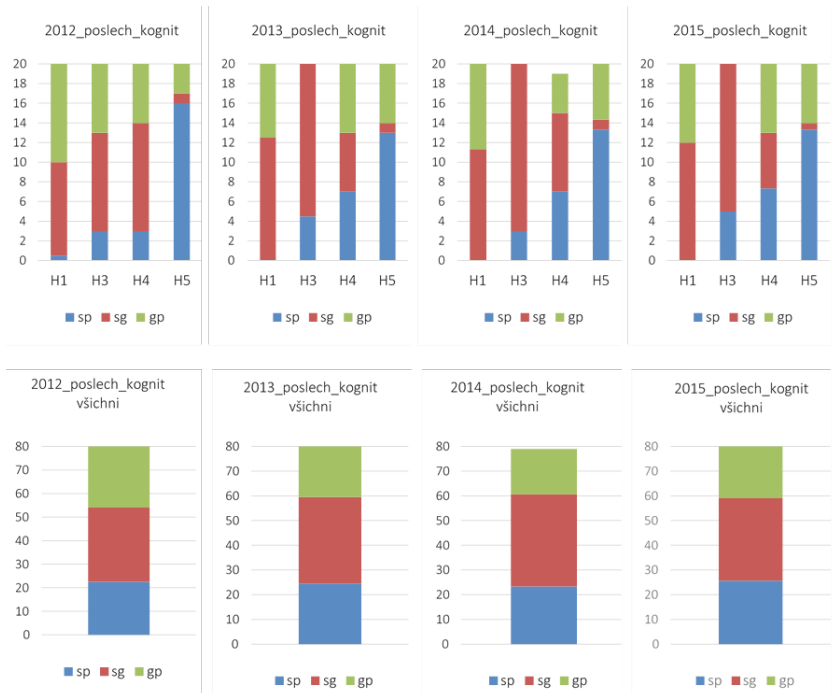
Pohled jednotlivých posuzovatelů na strukturu obsahu pro sloučené kategorie popisného modelu SERRJ v Poslechu



Grafy v obrázku 4b ukazují pohled posuzovatelů jako jednotlivců na to, jak stabilní a srovnatelný je obsah testových verzí 2012–2015. Pozorujeme, že individuální posuzovatelé vidí strukturu poslechových subtestů jako velmi podobnou v čase, i když lze zaznamenat určité rozdíly. Pokud bychom měli zjištění zobecnit, aniž bychom tvrdili, že jde o rozdíl signifikantní, pak se zdá, že u posuzovatelů (s výjimkou H3) slabě narůstá podíl kategorie DEF v čase (interpretace textu, porozumění myšlenkám v textu) oproti C (práce s informacemi v textu).

Obrázek 5a

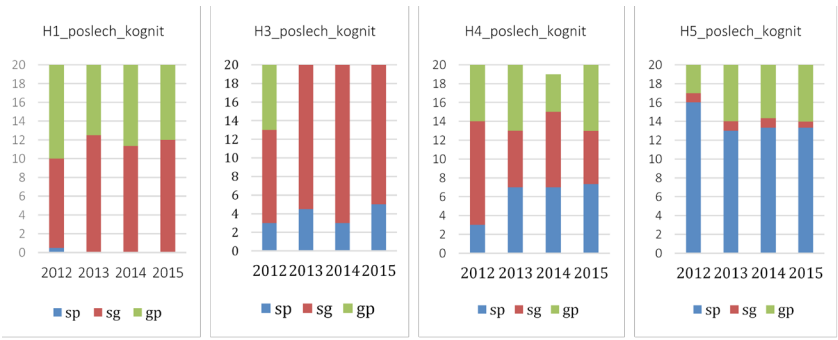
Struktura obsahu pro sloučené kategorie modelu kognitivních procesů v Poslechu



Pro poslechové subtesty posuzované v modelu kognitivních procesů pozorujeme poměrně velké odlišnosti v tom, jak posuzovatelé nahlížejí strukturu subtestů 2012–2015 (obr. 5a). Je zřejmé, že všichni čtyři posuzovatelé pozorují u subtestu 2012 odlišnou strukturu oproti třem zbývajícím verzím. Obrázek 5b ukazuje, že pro jednotlivé posuzovatele

jsou subtesty z hlediska zastoupení položek vyžadujících při řešení určité kognitivní procesy poněkud odlišné. Pokud bychom se na subtesty podívali skrze superposuzovatele (dolní část obr. 5a), pak je struktura obsahu subtestů 2012–2015 velmi podobná, ne však identická, neboť i zde se projevuje odlišnost subtestu 2012 a zároveň posuzovatelé se vzájemně liší v pohledu na strukturu subtestů, i když jako jednotlivci hodnotí subtesty jako podobné, až identické, s výjimkou roku 2012.

Obrázek 5b
Pohled jednotlivých posuzovatelů na strukturu obsahu pro model kognitivních procesů v Poslechu

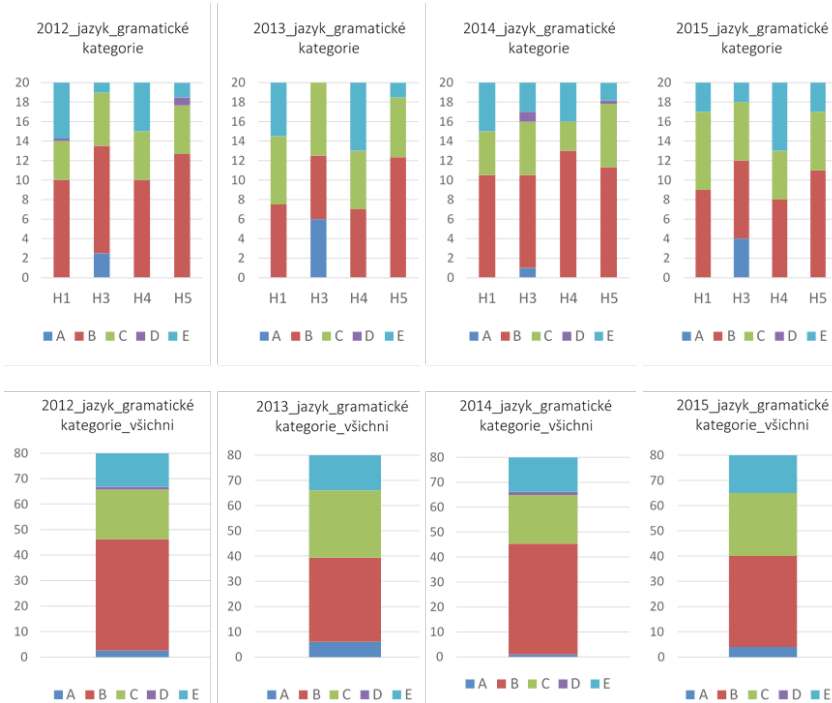


Subtest Gramatika 2012–2015

V subtestu Gramatika v modelu gramatických kategorií bylo patrně přiřazení deskriptorů k položkám pro posuzovatele obzvláště obtížné, což vyvozujeme z toho, jak často se uchylovali k přiřazení dvou deskriptorů k jedné položce. Při úvahách o sloučení do nadřazených kategorií však nebylo možné nalézt smysluplné nadřazené nebo zastřešující kategorie, byly tedy pro analýzy ponechány v původní podobě. Na obrázku 6a lze vidět, že dvojice verzí 2012 a 2014 a verzí 2013 a 2015 jsou si vzájemně podobné (horní část obr. 6a), podobnost všech čtyř verzí už však patrná tolik není. U superposuzovatele (dolní část obr. 6a) spatřujeme rozdíl zejména v odlišném zastoupení deskriptorů B (lexikum, frazeologie) a C (morfosyntax). Ani posuzovatelé jako jednotlivci (obr. 6b) nevidí subtesty 2012–2015 jako podobné, shodují se tedy na tom, že podíl gramatických kategorií ověřovaných v jednotlivých subtestech 2012–2015 je odlišný.

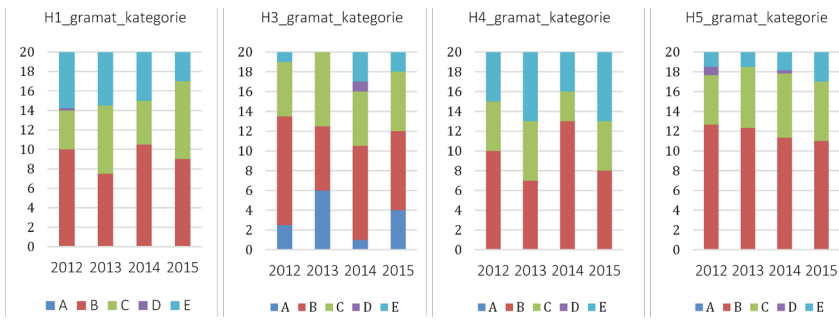
Obrázek 6a

Struktura obsahu v modelu s gramatickými kategoriemi v subtestu Gramatika



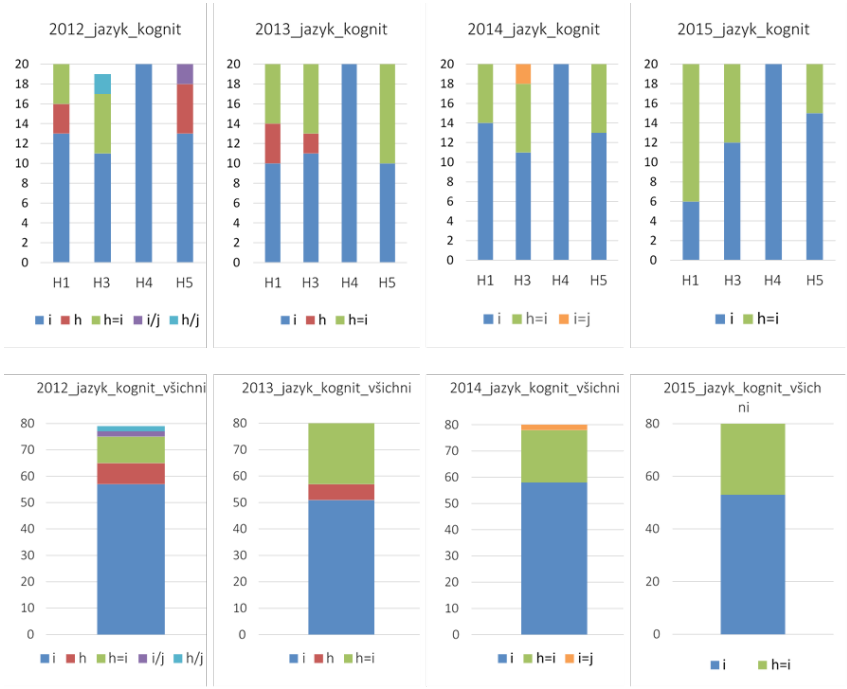
Obrázek 6b

Pohled jednotlivých posuzovatelů na strukturu obsahu modelu s gramatickými kategoriemi v subtestu Gramatika



Obrázek 7a

Struktura obsahu pro kategorie modelu kognitivních procesů v subtestu Gramatika

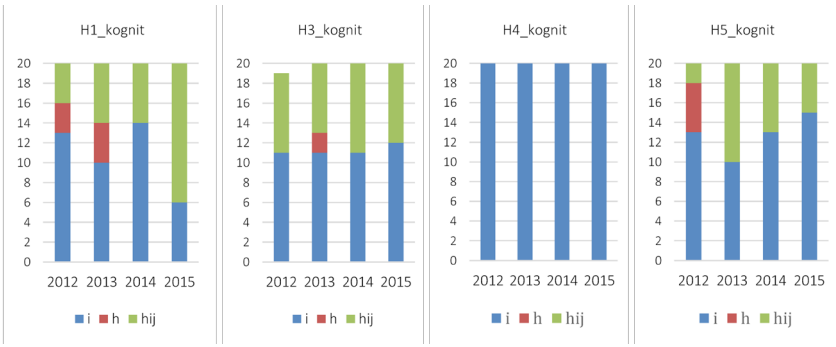


U subtestu Gramatika v modelu kognitivních procesů ukazují grafy v horní části obrázku 7b, že posuzovatel H4 hodnotí subtesty 2012–2015 jako naprosto identické a monolitické z hlediska podílu deskriptorů, avšak pro ostatní tři posuzovatele už toto neplatí. Ti hodnotí strukturu obsahu napříč roky jako odlišnou, s kolísavým podílem popisných kategorií. Ani jako skupina se posuzovatelé příliš neshodují a struktura subtestů 2012–2015 se nejeví jako shodná ani při aplikaci pohledu superposuzovatele (dolní část obr. 7a).

V deskriptivním modelu kognitivních procesů byly použity tři kategorie odvozené z Purpurova modelu jazykové kompetence (2004): H – znalost gramatické formy, I – znalost gramatického významu a J – znalost pragmatického významu. Posuzovatelé však v subtestech neshledali mnoho položek, které by měřily formu a význam odděleně, přiřazovali proto položkám kombinaci deskriptorů. Položky ověřující

pouze gramatickou formu (H) byly zastoupeny jen minimálně, a to jen u některých hodnotitelů a subtestů. Ani položky měřící pragmatický význam (J) nebyly podle posuzovatelů v subtestech zastoupeny nijak významně, a už vůbec ne jako položky ověřující pouze znalost pragmatického významu. V grafické analýze proto zobrazujeme podíly položek, které byly přiřazeny buď k deskriptoru H, nebo k deskriptoru I a podíl položek s vícečetným přiřazením (HIJ). Převažují položky ověřující znalost gramatického významu (H) a položky, u kterých nelze jednoznačně říci, o jaký typ znalosti jde (HIJ).

Obrázek 7b
Pohled jednotlivých posuzovatelů na strukturu obsahu v modelu kognitivních procesů v subtestu Gramatika

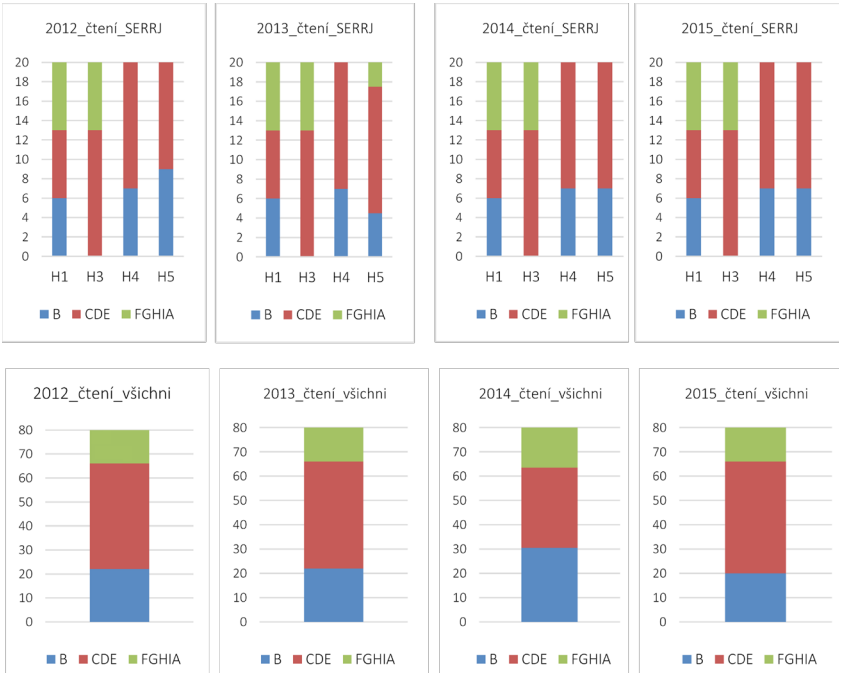


Subtest Čtení 2012–2015

Verze subtestů Čtení v modelu podle SERRJ se zdají strukturou obsahu podobné dvojicím posuzovatelů (obr. 8a). Jako velmi podobné se zdají subtesty posuzovatelům H4 a H5, posuzovatelé H1 a H3 se shodují na posílu kategorie FGHIA, která ale není v subtestech téměř zastoupená podle druhé dvojice. Pohled superposuzovatele však tyto rozdíly stírá a testové verze Čtení by při sloučeném pohledu vycházely jako srovnatelné napříč roky 2012–2015. Zajímavé je, že jednotliví posuzovatelé hodnotí testové verze jako velmi podobné napříč roky 2012–2015 (obr. 8b).

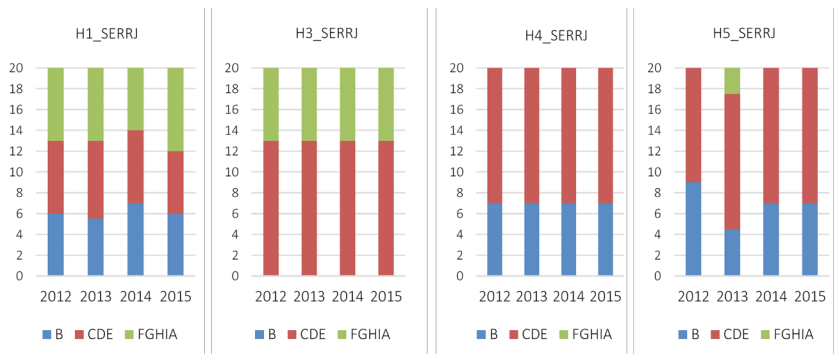
Obrázek 8a

Struktura obsahu pro popisný model SERRJ v subtestu Čtení



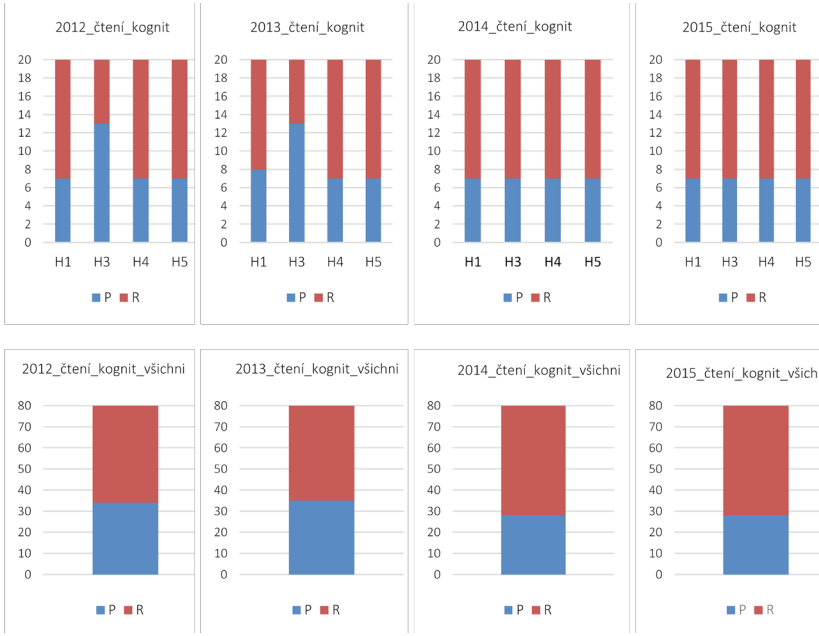
Obrázek 8b

Pohled jednotlivých posuzovatelů na strukturu obsahu pro popisný model SERRJ v subtestu Čtení



Obrázek 9a

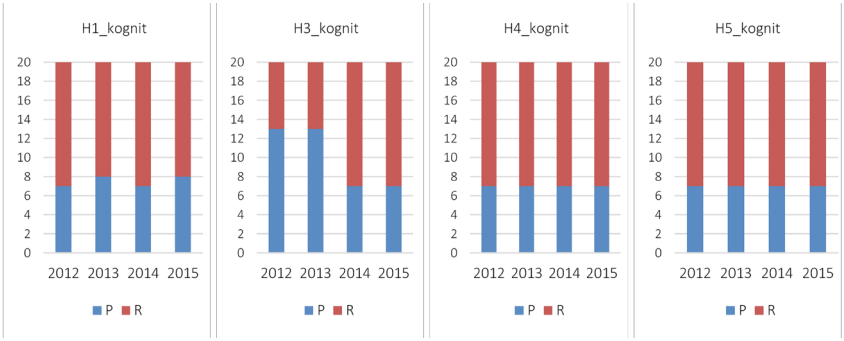
Struktura obsahu pro kategorie popisného modelu kognitivních procesů v subtestu Čtení



U verzí subtestu Čtení byla v modelu kognitivních procesů posuzovateli identifikována identická struktura obsahu (horní část obr. 9a), s malou odlišností v případě posuzovatele H3 u subtestů 2012 a 2013. Tyto rozdíly mizí při pohledu superposuzovatele (dolní část obr. 9a).

Podíváme-li se na to, jak strukturu obsahu subtestu Čtení napříč verzemi 2012–2015 vidí jednotliví posuzovatelé (obr. 9b), pozorujeme, že posuzovatelé H4 a H5 se zcela shodují jak ve dvojici, tak zároveň hodnotí všechny testové verze jako zcela identické z hlediska zastoupení deskriptorů kognitivních procesů. Jednoznačně převládá podrobné čtení (P) nad rychlým (R). Velmi podobně hodnotí subtesty i posuzovatel H1, i když on sám vidí nepatrné odlišnosti u některých verzí. H3 hodnotí jako identické verze 2012 a 2013, a také dvojici subtestů 2014 a 2015, ovšem podle něj je zastoupení deskriptorů těchto dvou dvojic subtestů inverzní.

Obrázek 9b
Pohled jednotlivých posuzovatelů na strukturu obsahu pro model kognitivních procesů v subtestu Čtení



Přehled četnosti shody posuzovatelů na popisných kategoriích

Před provedením výpočtů shody posuzovatelů v balíčku AgreeStat jsme vypracovali přehled četností shod posuzovatelů v každém ze tří subtestů a v obou modelech, celkemtedy 24 dílčích shrnutí (tab. 2–4). Shoda byla definována jako shoda posuzovatelů na přiřazení položky k určitému deskriptoru, přičemž rozlišujeme tři úrovně shody: shodu dvou posuzovatelů (max. počet případů shody je 20), shodu jednoho posuzovatele s ostatními třemi posuzovateli (max. počet případů shody je 60) a shodu všech čtyř posuzovatelů (max. počet případů shody je 120). U položek přiřazených k více deskriptorům jsme pracovali s poměrem, tedy pokud např. položka 3 byla přiřazena posuzovatelem H1 k deskriptoru A a posuzovatelem k deskriptoru A a B, pak byla shoda 0,5; pokud oba přiřadili položku k deskriptoru A, byla shoda 1. Stejně jsme postupovali na všech třech úrovních shody. Pro posuzování četnosti této absolutní shody jsme využili sloučené kategorie popisných modelů.

Přehled četností shody zobrazuje, zda se posuzovatelé shodují v tom, co je ověřováno konkrétní položkou, tedy v tom, která z popisných kategorií dané položce nejlépe odpovídá. Tento pohled na shodu posuzovatelů doplňuje pohled na strukturu obsahu o informaci o míře absolutní shody na tom, co položky v určitém subtestu ověřují. Zde tedy neporovnáváme, zda posuzovatelé nahlížejí na subtesty jako na identické z hlediska struktury, nýbrž nás zajímá, jak vysoká je shoda v rámci jednoho subtestu, který subtest a v kterém modelu byl pro posuzovatele nejjednoznačnější a kteří posuzovatelé a v jakém modelu se shodovali.

Toto doplnění pohledu na analýzu struktury obsahu umožňuje kriticky zhodnotit fungování interakce mezi posuzovateli, popisnými modely a subtesty, resp. položkami. V případě shody i neshody je totiž možné vysvětlení hledat na straně kterékoli z těchto tří těchto složek posuzování, nebo v jejich kombinaci. Neshoda může být vysvětlena nekonzistentností nebo nejasnou interpretací na straně posuzovatele, nefunkčností popisného modelu či jeho nevhodnou interpretací, nebo také problematičností položek či subtestů. Ověření konzistentnosti neboli reliability posuzovatelů by bylo možné provést například opakováním posuzování se stejnými položkami a stejnými popisnými modely a způsobem zaškolení; ověření funkčnosti popisných modelů by bylo možné provést opět opakováním procesu posuzování s jinými posuzovateli, ale stejnými modely a subtesty, případně s jinou sadou položek, atd. Možností ověření reliability je více, v tomto projektu však nebylo možné toto provést, ani nebylo primárním cílem navrhnout konkrétní nástroj nebo validovat maturitní zkoušku.

Shoda posuzovatelů u Poslechu (tab. 2) v modelu dle SERRJ ukazuje na problematickou interpretaci posuzovatele H3. Posuzovatelé H1, H4 a H5 se shodují ve vysoké míře (bez H3 jde o cca 80–100% shodu), shodu v celé skupině výrazně snižuje právě vliv H3. V modelu kognitivních procesů je u posuzovatelů shoda méně patrná, posuzovatel H3 se výrazněji neodlišuje.

V subtestu Gramatika (tab. 4) se u modelu gramatických kategorií shodují posuzovatelé z více než 70 % na přiřazení položek deskriptorů u subtestu 2014, u ostatních subtestů se shodují u 55–65 % přiřazení. Podobně je tomu u modelu kognitivních procesů, i zde se posuzovatelé shodli nejvíce u subtestů 2014 (přes 80 %) a 2012. Nutno však upozornit na to, že v jejich přiřazení výrazně převažovala jedna ze tří popisných kategorií.

Tabulka 2
Přehled četností shody posuzovatelů – Poslech

Poslech 2012 SERRJ				
Shoda ve skupině 76/120 63 %				
	H1	H3	H4	H5
H1		7	20	20
H3	7		8	4,5
H4	20	8		16,5
H5	20	4,5	16,5	
Tot/60	**47**	**19,5**	**44,5**	**41**
%	**78**	**33**	**76**	**68**

Poslech 2013 SERRJ				
Shoda ve skupině 71/120 59 %				
	H1	H3	H4	H5
H1		7	19,5	15,7
H3	7		7	5,7
H4	19,5	7		16,2
H5	15,7	5,7	16,2	
Tot/60	**42,2**	**19,7**	**42,7**	**37,5**
%	**70**	**33**	**71**	**62**

Poslech 2014 SERRJ				
Shoda ve skupině 79/120 66 %				
	H1	H3	H4	H5
H1		11	16,5	15,5
H3	11		9,5	9,5
H4	16,5	9,5		17
H5	15,5	9,5	17	
Tot/60	**43**	**30**	**43**	**42**
%	**72**	**50**	**72**	**70**

Poslech 2015 SERRJ				
Shoda ve skupině 75/120 63 %				
	H1	H3	H4	H5
H1		7	16	19
H3	7		10	7
H4	16	10		16
H5	19	7	16	
Tot/60	**42**	**24**	**42**	**42**
%	**70**	**40**	**70**	**70**

Poslech 2012 kognitivní procesy				
Shoda ve skupině 55/120 **46 %**				
	H1	H3	H4	H5
H1		14	13	3,5
H3	14		15	6
H4	13	15		5
H5	3,5	6	5	
Tot/60	**30,5**	**35**	**33**	**14,5**
%	**51**	**58**	**55**	**24**

Poslech 2013 kognitivní procesy				
Shoda ve skupině 55,5/120 **46 %**				
	H1	H3	H4	H5
H1		8,5	13	7
H3	8,5		7,5	5,5
H4	13	7,5		14
H5	7	5,5	14	
Tot/60	**28,5**	**21,5**	**34,5**	**26,5**
%	**48**	**36**	**58**	**44**

Poslech 2014 kognitivní procesy				
Shoda ve skupině 40,5/120 **34 %**				
	H1	H3	H4	H5
H1		7,5	6	6
H3	7,5		9	3
H4	6	9		9
H5	6	3	9	
Tot/60	**19.5**	**19,5**	**24**	**18**
%	**33**	**33**	**41**	**31**

Poslech 2015 kognitivní procesy				
Shoda ve skupině 53,5/120 **45 %**				
	H1	H3	H4	H5
H1		7	12	6,5
H3	7		8,5	5,5
H4	12	8,5		14
H5	6,5	5,5	14	
Tot/60	**32,5**	**10**	**30,5**	**21**
%	**54**	**17**	**51**	**35**

Tabulka 3
Přehled četností shody posuzovatelů – Gramatika

Gramatika 2012 SERRJ				
Shoda ve skupině 75,8/120 63 %				
	H1	H3	H4	H5
H1		12,7	13,7	13,3
H3	12,7		11,8	12,2
H4	13,7	11,8		12,2
H5	13,3	12,2	12,2	
Tot/60	**39,7**	**36,7**	**37,7**	**37,5**
%	**66**	**61**	**63**	**63**

Gramatika 2013 SERRJ				
Shoda ve skupině 66,5/120 55 %				
	H1	H3	H4	H5
H1		9,5	12	12
H3	9,5		11	10,5
H4	12	11		
H5	12	10,5	11,5	
Tot/60	**33,5**	**31**	**34,5**	**34**
%	**56**	**52**	**58**	**57**

Gramatika 2014 SERRJ				
Shoda ve skupině 87/120 73 %				
	H1	H3	H4	H5
H1		15,5	14	16
H3	15,5		14	13,5
H4	14	14		14
H5	16	13,5	14	
Tot/60	**45,5**	**43**	**42**	**43,5**
%	**76**	**72**	**70**	**73**

Gramatika 2015 SERRJ				
Shoda ve skupině 76,5/120 65 %				
	H1	H3	H4	H5
H1		11	12	14
H3	11		13	13
H4	12	13		13,5
H5	14	13	13,5	
Tot/60	**37**	**37**	**38,5**	**40,5**
%	**62**	**62**	**64**	**68**

Gramatika 2012 kategorie				
Shoda ve skupině 94/117 80 %				
	H1	H3	H4	H5
H1		15,5	15	17
H3	15,5		14	16
H4	15	14		16,5
H5	17	16	16,5	
Tot/60	**47,5**	**45,5**	**45,5**	**49,5**
%	**79**	**76**	**76**	**83**

Gramatika 2013 kategorie				
Shoda ve skupině 89/120 74 %				
	H1	H3	H4	H5
H1		14	13	15
H3	14		14,5	17,5
H4	13	14,5		15
H5	15	17,5	15	
Tot/60	**42**	**46**	**42,5**	**47,5**
%	**70**	**77**	**71**	**79**

Gramatika 2014 kategorie				
Shoda ve skupině 97,5/120 81 %				
	H1	H3	H4	H5
H1		16	17	16,5
H3	16		15,5	16
H4	17	15,5		16,5
H5	16,5	16	16,5	
Tot/60	**49,5**	**47,5**	**49**	**49**
%	**84**	**81**	**83**	**83**

Gramatika 2015 kategorie				
Shoda ve skupině 75/120 63 %				
	H1	H3	H4	H5
H1		13,5	9,5	12
H3	13,5		12	14,5
H4	9,5	12		13,5
H5	12	14,5	13,5	
Tot/60	**35**	**40**	**35**	**40**
%	**58**	**67**	**58**	**67**

Tabulka 4
Přehled četností shody posuzovatelů – Čtení

Čtení 2012 SERRJ				
Shoda ve skupině 37/120 **31 %**				
	H1	H3	H4	H5
H1		7	20	20
H3	7		8	4,5
H4	20	8		16,5
H5	20	4,5	16,5	
Tot/60				
%	**32**	**22**	**32**	**38**

Čtení 2013 SERRJ				
Shoda ve skupině 75/120 62 %				
	H1	H3	H4	H5
H1		14,5	7,5	9,3
H3	14,5		13	13
H4	7,5	13		17,5
H5	9,3	13	17,5	
Tot/60	**31,3**	**40,5**	**38**	**39,8**
%	**53**	**69**	**64**	**68**

Čtení 2014 SERRJ				
Shoda ve skupině 66/120 56 %				
	H1	H3	H4	H5
H1		14	7	7
H3	14		14	14
H4	7	14		20
H5	7	14	20	
Tot/60	**28**	**42**	**41**	**41**
%	**47**	**71**	**69**	**69**

Čtení 2015 SERRJ				
Shoda ve skupině 74,5/120 62 %				
	H1	H3	H4	H5
H1		13	9,5	12
H3	13		12	14,5
H4	9,5	12		13,5
H5	12	14,5	13,5	
Tot/60	**34,5**	**39,5**	**35**	**40**
%	**58**	**66**	**58**	**67**

Čtení 2012 kognitivní procesy				
Shoda ve skupině 102/120 85 %				
	H1	H3	H4	H5
H1		14	20	20
H3	14		14	14
H4	20	14		20
H5	20	14	20	
Tot/60	**54**	**42**	**54**	**54**
%	**90**	**70**	**90**	**90**

Čtení 2013 kognitivní procesy				
Shoda ve skupině 101/120 84 %				
	H1	H3	H4	H5
H1		15	19	19
H3	15		14	14
H4	19	14		20
H5	19	14	20	
Tot/60	**53**	**43**	**53**	**53**
%	**88**	**72**	**88**	**83**

Čtení 2014 kognitivní procesy				
Shoda ve skupině 120/120 100 %				
	H1	H3	H4	H5
H1		20	20	20
H3	20		20	20
H4	20	20		20
H5	20	20	20	
Tot/60	**60**	**60**	**60**	**60**
%	**100**			

Čtení 2015 kognitivní procesy				
Shoda ve skupině 120/120 100 %				
	H1	H3	H4	H5
H1		20	20	20
H3	20		20	20
H4	20	20		20
H5	20	20	20	
Tot/60	**60**	**60**	**60**	**60**
%	**100**			

Tabulka 5
Vypočtené koeficienty procentuální shody PA, Gwetova koeficientu AC1 a s nimi asociované směrodatné chyby

Rok	2012		2013		2014		2015	
Dovednost Model	PA	AC1	PA	AC1	PA	AC1	PA	AC1
Poslech SERRJ	0,67	0,54	0,53	0,42	0,75	0,69	0,63	0,49
SE	0,06	0,08	0,03	0,03	0,06	0,07	0,06	0,10
Gramatika Kategorie	0,69	0,62	0,57	0,44	0,78	0,74	0,65	0,56
SE	0,07	0,10	0,09	0,12	0,08	0,10	0,08	0,11
Čtení SERRJ	0,34	0,13	0,62	0,46	0,62	0,46	0,59	0,42
SE	0,03	0,04	0,08	0,12	0,08	0,12	0,06	0,11
Poslech Kognit. proc.	0,52	0,28	0,45	0,18	0,40	0,12	0,48	0,22
SE	0,04	0,06	0,02	0,02	0,02	0,03	0,03	0,06
Gramatika Kognit. proc.	0,92	0,91	0,85	0,83	0,65	0,46	0,63	0,32
SE	0,05	0,06	0,05	0,07	0,07	0,15	0,06	0,15
Čtení Kognit. proc.	0,85	0,71	0,83	0,70	1,00	1,00	1,00	1,00
SE	0,05	0,10	0,06	0,12	0,00	0,00	0,00	0,00

U posuzování subtestu Čtení (tab. 5) je z pohledu SERRJ výrazná odlišnost pozorována u subtestu 2012. V tomto subtestu je shoda posuzovatelů výrazně nízká, vůbec nejnižší ze všech posuzovaných subtestů a modelů. Shoda u subtestů 2013–2015 se již výrazněji nevymyká, je kolem 60 %. Naopak v modelu kognitivních procesů Čtení je shoda posuzovatelů velmi vysoká, až mezi 84 % a 100 %. Je však třeba připomenout, že zde šlo o pouhé dvě kategorie.

Shoda posuzovatelů: podíl přímé shody a Gwetův koeficient AC1

Pro kvantitativní posouzení shody posuzovatelů jsme zvolili koeficient AC1, který se dokáže vyrovnat s kappa paradoxy, tj. vlivem převahy některé z kategorií (*high trait prevalence*) a nerovnoměrným rozložením celkové

pravděpodobnosti volby kategorií (*marginal probability*) a s hyperkorekcí náhodné shody. Spolu s ním reportujeme také procentuální podíl přímé shody, který doplňuje náhled na strukturu dat a napomáhá jejich interpretaci. V tabulce 5 uvádíme vypočtené hodnoty obou koeficientů pro všechny subtesty a oba popisné modely a s nimi související směrodatné chyby udávající míru spolehlivosti těchto koeficientů. Ve výpočtech jsme pracovali se sloučenými kategoriemi popisných modelů. Veškeré výstupy, včetně vypočtené směrodatné chyby a intervalu spolehlivosti byly provedeny v balíčku AgreeStat. Hodnoty koeficientů AC1 udávají, do jaké míry se posuzovatelé shodnou jako skupina na přiřazení popisné kategorie k položkám v rámci jednotlivých roku 2012–2015, nevypovídají však nic o tom, zda jsou testové verze srovnatelné napříč těmito roky.

Maximální možná hodnota koeficientů je 1. Nejnižší vypočtené hodnoty dosahují hodnoty koeficientu u modelu kognitivních procesů Poslechu a také u modelu Čtení dle SERRJ. Vypočtené hodnoty koeficientů shody však samy o sobě nic neříkají, je třeba je interpretovat, určit, zda je jejich hodnota akceptovatelná. Akceptovatelnost je relativní a vždy souvisí s účelem a závažností celého šetření a také s povahou dat. Hodnotu koeficientů ovlivňuje počet posuzovatelů, počet posuzovaných subjektů/objektů, počet popisných kategorií a distribuce těchto kategorií. Na jejich interpretaci proto neexistuje jednoduchý návod. Podle Gweta[35] je důležitým aspektem velikost směrodatné chyby (SE) koeficientu shody, tedy vzdálenosti hodnoty vypočteného koeficientu od průměru skutečné hodnoty (true value), neboť udává míru nejistoty spojenou s vypočteným koeficientem shody. Měla by tvořit maximálně 15 % z celkové hodnoty koeficientu, jinak jsou vypočtené hodnoty zatížené velkou chybou, a tudíž nepřesné.

Některé hodnoty koeficientů AC1 v tabulce 5 jsou poměrně nízké, případně je jejich směrodatná chyba nezanedbatelná (zvýrazněné buňky). Z tabulky 5 je dále patrné, že přibližně čtvrtina všech koeficientů je zatížena směrodatnou chybou větší než zmíněných 15 %, z toho většina se pojí ke koeficientům AC1. Se směrodatnou chybou je úzce propojen i interval spolehlivosti (CI – *confidence interval*) udávající rozmezí, ve kterém se s určitou pravděpodobností (zde byla zvolena 95% hladina), nalézá pravá hodnota koeficientu. Směrodatná chyba hraje klíčovou roli v interpretaci koeficientů, kterou navrhuje Gwet (2014).

35 K. Gwet's Inter-Rater Reliability Blog - http://inter-rater-reliability.blogspot.com/

Pro interpretaci koeficientů, tedy výrok, zda je zjištěná hodnota koeficientu shody dostatečně vysoká pro účel, kvůli kterému byla analýza prováděna, potřebujeme kromě obsahových kritérií také statistické srovnávací ukazatele (*benchmarks*). Pravděpodobně mezi nejvíce citované a využívané patří ukazatele Landise a Kocha (1977) a Fleisse (1981), jejichž aplikace je poměrně přímočará. V podstatě jde o stanovení určitého rozsahu hodnot vypočtených koeficientů, pojmenování těchto intervalů a interpretaci síly této shody porovnáním hodnoty koeficientu s intervalem. Například hodnoty v intervalu 0,4–0,6 bývají označovány jako mírná shoda, hodnoty mezi 0,6–0,8 jako dobrá shoda atd. (Landis & Koch, 1977). Záleží na záměru výzkumu a využití výsledků, pro jaký srovnávací ukazatel se výzkumník při interpretaci míry shody rozhodne a s jakou úrovní míry shody se spokojí. V našem případě, kdy nejde o shodu samotnou, nýbrž o jeden z pohledů na strukturu obsahu a ověření funkčnosti popisných modelů, a kdy máme ještě další pohled na strukturu subtestů a shodu posuzovatelů, bychom vlastně označili většinu získaných koeficientů jako dostatečně informativních, a ani bychom je nepotřebovali interpretovat pomocí srovnávacích ukazatelů. Přesto jsme se rozhodli tak učinit z výzkumných důvodů.

Pokud bychom tedy aplikovali ukazatele Landise a Kocha (1977) nebo Fleisse (1981), hodnota většiny vypočtených koeficientů shody by spadala do kategorie mírné až podstatné shody, u některých koeficientů by byla shoda interpretována jako vynikající, což by pro naše účely zcela postačovalo. Gwet (2014) však označuje tyto srovnávací ukazatele za nedostatečně reflektující design výzkumu a jeho podmínky (počet posuzovatelů, subjektů a posuzovaných kategorií). Podle něj mají tyto podmínky vliv na velikost vypočteného koeficientu shody (s. 165) a chyby s ním asociované, což může vést k neúplné interpretaci tohoto koeficientu a jeho statistické významnosti, nebo dokonce k validaci hodnot s velkým rozpětím chyby (*large error margin*). Uvádí příklad, kdy koeficientu s hodnotou 0,5 u studie s větším množstvím posuzovatelů, subjektů a kategorií je přiřazena interpretace *mírná shoda*, zatímco stejná hodnota koeficientu u studie s menším počtem posuzovatelů, subjektů a kategorií ani nedosahuje statistické významnosti, což znamená, že jeho „pravá“ hodnota může být i blízká 0 (Gwet, 2014, s. 165). Podle Gweta je pak jediným indikátorem „pravé“ hodnoty koeficientu velikost chyby s ním spojené a u malých vzorků a statisticky nesignifikantních hodnot koeficientů nelze tyto hodnoty označit jinak než jako slabou míru shody.

Závažné je to zejména v případech, kdy se na základě interpretace shody rozhoduje o důležitých procesech, např. v lékařském prostředí. Gwet navrhuje přistupovat k interpretaci probabilisticky a popisuje interpretační škálu s tzv. ukazateli kumulativních pravděpodobností příslušnosti (*cumulative membership probabilities*). Vychází z hladin Fleissových ukazatelů a bere v úvahu charakteristiky výzkumného designu (počet posuzovatelů, posuzovaných jevů/subjektů a kategorií). K interpretaci výsledných hodnot ukazatelů Gwet doplňuje, že výzkumník sám musí zvolit z navrženého interpretačního schématu takovou hladinu ukazatele, která nejlépe odpovídá požadované míře přesnosti výzkumu. Podobně smýšlí Krippendorff (2004, s. 221), když říká, že dokonalá reliabilita je v praxi nedosažitelná a výzkumník musí vyhodnotit, jakou hodnotu reliability (při použití určitého koeficientu) může v kontextu svého výzkumu akceptovat. Míra akceptovatelnosti se liší zejména podle toho, jaké důsledky může mít využití dat nebo závěrů s nižší mírou spolehlivosti neboli „potřeba přesnosti se zvyšuje s narůstajícím významem důsledků rozhodnutí a interpretací" (AERA, APA, & NCME, 2014, s. 33, překlad autorky).

Pokud bychom chtěli aplikovat Gwetův pohled na interpretaci koeficientů shody posuzovatelů, využili bychom jím navržené ukazatele kumulativní pravděpodobnosti příslušnosti. V tabulce 6 představujeme ukázku hodnot koeficientů shody pro Poslech 2012 dle modelu SERRJ, příslušné směrodatné chyby a intervaly spolehlivosti, v tabulce 7 pak Gwetovy ukazatele kumulativní pravděpodobnosti příslušnosti asociované s vypočtenými koeficienty. V ukázce v tabulce 6 byl například pro Poslech 2012 podle SERRJ modelu vypočten koeficient shody AC1 = 0,535, s chybou 0,081, což při 95% hladině významnosti znamená, že pravá hodnota koeficientu je mezi 0,366 a 0,704. V tabulce 4 pak vidíme, že při interpretaci pomocí kumulativní pravděpodobnosti příslušnosti a hranici spolehlivosti nastavené také na 95 % musíme označit míru shody za slabou – 95% hladina pravděpodobnosti spadá až do úrovně ukazatele <0,4, tedy slabé úrovně shody. Dále lze konstatovat, že máme cca 20% pravděpodobnost, že je míra shody vynikající a 56% pravděpodobnost průměrné a dobré shody.

Tabulka 6
Příklad interpretace koeficientů pomocí kumulativní pravděpodobnosti příslušnosti (Poslech 2012 v modelu podle SERRJ)

		SE	95 % CI	p-value
Gwetův AC1	0,535	0,081	0,366 to 0,704	0,000
Procentuální shoda	0,667	0,057	0,548 to 0,785	0,000
Kumulativní pravděpodobnosti příslušnosti	Interpretace míry shody		Gwetův AC1	Procentuální shoda
>0,75	Vynikající		0,196	0,297
0,4–0,75	Průměrná až dobrá		0,564	0,784
<0,4	Slabá		1,000	1,000

Vzhledem k tomu, že šlo o testové verze použité pro analýzu struktury obsahu tak, jak byly realizovány v ostrém testování, tedy nijak jsme do nich nezasahovali (nevyřazovali jsme např. položky s ne zcela dobrými psychometrickými parametry), neměli jsme žádná očekávání, co se týče míry shody. Primárním cílem bylo zjistit, zda je tento způsob analýzy obsahu funkční, tedy skýtá možnost zjistit, jak obsahovou strukturu vidí jednotliví posuzovatelé, a jak se jeví skupině posuzovatelů, zda jsou popisné modely vhodným nástrojem, jaká mohou být úskalí metody analýzy struktury obsahu, pojmenovat je a navrhnout opatření pro jejich řešení. Pro odpovědi na tyto otázky není potřeba vysoká míra shody posuzovatelů. V tomto smyslu jsme přistoupili i k interpretaci výsledků.

Tabulka 7
Interpretace koeficientů shody pomocí kumulativní pravděpodobnosti příslušnosti

Kumulativní pravděpodobnost, že se pravá hodnota koeficientu nachází alespoň mezi 0,4 a 0,75								
	SERRJ				Kognitivní procesy			
	2012	2013	2014	2015	2012	2013	2014	2015
Poslech	0,56	0,48	0,84	0,58	0,36	0,14	0,03	0,18
Gramatika	0,99	0,89	1,00	0,99	1,00	0,99	0,72	0,39
Čtení	0,10	0,60	0,60	0,53	0,82	0,90	1,00	1,00

V přehledu v tabulce 7 uvádíme hodnoty udávající kumulativní pravděpodobnost příslušnosti koeficientu shody AC1 do kategorie střední až dobré shody. Většina koeficientů vykazuje více než 50% pravděpodobnost příslušnosti do kategorie průměrné až dobré shody, což nám pro náš výzkumný projekt s ohledem na jeho cíle připadá jako dostatečné. V reálných podmínkách využití pro maturitní zkoušku předpokládáme jiný design provedení obsahové analýzy (jeho podobu částečně navrhujeme v části III) a lze tedy očekávat dosažení hodnot vyšších.

4.2 Konstruktová ekvivalence: exploratorní faktorová analýza

Jak jsme již předeslali, cílem tohoto projektu nebyla validace slovenské maturitní zkoušky z anglického jazyka na úrovni B1, ani podání důkazu o srovnatelnosti či naopak nesrovnatelnosti testových verzí 2012–2015. Cílem projektu bylo zjistit, jaké z existujících metod by byly funkční, vhodné a relativně snadno implementovatelné do procesu vývoje a sestavování verzí testu receptivních dovedností v jakékoli podobné zkoušce tak, aby díky jejich využití mohl poskytovatel sám srovnatelnosti dosahovat a prokazovat ji.

Zmínili jsme již, že je velmi nesnadné, ne-li prakticky nemožné, sestavit po všech stránkách srovnatelné testové verze (obsahově, konstruktově i psychometricky). I proto je nutné zdůraznit, že metoda analýzy struktury obsahu jako metoda založená na kvalitativním posouzení obsahové struktury testových verzí vůči předem danému rámci (zde popisnému modelu) poskytla jedinečný a důležitý pohled na strukturu testu. Nepředpokládáme však, že by se využívala jako jediná. Tento obsahový pohled by měl být doplněn i pohledem kvantitativním. Právě takový pohled nabízí například faktorová analýza při zkoumání konstruktové ekvivalence testových verzí.

Výhodou maturitní zkoušky na Slovensku je velikost populace konající tuto zkoušku, a tedy i velké množství dat, která lze analyzovat. Proto jsme i my v dalším kroku využili tato data – zcela anonymizované žákovské odpovědi na položky – k provedení faktorové analýzy struktury konstruktu ověřovaného v testových verzích 2012–2015. Vycházíme z předpokladu, že tyto verze založené na stejných specifikacích (jakkoli jsou veřejně dostupné informace o nich a dalších aspektech související s konstruktem zkoušky a s celým jejím vývojem a validací nedostatečně podrobné) mají za cíl měřit stejný konstrukt stejným způsobem.

Konstruktovou (konfigurální) ekvivalenci testových verzí můžeme zkoumat také prostřednictvím struktury faktorů, jež nabídne faktorová analýza. Faktory vzniknou na základě vztahů manifestních proměnných – odpovědí na položky. Faktory (latentní rysy, složky konstruktu, subkonstrukty) lze vysvětlit buď pomocí hypotézy, tj. vstupního modelu specifikovaného pro CFA, nebo zpětnou interpretací, pokud použijeme EFA. V podstatě lze také EFA a CFA zkombinovat tak, že se nejprve provede EFA a nalezená struktura faktorů, resp. její interpretace, se následně využije jako specifikační model, jenž by měl být potvrzen pomocí CFA.

Původním záměrem této fáze projektu bylo provést porovnání konstruktů ve verzích 2012–2015 pomocí konfirmatorní faktorové analýzy (CFA) a využít pro specifikaci vstupního modelu CFA strukturu obsahu, kterou identifikovali posuzovatelé. S ohledem na několik okolností jsme se nakonec rozhodli provést exploratorní faktorovou analýzu. Mezi tyto okolnosti patří například otázka dostupnosti vhodného softwaru, binární povaha dat vyžadující využití tetrachorické korelace[36] a nižší míra spolehlivosti, již jsme zjistili u některých výstupů a popisných modelů (viz kap. 4). Proto jsme nakonec využili EFA. Smyslem EFA je nalézt v datovém souboru vzorec ve vztazích mezi manifestními proměnnými (položkami), jenž se dá vysvětlit společným faktorem, a na základě sdíleného rozptylu redukovat počet proměnných na nižší počet tzv. latentních proměnných – společných faktorů. EFA nevyžaduje předem specifikovat model, neprovádí jeho konfirmaci, nýbrž hledá nejmenší přijatelný počet faktorů, které vysvětlují vztahy mezi položkami. Strukturu obsahu identifikovanou posuzovateli v analýze struktury obsahu jsme přesto využili, a sice jako sekundární označení proměnných – položek (např. R – rychlé čtení versus P – podrobné čtení), což nám napomáhalo při interpretaci výstupů EFA a při rozhodování, které z nabízených řešení, tj. která z nabízených faktorových struktur koresponduje nejlépe s vysvětlením a interpretací nalezených faktorů.

Využili jsme volně dostupný software implementovaný v balíčku OPLM (Verhelst & Glas, 1995) založený na analýze hlavních faktorů – *principal factor analysis* (Harman, 1976), který akceptuje i korelační matice, jež nejsou pozitivně semidefinitní[37], což bývá u tetrachorických korelací relativně časté (Verhelst, 2019, osobní komunikace).

36 Volně dostupný software Jamovi, který umí provádět i CFA, například nezahrnuje možnost pracovat se vstupní tetrachorickou korelací, která je nutná pro analýzu binárních, nespojitých dat (odpovědi kódované jako 0 nebo 1).

37 Pozitivně semidefinitní matice má vlastní čísla větší nebo rovna 0.

4.2.1 Výsledky exploratorní faktorové analýzy

Exploratorní faktorovou analýzu jsme provedli pro všechny tři subtesty, všechny čtyři roky a oba popisné modely, tedy celkem pro 24 datových souborů. V každém souboru bylo 20 položek, na které odpovídalo více než 20 000 respondentů. Jedná se tedy o velmi velké soubory dat.

Nejprve jsme připravili korelační matice pro položky ve všech datových souborech (celkem 24 matic). Zkoumali jsme, jak spolu položky v subtestech korelují a do jaké míry jsou tedy vhodné pro faktorovou analýzu. Podle Watsona (2017) by se měly ideálně hodnoty korelačních koeficientů pohybovat mezi 0,20 a 0,80. Hodnoty pod 0,20 mohou znamenat, že položky měří něco jiného než zbytek souboru. Hodnoty nad 0,80 mohou indikovat multikolinearitu, vzájemnou závislost položek a mohou způsobovat nestabilitu odhadů a komplikovat interpretaci faktorové struktury. Obvykle se doporučuje takové položky z faktorové analýzy vyřadit. V našem případě by to však znamenalo vyřadit téměř polovinu položek ze subtestu Poslech 2012, v dalších subtestech by se jednalo o vyřazení tří až pěti položek (z dvaceti). Podrobné prozkoumání korelačních matic však bylo velmi užitečným krokem, který nám pomohl pochopit možné příčiny ne zcela jednoznačných výsledků EFA, problematické interpretace faktorů a propojit vše do závěrů o povaze obsahu a konstruktu testových verzí 2012–2015.

I přes výše zmíněná doporučení a zjištěné problémy v datech jsme z výzkumných důvodů EFA provedli. Prezentujeme výsledky, jsme si však plně vědomi toho, že byly výrazně ovlivněny povahou vstupních dat a že pokud by byla EFA prováděna za účelem prokázání ekvivalence konstruktu u ostrých verzí maturitní zkoušky, jejichž skóry by pak měly být vyrovnávány, bylo by nutné mít data, která naplňují kritéria nutná pro provedení EFA (tzn. např. přistoupit k vyřazení některých položek[38]).

Korelační matice jsme využili jako vstupní data pro EFA. EFA jsme prováděli v balíčku OPLM, dříve volně dostupném na webové stránce Cito, nyní dostupném na písemné vyžádání pověřeného pracovníka Cito[39]. Faktory jsme extrahovali pomocí metody hlavních faktorů (PFA – *principal factor analysis*), zabudované v OPLM, a nerotovali jsme je. Zjistili jsme následující:

38 Tento postup je ovšem hůř obhajitelný v případě high-stakes testu, jakým je MZ. Proto by k takovéto situaci nemělo docházet a vyrovnávání skórů by mělo probíhat na takových testových souborech, jež nevykazují problémy tohoto typu.

39 Výpočty jsme provedli v rámci stáže u jednoho z autorů OPLM, Normana Verhelsta, v únoru 2019.

U Poslechu se jako dominantní jeví jeden faktor (výrazně vyšší vlastní číslo, cca 6–7) a dále se objevují faktory, které lze považovat za méně zřetelné; zmiňujeme je proto, že i jejich vlastní čísla jsou nad hodnotou 1. Faktorové zátěže s nimi spojené jsou však zanedbatelné. Dominantní faktor 1 lze vysvětlit všemi položkami z části 3 subtestu, jejich faktorové zátěže dosahují velmi vysokých hodnot, které se podle posuzovatelů pojí s porozuměním hlavním myšlenkám a závěrům. Pojí se k němu ale i některé položky z částí 1 a 2, kde posuzovatelé identifikovali spíše kategorii práce s informacemi a porozumění hlavním bodům textu. Hodnoty faktorových zátěží ani sdíleného rozptylu však nejsou tak vysoké, jako u položek z části 3. Nejobtížněji interpretovatelný je subtest 2012, u ostatních subtestů bychom vyvodili závěr, že se jako akceptovatelná jeví struktura jednofaktorové, kde převažuje faktor interpretace textu, porozumění myšlenkám v textu. U subtestů se nízké hodnoty faktorových zátěží projevují převážně u položek, u nichž byla identifikována nízká míra citlivosti (diskriminace ULI) a/nebo bodově biseriální korelace (pbi).

U Gramatiky se jako výrazně dominantní jeví jeden faktor (výrazně vyšší vlastní číslo, cca 6–8). Dále se objevují jeden až dva faktory s vlastními čísly nad 1, které lze považovat za méně zřetelné. Dominantní faktor 1 lze vysvětlit položkami napříč oběma částmi subtestu, výrazně vyšší hodnoty faktorových zátěží se koncentrují zejména do části 2, kde posuzovatelé z velké části označili jako dominantní kategorie morfosyntax a syntax (včetně prostředků textové návaznosti). U subtestů Gramatika se nízké hodnoty faktorových zátěží projevují převážně u položek, u nichž byla identifikována nízká míra citlivosti (diskriminace ULI) a/nebo bodově biseriální korelace (pbi).

U Čtení je ze všech tří dovedností jednofaktorové struktura nejvýraznější. Zde může mít vliv i to, že subtesty Čtení obsahovaly nejméně problematických položek. Dominantní faktor 1 má vlastní číslo 7–8, dále se objevují dva faktory s vlastním číslem nad 1. Nicméně všechny položky ve všech verzích se váží výrazně k faktoru 1. Zde se zdá, že faktor 1 nelze interpretovat pomocí původních popisných kategorií (práce s informacemi versus práce s interpretací textu, rychlé nebo podrobné čtení). Můžeme zde opět vnést do hry pochybnost o tom, zda je konstrukt dovedností skutečně dělitelný (smysluplně) na dovednosti.

4.3 Srovnatelnost deskriptivních statistik testových verzí

Veškeré testové materiály použité při slovenské MZ, tedy včetně testů z cizích jazyků, musí být zveřejněny ihned po testování. Není proto možné testové verze, které byly použity v ostrém testování na celé populaci, použít například v pretestacích verzí dalších, a tím propojit informace o úlohách i testovaných z více zkušebních termínů, a získat tak možnost přímo porovnávat obtížnost úloh a schopnosti testovaných. Není nám známo z veřejně dostupných informací, jak probíhá vývoj testu včetně pretestací, zda jsou využívány analýzy IRT, případně nekompletní design, jež by umožňovaly porovnání populací nebo testových verzí či úloh. Pokud se nic takového neděje a také vzhledem k tomu, že se všechny testy ihned po realizaci zveřejňují, pak pravděpodobně není možné provést přímé, neexperimentální porovnání úrovně měřených schopností u populací v různých letech a psychometricky podložené sestavování ekvivalentních testových verzí ve smyslu ekvivalence skórů, tj. výsledků studentů konajících zkoušku v různých letech. Z toho vyplývá, že i interpretace rozdílů mezi psychometrickými charakteristikami testových verzí, jako např. distribuce skórů nebo obtížnosti, je limitována. Nevíme, zda a do jaké míry lze populace v jednotlivých ročnících považovat za srovnatelné (ekvivalentní), zejména z pohledu distribuce měřené vlastnosti (úrovně jazykové způsobilosti). Dále nevíme, zda jsou testové verze realizované v různých zkušebních termínech opravdu stejně obtížné a zda položky mají stejnou diskriminační schopností. Nelze tedy zjistit, zda např. mezi úspěšnostmi populací v roce 2012 a 2013 existuje statisticky významný rozdíl[40], a čím je způsoben; zda odlišnou obtížností položek, nebo rozdílnou úrovní schopností žáků, nebo kombinací obou těchto faktorů. V následujícím porovnání proto neprovádíme analýzu příčin nebo statistické významnosti případných odlišností ani jejich interpretaci, pouze je konstatujeme.

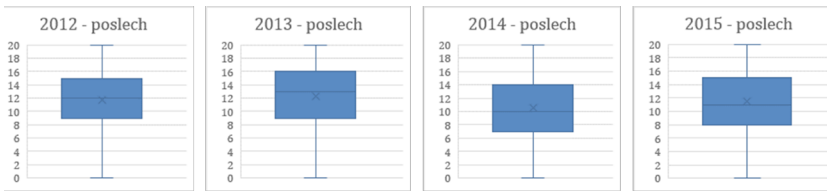
4.3.1 Porovnání populací z let 2012–2015

Při pohledu na populace 2012–2015 jsme se tedy mohli opírat pouze o vnější charakteristiky populací (zastoupení podle typu škol, pohlaví) a deskriptivní statistiky pro subtesty a podskupiny. Přehled jsme doplnili

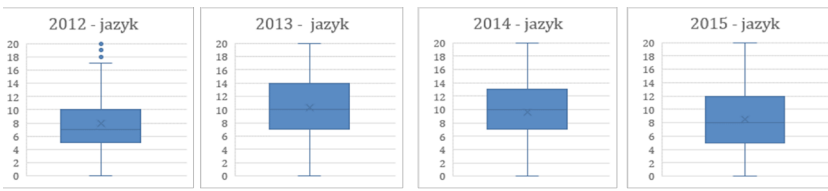
40 Je možné, že takovéto mechanismy jsou ve slovenské maturitní zkoušce implementovány, ale z veřejně dostupných informací to nevyplývá.

o krabicové grafy pro porovnání rozptylu a distribuce skórů pro všechny subtesty (grafy na obr. 10–13). Z důvodu proveditelnosti analýzy struktury obsahu a EFA a při interpretaci výsledků jsme uvažovali o populacích jako o ekvivalentních.

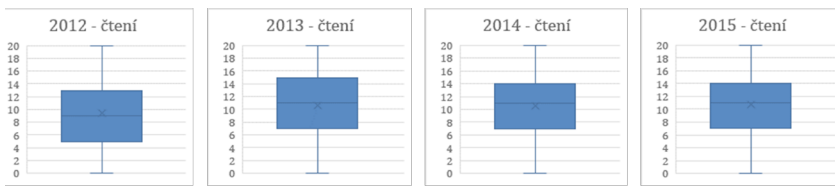
Obrázek 10
Krabicové grafy distribuce skórů pro subtest Poslech 2012–2015



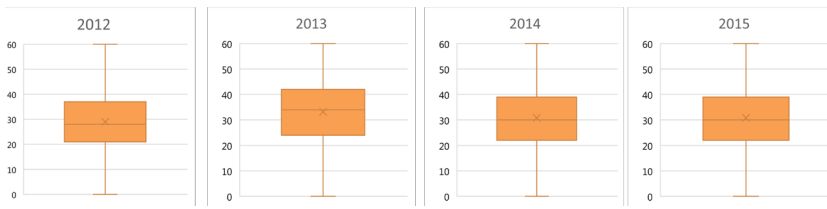
Obrázek 11
Krabicové grafy distribuce skórů pro subtest Gramatika 2012–2015



Obrázek 12
Krabicové grafy distribuce skórů pro subtest Čtení 2012–2015



Obrázek 13
Krabicové grafy distribuce skórů pro kompletní testové verze 2012–2015



Procentuální zastoupení žáků podle zřizovatele školy a pohlaví nevykazuje v meziročním srovnávání téměř žádné rozdíly[41], jak je patrné z údajů v tabulce 8. Snižující se počty maturujících žáků pravděpodobně reflektují demografický pokles, procentuální podíl ale zůstává velice podobný.

Tabulka 8
Podíl žáků podle zřizovatele školy a pohlaví

Rok	2012	2013	2014	2015
Počet žáků celkem	26332	25491	23034	21791
Státní školy	88,0 %	87,0 %	87,5 %	87,4 %
Soukromé školy	9,7 %	10,6 %	10,0 %	10,0 %
Církevní školy	2,4 %	2,4 %	2,6 %	2,6 %
Chlapci	53,4 %	53,0 %	53,2 %	54,0 %
Dívky	46,4 %	47,0 %	46,8 %	46,0 %

4.3.2 Deskriptivní charakteristiky subtestu Poslech v letech 2012–2015

Předpokládejme, že populace sdílí stejnou distribuci měřených schopností a že není podstatný rozdíl ve struktuře populací. Podíváme-li se takto na deskriptivní statistiky v tabulce 9 a na krabicové grafy distribuce skórů pro subtesty Poslech 2012–2015 na obrázku 10, můžeme konstatovat, že nejsnazší byl subtest 2013, nejobtížnější subtest 2014, průměrná úspěšnost populací v subtestu se lišila o cca 4 až 8 procentních bodů, poloha průměrné úspěšnosti žáků kolísá mezi 10 a 13 body, poloha mediánu mezi 11 a 12 body. Tento rozdíl v průměrné obtížnosti subtestů (resp. průměrné úspěšnosti žáků) je patrný i z příslušných krabicových grafů (obr. 10). U distribuce skórů pozorujeme různou míru špičatosti – nejvíce je to patrné u subtestu 2012, kde je střední část grafu (prostor mezi 1. a 3. kvartilem) nejužší, dále různou hodnotu mediánu (linie ve střední části grafu) a průměru (značka x je pod mediánem u subtestů 2012 a 2013 a nad mediánem u subtestů 2014 a 2015).

41 V souvislosti se změnou legislativy s platností od r. 2012 je pro gymnázia povinná zkouška z anglického jazyka na B2, SOŠ včetně konzervatoří jsou tedy jediným typem školy konající zkoušku B1. Z tohoto důvodu neuvádíme v přehledu zastoupení žáků podle typu školy.

Tabulka 9
Deskriptivní statistiky pro Poslech 2012–2015

Poslech	2012	2013	2014	2015
Počet žáků	26332	25488	23034	21785
Průměrná úspěšnost žáků %	58,7	61,3	53,1	57,8
Rozdíl v úspěšnosti žáků ve srovnání s rokem 2012	–	2,6	-5,6	0,9
Rozdíl v úspěšnostech žáků (2012–2013, 2013–2014, 2014–2015)	–	2,6	-8,2	4,7
Směrodatná odchylka	18,7	22,4	22,1	21,5
Cronbachovo alpha	0,723	0,824	0,803	0,808

4.3.3 Deskriptivní charakteristiky subtestu Gramatika v letech 2012–2015

Také u subtestu Gramatika se verze 2013 jeví jako nejsnazší, naopak verze 2012 byla relativně obtížná, žáci v ní dokonce dosahovali nejnižší průměrné úspěšnosti ze všech subtestů za celé sledované období. Celkově můžeme konstatovat, že u Gramatiky meziročně velmi kolísá průměrná úspěšnost žáků (tab. 10). Potvrzuje to i pohled na krabicové grafy (obr. 11), kde pozorujeme odlišnou distribuci skórů mezi roky, jednoznačně nejšpičatější pro rok 2012, kde dokonce vidíme i několik odlehlých hodnot (tečky nad horním vousem). Průměrná úspěšnost žáků se pohybuje mezi 7 a 10 body, poloha mediánu kolísá mezi 7 a 11 body.

Tabulka 10
Deskriptivní statistiky pro Gramatiku 2012–2015

Gramatika	2012	2013	2014	2015
Počet žáků	26323	25491	23034	21791
Průměrná úspěšnost žáků %	39,4	51,6	48,1	42,8
Rozdíl v úspěšnosti žáků ve srovnání s rokem 2012	–	12,2	8,7	3,4
Rozdíl v úspěšnostech žáků (2012–2013, 2013–2014, 2014–2015)	–	12,2	-3,5	-5,3
Směrodatná odchylka	19,0	21,1	19,9	23,9
Cronbachovo alpha	0,743	0,800	0,779	0,845

4.3.4 Deskriptivní charakteristiky subtestu Čtení v letech 2012–2015

Podobně jako u subtestu Gramatika, i ve Čtení vychází jako neobtížnější verze z roku 2012. Ve verzích 2013–2015 dosáhli žáci přibližně stejné průměrné úspěšnosti (tab. 11). Z hlediska distribuce skórů jsou verze 2014 a 2015 v podstatě identické, verze 2013 se od nich liší pouze nepatrně vyšším rozptylem skórů mezi 1. a 3. kvartilem. Verze 2012 je odlišná distribucí skórů, resp. zejména špičatostí, hodnotou mediánu a průměru a jejich vzájemnou polohou (obr. 12). Průměrná úspěšnost žáků se pohybuje mezi 9 a 11 body, poloha mediánu mezi 9 a 10 body.

Tabulka 11
Deskriptivní statistiky pro Čtení 2012–2015

Čtení	2012	2013	2014	2015
Počet žáků	26323	25491	23034	21791
Průměrná úspěšnost žáků %	47,0	53	52,8	53,5
Rozdíl v úspěšnosti žáků ve srovnání s rokem 2012	x	6	5,8	6,5
Rozdíl v úspěšnostech žáků (2012–2013, 2013–2014, 2014–2015)	0	6	-0,2	0,7
Směrodatná odchylka	24,6	24,9	22,3	23,0
Cronbachovo alpha	0,851	0,865	0,812	0,828

4.3.5 Deskriptivní charakteristiky celých testů 2012–2015

Primárně jsme se zaměřili na analýzu tří subtestů řečových dovedností, neboť jen napříč subtesty lze konstrukt a obsah smysluplně porovnávat. Níže uvádíme i přehled za celý test pro všechny čtyři verze testů (tab. 12), neboť že klíčové rozhodnutí *prospěl/neprospěl*, stanovené na 33 % z maximálního skóru, je vzhledem k využití kompenzačního přístupu počítáno pro celý test dohromady. Stručně můžeme shrnout, že dílčí rozdíly identifikované v jednotlivých verzích subtestů se díky kompenzačnímu přístupu[42] stírají, i když stále lze pozorovat jisté rozdíly. Vzhledem ktomu, že populace ani testové verze 2012–2015 nejsou propojené (např. přes společné

42 Kompenzační přístup znamená, že slabý výkon v jednom subtestu je vykompenzován lepším výkonem v jiném subtestu, neboť celkový skór se počítá jako prostý součet bodů v dílčích subtestech.

položky, stejnou skupinu testovaných nebo externí test) a není prokázané, že jsou populace ve sledovaných letech ekvivalentní, není možné spočítat, zda je rozdíl v psychometrických charakteristikách testových verzí statisticky významný ani zjišťovat, jak výsledky 2012–2015 korelují. Můžeme proto jen popsat pozorované rozdíly v popisných charakteristikách testových verzí a konstatovat, že toto porovnání je platné za předpokladu, že by byly populace 2012–2015 ekvivalentní.

Testová verze 2012 se jeví jako nejobtížnější (žáci dosahují nejnižší průměrné úspěšnosti), nejsnazší je verze 2013. Rozdíl mezi průměrnými úspěšnostmi studentů je cca 3 až 7 procentních bodů, tedy asi 1,5 bodu hrubého skóru. Verze 2014 a 2015 vykazují podobné charakteristiky, obě jsou o něco obtížnější než verze 2013 a o něco snazší než verze 2012 (viz obr. 13). Nicméně pohled na testovou verzi jako celek zakrývá důležité odlišnosti ve statistických parametrech napříč subtesty 2012–2015, čemuž se věnujeme podrobněji v oddíle 4.4.6.

Tabulka 12
Deskriptivní statistiky pro celé testové verze 2012–2015

Celý test	2012	2013	2014	2015
Počet žáků	26332	25491	23034	21792
Průměrná úspěšnost žáků %	48,4	55,3	51,3	51,4
Rozdíl v úspěšnosti žáků ve srovnání s rokem 2012	x	6,9	2,9	3
Rozdíl v úspěšnosti žáků mezi sousedními roky	0	6,9	-4	0,1
Směrodatná odchylka	17,6	19,7	18,2	19,5
Cronbachovo alpha	0,895	0,922	0,904	0,918

4.3.6 Podíl neúspěšných žáků v letech 2012–2015

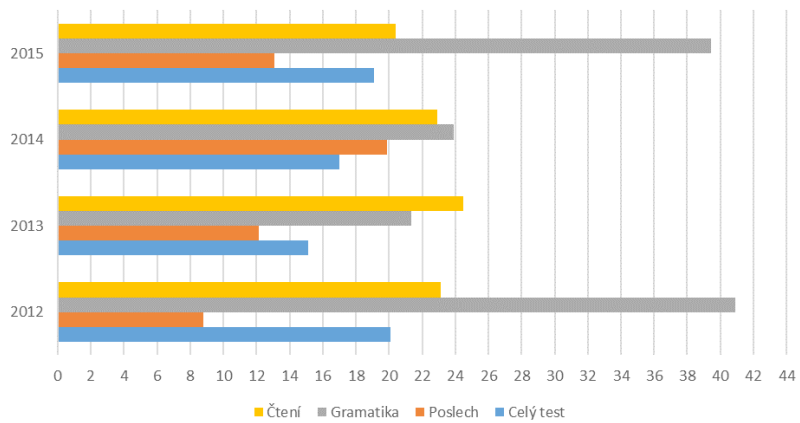
V návaznosti na předchozí oddíly doplňujeme pohled na psychometrické vlastnosti testových verzí a subtestů informacemi o podílu neúspěšných žáků. Musíme zmínit, že u slovenské maturitní zkoušky z anglického jazyka B1 platí hranice úspěšnosti 33 % v celém testu a tzv. kompenzační přístup. Není tedy důležité, jak žáci dopadnou v jednotlivých subtestech, nýbrž to, kolik bodů v celém testu získají. Mohou tedy kompenzovat slabší výkon v jednom subtestu lepším výkonem v subtestu jiném.

Tabulka 13
Podíl neúspěšných žáků v subtestech a testech 2012–2015

% neúspěšných	2012	2013	2014	2015
Poslech	8,81	12,16	19,88	13,07
Gramatika	40,93	21,34	23,92	39,45
Čtení	23,13	24,46	22,92	20,38
Celý test	20,08	15,10	17,01	19,11

V přehledu v tabulce 13 a v grafu na obrázku 14 vidíme, jak velký podíl žáků dosáhl skóru pod 33 % v jednotlivých subtestech a v celém testu. Pozorujeme odlišné podíly neúspěšných žáků napříč roky 2012–2015 a naopak vcelku podobný podíl neúspěšných žáků, pokud se díváme na celý test. Graf na obrázku 14 ukazuje tutéž informaci, a to pouze v grafické podobě. Je patrný vliv kompenzačního přístupu uplatňovaného při vyhodnocení výsledků žáků. Navíc, podíváme-li se na graf na obrázku 15, který ukazuje, jak spolu korelují výsledky žáků v jednotlivých subtestech a v subtestech a celém testu, je patrné, že pokud by byl namísto kompenzačního přístupu uplatněn přístup konjunkční (tedy žák musí prokázat minimální úspěšnost v každém subtestu), pak by podíl neúspěšných žáků byl daleko vyšší a různorodější, neboť by se neúspěšnost, respektive pravděpodobnost neúspěšnosti kumulovala s každým subtestem.

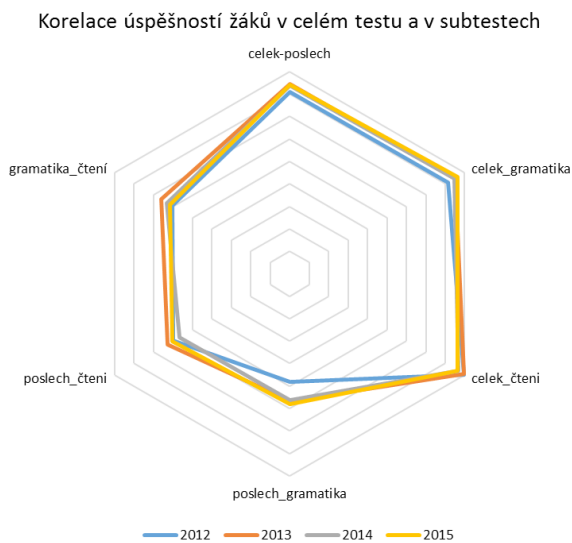
Obrázek 14
Porovnání podílu neúspěšných žáků v subtestech a testech 2012–2015



Tabulka 14
Korelace výsledků žáků v subtestech a v celých testových verzích 2012–2015

Korelace výsledků	2012	2013	2014	2015
Celek_Poslech	0,81	0,85	0,84	0,84
Celek_Gramatika	0,81	0,85	0,85	0,86
Celek_Čtení	0,90	0,89	0,87	0,86
Poslech_Gramatika	0,48	0,57	0,56	0,58
Poslech_Čtení	0,60	0,63	0,57	0,60
Gramatika_Čtení	0,60	0,66	0,63	0,62

Obrázek 15
Korelace výsledků v subtestech a v celých testových verzích 2012–2015



Graf na obrázku 15 a tabulka 14 podávají přehled o tom, jak korelují úspěšnosti žáků v jednotlivých subtestech a jak korelují výsledky v subtestech s výsledky v celém testu. Úspěšnost v jednotlivých subtestech vykazuje korelaci s úspěšnosti v celém subtestu vyšší než 0,80, avšak vzájemná korelace subtestů je poměrně nízká, mezi 0,48 a 0,63. To znamená,

že úspěšnost žáků např. v poslechu korelují s úspěšností žáků v celém testu, ale již méně koreluje úspěšnost žáků např. v poslechu a v gramatice. Pokud se podíváme např. na Poslech 2012 a 2015, vidíme, že úspěšnost v těchto dvou subtestech kompenzuje nižší úspěšnost žáků v ostatních subtestech v témže roce, zejména v Gramatice. Přitom vidíme, že právě v těchto dvou subtestech spolu výsledky žáků (jejich úspěšnost) celkově korelují nejslaběji. Také pozorujeme, že meziročně dochází k různé míře kompenzace, to znamená, že např. v roce 2012 byl poslech relativně nejsnazší a mohl kompenzovat relativně obtížnou gramatiku, naopak v roce 2014 k takové míře kompenzace nedocházelo.

4.3.7 Psychometrické charakteristiky položek a subtestů 2012–2015

Při interpretaci srovnatelnosti testových verzí je důležité brát v úvahu i psychometrické charakteristiky testů a položek v nich obsažených. Jejich vlastnosti mohou do značné míry interpretaci ovlivnit: konstrukční kvalita položek/úloh ovlivňuje žákovská řešení a má vliv i na posuzovatele, psychometrické vlastnosti se odrážejí v následných analýzách (např. EFA), v interpretaci deskriptivních statistik apod.

Ve výzkumných studiu´ích a projektech jsou často položky nebo subjekty s nevyhovujícími vlastnostmi z analýz vyřazovány. Instituce obvykle své testové nástroje podrobují přezkoumání již při jejich vývoji (před ostrým testováním) – v průběhu vývoje položek, při moderacích, po pretestacích, při kalibraci pomocí IRT apod., tedy téměř v každém kroku vývoje testů. Přesto obvykle nelze zcela zabránit tomu, aby se v testech po ostrém testování vyskytly položky s psychometrickými charakteristikami, jež nelze považovat za uspokojivé. Některé organizace, i díky tomu, že nemusí své testy zveřejňovat ihned po zkušebním termínu, mají možnost uplatnit některá z existujících nápravných opatření. Tato opatření mohou být různé povahy, od vyřazení položky a přepočtu výsledků až např. po zařazení tzv. *emergency item*, tedy určité náhradní položky, která je do testu vkládána, ale v případě bezproblémového chování ostatních položek není započítávána do výsledků; naopak pokud se v testu vyskytnou problematické položky, jsou problémové položky vyřazeny a nahrazeny výsledky těchto náhradních položek.

My jsme v našem projektu zvolili přístup neexperimentální, záměrně jsme se rozhodli neměnit nic na materiálech a výstupech, které byly využity při ostrém testování, a to s plným vědomím rizika vlivu

na výsledky, které toto rozhodnutí s sebou nese. Uvádíme tedy v tabulce 15 pouze přehled problematických položek. Vycházíme přitom především z analytických zpráv NÚCEM[43] a přebíráme i jeho kritéria pro výběr problematických položek.

Tabulka 15
Počet položek vykazujících neuspokojivé statistické parametry

	Diskriminace[1] ULI < 0,30	Obtížnost[2] < 0,20	Obtížnost > 0,80	Point biserial korelace[3] < 0,30	Cronbach alpha[4]
P2012	5	3	0	10	0,723
G2012	6	1	1	9	0,743
Č2012	0	1	0	2	0,851
Celek 2012	11	5	1	21	0,895
P2013	2	0	1	5	0,824
G2013	1	0	2	6	0,800
Č2013	0	1	0	0	0,865
Celek 2013	3	1	3	11	0,922
P2014	4	0	1	8	0,803
G2014	2	2	1	8	0,779
Č2014	0	1	0	2	0,812
Celek 2014	6	3	2	18	0,904
P2015	4	0	3	4	0,808
G2015	1	0	0	3	0,845
Č2015	0	0	0	1	0,828
Celek 2015	5	0	3	8	0,918

[1] Zde uvádíme diskriminaci ULI pouze pro položky s obtížností mezi 0,20 a 0,80.

[2] Obtížnost je zde chápána jako podíl správných odpovědí, tedy např. obtížnost 0,20 znamená, že 20 % žáků vyřešilo položku správně.

[3] Bodově biseriální korelace udává, jak koreluje položka se zbytkem testu, tedy zda měří totéž, co zbytek testu.

[4] Cronbachovo alpha udává míru vnitřní konsistence (soudržnosti) testu.

43 https://www.nucem.sk/sk/merania/narodne-merania/maturita

Vidíme, že většina subtestů obsahuje hned několik problematických položek. Jsou to položky s relativní obtížností pod nebo nad obvykle akceptovaný interval 0,20–0,80, položky problematické z hlediska schopnosti rozlišit mezi žáky celkově dobrými v řešení subtestu a celkově slabými, a položky s nízkou mírou korelace se zbytkem testu. I toto mohl být jeden z faktorů, který ovlivňoval průběh a výsledky analýzy struktury obsahu a faktorové analýzy (např. nejednoznačnost formulace položek, cílení na dovednosti, které nejsou součástí konstruktu popsaného popisnými modely apod.). Vidíme také relativně nízké hodnoty koeficientu vnitřní soudržnosti testů (Cronbachovo alpha) u subtestů (0,723–0,865), poněkud vyšší pak u celých testových verzí. Chráska (2007, s. 21) definuje Cronbachovo alpha jako koeficient udávající vnitřní soudržnost (konzistenci) testu, stejné významové zaměření položek, jež můžeme chápat i jako příslušnost ke shodnému konstruktu. Problematiku zmiňujeme v souvislosti s výstupy exploratorní faktorové analýzy v kapitole 4 a také v závěrečné kapitole 5.

Pohled na psychometrické vlastnosti testů a položek vhodně doplňuje výstupy EFA. Z obou pohledů vyplývá, že subtesty obsahují položky, které se svými vlastnostmi a chováním odlišují od zbytku testu (resp. subtestu), a pravděpodobně měří něco jiného než ostatní položky. Pravděpodobně jsou tedy konstruktově irelevantní, nebo mají jiné konstrukční nedostatky.

4.4 Shrnutí empirické části výzkumu

Hlavními cíli projektu bylo: 1. zmapovat metody umožňující vznik srovnatelných testových verzí, tedy verzí, jejichž skóry mají shodnou interpretaci ve smyslu měřeného rysu, je možné je srovnávat a za předem stanovených (a posléze i naplněných) podmínek považovat za ekvivalentní; 2. vyzkoušet a ověřit praktičnost a funkčnost vybraných metod na čtyřech realizovaných verzích slovenské MZ z AJ B1; 3. na základě teoretického výzkumu a empirické části navrhnout takové metody nebo jejich kombinaci, které by mohly být využívány v kontextu vývoje maturitních testů na Slovensku, a to ve fázích před ostrým testováním nebo po něm. Jedním z kritérií pro výběr některých metod a postupů mezi

doporučené[44] bylo kromě již zmíněných výsledků empirické části také kritérium, aby zavedení těchto metod nemuselo být podmíněno legislativními změnami, jinými slovy jejich případné zavedení by předpokládalo pouze interní opatření realizované poskytovatelem zkoušky.

Pracovali jsme s reálnými daty za čtyři ročníky maturitní zkoušky (více než 20 000 respondentů pro každou testovou verzi), neprováděli jsme žádná experimentální šetření ve speciálním designu. Tato data nám poskytl NÚCEM.

Zaměřili jsme se na vybrané metody často citované v odborné literatuře zabývající se srovnatelností testových verzí nebo ekvivalencí skórů, konkrétně na analýzu struktury obsahu z pohledu jeho srovnatelnosti realizovanou pomocí expertního posuzování, dále na faktorovou analýzu pro zkoumání srovnatelnosti konstruktu a na porovnání psychometrických vlastností testových verzí a populací testovaných.

Výběr metod byl ovlivněn snahou realizovat výzkum s daty získanými tak, jak jsou k dispozici poskytovateli zkoušky NÚCEMu, a s vědomím všech omezení, která tato rozhodnutí přináší. Zvolené metody, resp. jejich provedení tím ovlivněny nebyly. Toto rozhodnutí však přineslo komplikace při interpretaci výsledků, např. do jaké míry lze populace z let 2012 až 2015 považovat za ekvivalentní z hlediska rozložení sledovaného konstruktu – úrovně jazykové způsobilosti, jak naložit s položkami s ne zcela ideálními psychometrickými vlastnostmi, jak přistoupit k nejednoznačným postupům posuzovatelů při obsahové analýze, jak interpretovat faktory, jsou-li do faktorové analýzy zahrnuty položky vykazující nízkou korelaci se zbytkem testu.

Výsledky a jejich interpretace se totiž řetězí. Analýza struktury obsahu je sice provedením zcela nezávislá na exploratorní faktorové analýze, neboť první pracuje s popisnými modely a zněním testových položek, druhá pracuje s reálnými odpověďmi žáků na tyto položky. Nicméně konstrukční kvalita nástroje (testových verzí, resp. položek v nich obsažených) má vliv na hodnocení posuzovatelů a také na psychometrické vlastnosti položek, které se projeví při faktorové analýze atd. Znamená to komplikace a případnou kumulaci nepřesností napříč analýzami a interpretací výsledků. Na druhé straně tyto problémy vedly k potvrzení toho, jak důležité je provádět validaci všech kroků vývoje

44 Doporučením myslíme pouze označení metody, která by mohla být využitelná v situaci nebo v kontextu podobném tomu, v jakém se nachází slovenská maturitní zkouška; v žádném případě nechceme ze své pozice doktoranda určovat, jak by měla postupovat konkrétní instituce.

a použití testu a jak propojené jsou kroky vývoje testů: jasný účel testu, podrobná, teoreticky podložené a empiricky ověřené specifikace testů, správně nastavený proces vývoje testových úloh, jejich revize, moderace, pretestace provedené takovým způsobem, aby kromě zpětné vazby o psychometrických vlastnostech položek jejich design umožnil následné vyrovnávání skórů, s dopředu jasnou představou o paradigmatu analýz – klasickou teorií testů nebo teorií odpovědi na položku, a s možností vyrovnávání obhajitelného v dlouhodobé perspektivě.

4.4.1 Zjištění z analýzy struktury obsahu

Naším cílem bylo vyzkoušet metodu analýzy srovnatelnosti obsahu a postupy s ní související. Pro analýzu byli oslovení zkušení posuzovatelé, vytvořeny dva popisné modely, posuzovatelé byli vyškoleni v práci s nimi a posuzovány byly čtyři testové verze slovenské maturitní zkoušky z anglického jazyka B1 realizované v letech 2012–2015. Na základě získaných dat a vyhodnocení průběhu celého procesu shrnujeme následující zjištění:

Reflexe metody a modelů

Provedení analýzy struktury obsahu bylo poměrně časově náročné pro posuzovatele (posouzení celkem 240 položek ve čtyřech testových verzích) i pro výzkumníka (design sběru dat, školení posuzovatelů, slučování a čištění dat a následující několikeré opakované zpracování). V reálné situaci by pravděpodobně byla časová zátěž daleko nižší, předpokládáme-li, že by byly posuzovány nikoli čtyři testové verze, nýbrž pouze dvě, případně dokonce jedna, pokud by se pracovalo s jednou, stále stejnou referenční verzí, jejíž struktura obsahu by již byla dána, a bylo by třeba provést analýzu struktury obsahu pouze pro novou verzi.

Data byla analyzována několika způsoby, jejichž kombinace poskytla velmi komplexní pohled na obsah testových verzí: analýza četnosti shody, grafické zobrazení zastoupení popisných kategorií a porovnání pohledu jednoho posuzovatele na čtyři verze a pohled skupiny posuzovatelů na jednotlivé verze, výpočet shody hodnotitelů pomocí koeficientů procentuální shody a AC1. Zároveň byl tento pohled využit při interpretaci výstupů exploratorní faktorové analýzy.

Ve struktuře dat se projevily kappa paradoxy, tj. převaha určitých kategorií a vysoká míra shody na některé z kategorií, což však nepovažujeme za negativní nález, nýbrž to odpovídá povaze posuzovaných dat. Všechny kroky a procesy (pilotáž, školení, kódování, čištění dat, slučování kategorií atd.) byly aplikovány v souladu s teorií i praxí popsanou v literatuře a vycházející i z vlastní praxe výzkumníka.

Data získaná z expertního posuzování bylo obtížné připravit pro zpracování, analyzovat a interpretovat vzhledem k ne zcela jednoznačným vstupním informacím od posuzovatelů. Data byla do určité míry nesystematicky zatížena dvojznačností (položky přiřazené k více než jednomu deskriptoru), v některých subtestech také odlišným přístupem některých posuzovatelů k interpretaci popisných kategorií nebo ke vztahu položek k těmto kategoriím. Z tohoto důvodu jsme přistoupili k redefinici popisných kategorií, resp. k jejich sloučení do nadřazených, obecněji formulovaných kategorií. Tato počáteční dvojznačnost v datech mohla ovlivnit výsledky a jejich interpretaci, proto navrhujeme v kapitole 5 některé kroky související zejména s vytvářením popisných modelů a školení posuzovatelů, jež by mohly tento problém eliminovat.

Přestože posuzovatelé byli vyškoleni v práci s popisnými modely a byl kladen důraz na shodu v interpretaci popisných kategorií, při samostatné práci na posouzení obsahu subtestů se odlišná interpretace kategorií projevila. Toto mohlo být způsobeno několika faktory. Některé z nich připisujeme způsobu, jakým bylo provedeno školení, tj. jednokolové distanční posuzování. Dalším faktorem, který mohl hrát roli, je samotná povaha úkolu posuzovatelů, který byl založen na subjektivním posuzování latentních charakteristik položek a na práci s abstraktními modely. Uvažovat lze i o pochybnosti vyslovené např. Weirem (2005), do jaké míry je skutečně možné považovat procesy pojmenovávané popisnými kategoriemi modelů za oddělitelné, samostatně existující, tj. diskrétní kategorie neboli zda ve skutečnosti neplatí spíše to, že při řešení určitého úkolu (zde reprezentovaného položkami) probíhá více procesů zároveň. Tato pochybnost je reflektována i při interpretaci výsledků exploratorní faktorové analýzy, jež na základě dostupných žákovských dat naznačuje podobné chování subkonstruktů receptivních dovedností v subtestech Poslech, Gramatika a Čtení.

Reflexe zjištění

Jednotliví posuzovatelé hodnotí strukturu obsahu verzí většiny subtestů 2012–2015 jako podobnou, nikoli identickou. Shoda posuzovatelů jako celé skupiny na stejné struktuře obsahu napříč všemi verzemi 2012–2015 ale není jednoznačná. Také se zdá, že dochází k meziročním posunům ve struktuře obsahu. Verze z let po sobě následujících jsou si obvykle podobnější než verze od sebe časově více vzdálené.

U Poslechu lze říci, že jednotliví posuzovatelé považují verze za velmi podobné, i když ne zcela srovnatelné, a to v obou modelech. Jako skupina se ale neshodují na zastoupení konkrétních deskriptorů v subtestech – odlišně interpretují popisné kategorie.

U Gramatiky považují jednotliví posuzovatelé strukturu obsahu verzí 2012–2015 za odlišnou (pouze H4 v modelu kognitivních procesů považuje subtesty za shodné).

U Čtení jednotliví posuzovatelé považují verze za vcelku srovnatelné, a to v obou modelech, ale jako skupina se na struktuře obsahu neshodují v modelu dle SERRJ, naopak v modelu kognitivních procesů ano (kde jsou však pouze dvě popisné kategorie).

Zjištěné koeficienty shody AC1 interpretované pomocí Gwetova koeficientu kumulativní příslušnosti spadají z větší části do kategorie průměrné až dobré shody, s výjimkami jako např. Čtení v modelu kognitivních procesů, kde lze shodu považovat za výbornou, a naopak u Poslechu v modelu kognitivních procesů je shoda pouze slabá. Příčin může být velmi mnoho, od definice modelu, designu posuzování, obvyklosti používání popisných kategorií, kvality položek v posuzovaném subtestu a v neposlední řadě také skutečné rozdíly mezi testovými verzemi. Analýza struktury obsahu může ukázat na rozdíly, ovšem jejich vysvětlení je v rukou výzkumníka.

Důležité je ale především zhodnocení toho, zda by tato metoda představovala přínos při snaze dosáhnout srovnatelnosti testových verzí. Z tohoto pohledu konstatujeme, že proces analýzy struktury obsahu vyžadoval od výzkumníka mnohá rozhodnutí (nakládání s chybějícími daty, rozhodování o položkách s dvěma a více přiřazenými deskriptory, slučování popisných kategorií, rozhodnutí o případném vyřazení některého z posuzovatelů apod.), což mohlo mít vliv na konkrétní výsledky v tomto projektu, a proto v kapitole 6 doporučujeme postupy, které by umožnily se těmto komplikacím vyhnout. Dále je třeba upozornit,

že posuzovatelé pracovali se skutečnými testovými verzemi. Žádná položka nebyla vyřazena. Zdůrazňujeme tento fakt proto, že relativně velké množství položek vykazuje nevyhovující hodnoty psychometrických parametrů a v subtestech se objevují některé testologické nedostatky (například problematická testovací technika ve třetí části poslechu). Přes všechna výše uvedená úskalí tvrdíme, že analýza struktury obsahu zahrnující expertní posuzování, grafický pohled na strukturu obsahu, analýzu četnosti shody a výpočet shody a reliability posuzovatelů jako celek poskytuje velmi komplexní pohled na strukturu subtestů, umožňuje jejich porovnání a identifikaci problematických míst. Nástroje využité v tomto projektu považujeme za užitečné a využitelné i v jiných kontextech a doporučujeme zejména jejich kombinaci, tak jak to bylo provedeno i v tomto projektu.

Na základě získaných výsledků nelze jednoznačně tvrdit nebo popřít, že testové verze 2012–2015 jsou obsahově srovnatelné. Prokázání srovnatelnosti či nesrovnatelnosti však nebylo cílem tohoto projektu, cílem bylo zjistit, jaké metody pro řešení této otázky by byly využitelné.

4.4.2 Analýza konstruktové ekvivalence testových verzí 2012–2015

Exploratorní faktorová analýza byla zvolena jako metoda umožňující kvantitativní pohled na latentní strukturu subtestů – konstrukt ověřovaný prostřednictvím testových položek. Výstupy EFA lze pro jednotlivé subtesty porovnávat a zjistit, do jaké míry lze považovat konstrukty za srovnatelné. Vstupními daty pro EFA byly žákovské odpovědi, více než 20 000 respondentů pro každou testovou verzi, což je velmi vysoké číslo, a proto bychom mohli výsledky EFA považovat za průkazné. U všech subtestů jsme zjistili dominanci jednofaktorového řešení, u některých subtestů to byl faktor sycený položkami napříč částmi (např. Gramatika a Čtení), jinde byl faktor sycen položkami převážně z jedné části subtestu (poslech). V zásadě lze říci, že interpretace s využitím závěrů analýzy struktury obsahu by byla akceptovatelná, ačkoli lze také uvažovat o jednorozměrnosti těchto tří řečových dovedností, o čemž pojednáváme výše.

Nicméně EFA byla zproblematizována kvalitou vstupních dat. Korelační matice pro položky v datových souborech (celkem 24 matic) ukázaly nižší míru korelace zejména v subtestech Poslechu, což může indikovat, že položky s nízkými korelačními koeficienty měří něco jiného

než zbytek subtestu. Tyto položky (nejen v subtestu Poslech) byly položky, které vykazovaly v deskriptivní analýze nízké hodnoty bodově biserální korelace nebo citlivosti (diskriminace ULI). V EFA se to pak projevilo nižšími hodnotami sdíleného rozptylu a nízkými hodnotami faktorových zátěží, nebylo tedy možné je vztáhnout k nalezenému faktoru. Při výzkumu, jež by měl vést k přípravě designu pro vyrovnávání, k němuž je nutnou podmínkou prokázání konstruktové ekvivalence, by se nemělo stávat, že se v datech objeví tolik položek s nevyhovujícími hodnotami, tyto nedostatky by měly být odstraněny již v raných fázích vývoje testu.

Nálezy EFA jsou v souladu s tím, co bylo pozorováno během analýzy struktury obsahu, kdy se posuzovatelé také v některých případech neshodovali na přesné interpretaci toho, co je položkou ověřováno a jaké je zastoupení kategorií v subtestech.

5 Odpovědi na výzkumné otázky 1 a 2

V empirické části výzkumu jsme se pokusili odpovědět na dvě výzkumné otázky:

> RQ1: Do jaké míry jsou testové verze slovenské maturitní zkoušky z anglického jazyka B1 ekvivalentní z hlediska obsahu, konstruktu a psychometrických vlastností, jaké povahy jsou případně zjištěné odlišnosti a jak zásadní jsou pro interpretaci výsledků?
>
> RQ2: Jsou metody zjišťování srovnatelnosti testových verzí použité v tomto výzkumu dostatečně průkazné, spolehlivé a praktické, aby mohly být používány i v kontextu slovenské maturitní zkoušky?

5.1 Srovnatelnost testových verzí 2012–2015 (RQ1)

Obsahová srovnatelnost

Provedenou analýzou struktury obsahu jsme došli k závěru, že stávající realizované testové verze subtestů ve slovenské maturitní zkoušce z anglického jazyka B1 mají podobnou strukturu obsahu, nelze však jednoznačně tvrdit, že jsou obsahově zcela srovnatelné. Jednotliví posuzovatelé vyhodnotili jako identické pouze subtesty Čtení 2012–2015, a to v obou popisných modelech. Subtest Čtení, ovšem pouze v modelu kognitivních procesů s pouhými dvěma popisnými kategoriemi, byl také jediným subtestem, kde se posuzovatelé ve vysoké míře shodli i jako celá skupina na tom, že jsou subtesty srovnatelné. U ostatních subtestů je podle posuzovatelů struktura subtestů napříč roky odlišná, nepozorovali jsme ale žádný systematický trend těchto změn.

Míra shody posuzovatelů na struktuře obsahu se lišila napříč subtesty a modely. Vynikající míru shody na struktuře subtestů, nikoli však na jejich srovnatelnosti, měli posuzovatelé v subtestech Gramatika v obou modelech a u subtestu Čtení v modelu kognitivních procesů (jde o excelentní úroveň shody dle Gwetova interpretačního rámce; Gwet, 2014). U Poslechu a Čtení v modelech dle SERRJ byla shoda vyhodnocena jako průměrná až dobrá, u Poslechu v modelu kognitivních procesů jako slabá.

Konstruktová ekvivalence

Konstrukty všech tří subtestů v letech 2012–2015 vykazují spíše jednofaktorovou strukturu. V Poslechu je dominantní faktor sycen převážně položkami ze třetí části subtestu a několika položkami z ostatních částí, avšak ty mají mnohem nižší hodnoty faktorových zátěží. Subtest Poslech 2012 vykazuje jinou faktorovou strukturu než subtesty ostatní, které jsou si vzájemně podobnější. V Gramatice je dominantní faktor sycen různými položkami napříč oběma částmi subtestu, což je pochopitelné vzhledem k povaze tohoto subtestu zaměřeného spíše na jazykovou znalost (znalost lexika, morfosyntaxe, syntaxe atd.). Výrazně vyšší hodnoty faktorových zátěží se koncentrují zejména u položek v části 2. U Čtení byla zjištěná jednofaktorová struktura nejvýraznější, téměř všechny položky vykazovaly vysoké hodnoty faktorových zátěží.

Většina subtestů však vykazuje nezanedbatelné množství položek, které se ke společnému faktoru nevztahují, nebo jen velmi slabě, případně mají problematické psychometrické charakteristiky. Nelze tedy jednoznačně potvrdit strukturu konstruktu ani konstruktovou ekvivalenci testových verzí.

Srovnatelnost populací 2012–2015

Pro porovnávání struktury konstruktu a psychometrických vlastností položek v subtestech jsme museli učinit předpoklad, že populace z let 2012–2015 jsou ekvivalentní. Z pohledu složení populací se toto konstatování můžeme opřít o popisné statistiky populací. Ty jsou z hlediska zastoupení skupin (podle pohlaví, typu školy apod.) srovnatelné. Nicméně o distribuci měřené vlastnosti (úrovni jazykové kompetence) nemáme k dispozici žádné doplňující informace, data nejsou propojena a nesdílí tedy žádný referenční bod.

Psychometrické vlastnosti subtestů

U Poslechu a Gramatiky se objevilo nezanedbatelné množství položek s nízkou citlivostí a nízkými hodnotami bodově biserální korelace, což znamená, že tyto položky hůře nebo vůbec nerozlišovaly mezi žáky celkově dobrými celkově slabými v řešení celého subtestu (diskriminace ULI), a dále že jejich souvislost s tím, co měřil zbytek testu, byla velmi slabá (*point biserial correlation*). Právě tyto položky vykazovaly nízké hodnoty korelačního koeficientu *r* v korelační matici a zároveň nízké hodnoty faktorových zátěží. U Čtení se takovéto položky prakticky nevyskytovaly. Tato zjištění o nízké citlivosti položek a jejich nízké míře korelace s tím, co měří zbytek testu, velmi problematizuje provedení analýzy konstruktu a interpretaci výstupů. Přítomnost těchto položek zároveň mohla mít vliv i na chování posuzovatelů a analýzu struktury obsahu.

Z hlediska rozložení skórů jsme u porovnávaných testových verzí zjistili, že u Poslechu se jako nejsnazší jeví verze subtestu 2013, nejobtížnější je verze 2014. U subtestu Gramatika je nejobtížnější verze 2012, verze 2013 je nejsnazší. Také u Čtení vychází jako neobtížnější verze z roku 2012. Subtesty jsou odlišné distribucí skórů, špičatostí, hodnotou mediánu a průměru a jejich vzájemnou polohou.

U testových verzí jako celků se sice tyto rozdíly poněkud stírají, nicméně stále je nejobtížnější verze 2012 a nejsnazší verze 2013, verze 2014 a 2015 vykazují přibližně shodné hodnoty.

Nahlíželi jsme na testové verze 2012–2015 z různých úhlů pohledů a různými metodami. Na základě zjištění a s vědomím omezení, která vyplývají z toho, že testové verze ani design sběru dat nebyl nijak modifikován, se kloníme k závěru, že testové verze jsou si podobné z hlediska struktury obsahu a konstruktu, nevykazují však jednoznačně obsahovou a konstruktovou ekvivalenci. Testové verze navíc obsahují v různé míře položky s neuspokojivými psychometrickými vlastnostmi**.** To vše by u testových verzí maturitní zkoušky z anglického jazyka na úrovni B1 2012–2015 vyloučilo smysluplné provedení vyrovnávání testových skórů.

Shrnutí

Obsahová nesrovnatelnost může mít několik vzájemně souvisejících příčin. Oficiální testové specifikace zveřejněné na stránkách NÚCEMu, podle kterých se sestavují maturitní testy z anglického jazyka, jsou formulovány velmi obecně, což v důsledku umožňuje velkou volnost při naplňování subtestů konkrétním obsahem. Specifikace explicitně zmiňují, že test ověřuje úroveň jazyka B1 podle SERRJ, avšak popisy řečových činností (definice konstruktu) nejsou totožné s deskriptory B1 SERRJ a nenalezli jsme dokumentaci, která by prokazovala lokalizaci deskriptorů (modifikaci pro určitý kontext) a realizaci procesu přiřazení specifikací a celého testu k SERRJ. V rámci tvorby úloh a položek tak může docházet k odlišné interpretaci specifikací a naplnění subtestů obsahem, a pokud neexistuje propracovaná zpětná vazba typu analýzy obsahu pomocí expertního posouzení vyškolenými odborníky, může být každá verze naplňována poněkud odlišně. Vliv na analýzu obsahu mohly mít i položky vykazující neuspokojivé psychometrické charakteristiky, což mohlo být způsobeno ne zcela vhodnými konstrukčními vlastnostmi položek.

Zároveň připouštíme i možné nedokonalosti v nástrojích a provedení analýzy struktury obsahu ze strany výzkumníka, což reflektujeme v kapitole 5.

Výstupy z EFA potvrzují nejasnosti zjištěné při analýze struktury obsahu a analýze psychometrických vlastností testových verzí a ani konstruktová ekvivalence nebyla dostatečně prokázána. V návaznosti na analýzu struktury obsahu konstatujeme, že testové verze sice vycházející ze shodných specifikací, ty ale umožňují natolik širokou interpretaci, že nemusí dostatečně vést tvůrce úloh a sestavovatele testu k dostatečně podobné operacionalizaci konstruktu, což se projevilo v datech získaných od posuzovatelů a ve výsledcích exploratorní faktorové analýzy. Testové verze navíc obsahují položky s ne zcela uspokojivými psychometrickými charakteristikami, což také mohlo ovlivňovat strukturu konstruktu a jeho intepretaci ze strany testovaných, což mohlo způsobit přítomnost konstruktově irelevantní variance a následně se projevit při faktorové analýze jako nejasná struktura faktorů a vztahů mezi položkami a faktory, nízké faktorové zátěže apod.

Konstatujeme proto, že za použití výše uvedených metod v neexperimentálním designu jsme nebyli schopni spolehlivě potvrdit obsahovou a konstruktovou ekvivalenci testových verzí 2012–2015. Testové verze z hlediska struktury obsahu, konstruktu a psychometrických vlastností zcela srovnatelné nejsou, a z tohoto důvodu by nebylo možné provést vyrovnávání skórů. Domníváme se dále, že není možné prokázat srovnatelnost či odlišnosti testových verzí ex post, tzn. po realizaci testování, bez promyšlených kroků týkajících se sběru dat v designu vhodném pro pozdější vyrovnávání, a případně i bez provedení změn v různých fázích procesu vývoje testů. Možné návrhy těchto opatření diskutujeme v kapitole 5.

5.2 Zhodnocení použitých metod (RQ2)

Zdůrazňujeme, že cílem tohotoo projektu nebylo srovnatelnost testových verzí prokázat, nýbrž na reálných datech, ve skutečném kontextu a se všemi omezeními vyzkoušet některé metody vedoucí k zajištění srovnatelnosti testových verzí a zjistit, zda a jak fungují v tomto kontextu a jaké aspekty by měly být vylepšeny. Jako vysoce funkční se ukázala kombinace kvalitativního a kvantitativního přístupu a vyhodnocení z několika perspektiv (grafické přehledy o struktuře obsahu, tabulkové přehledy četností shod a statistický výpočet shody), neboť výsledky se vzájemně doplňovaly, což napomáhalo při interpretaci zjištění. Kromě komplexního pohledu na testové verze bylo možné zjistit i problematická místa v aplikaci těchto metod a postupů a následně tyto problematické aspekty reflektovat. V použitých metodách, především v kombinaci analýzy struktury obsahu vyhodnocené pomocí tabulkových a grafických přehledů a koeficientu AC1, faktorové analýzy (zde exploratorní) a analýzy psychometrických charakteristik testových verzí spatřujeme jednoznačný potenciál a přínos pro řešení otázky srovnatelnosti testových verzí. Doporučujeme tyto metody zavést jako rutinní krok do procesu vývoje úloh a sestavování testových verzí, a to jak pro posuzování shody obsahu testových verzí se specifikacemi, tak pro meziroční porovnávání struktury obsahu napříč testovými verzemi. Výstupy z těchto analýz poskytují také cennou zpětnou vazbou pro tvůrce a sestavovatele testů, neboť informují o míře shody mezi záměrem tvůrců (vtěleným do specifikací a popisu konstruktu a realizovaným testovými úlohami) a skutečnosti,

jež se projeví v reálných datech od testovaných. I přes některá problematická místa, o nichž píšeme v kapitole 4 a jejichž změny navrhujeme v kapitole 5, je považujeme za přínosné a za funkční postup na cestě k srovnatelným testovým verzím a ekvivalentním skórům. O jejich spolehlivosti a praktičnosti do budoucna si netroufáme vyslovovat soudy, neboť způsob a rozsah jejich uplatnění nedokážeme predikovat.

III. APLIKAČNÍ ČÁST

6 Návrhy procesů pro vývoj srovnatelných testových verzí (RQ3)

V předchozích částech výzkumu, tedy v teoretické části I, jež vycházela z dostupné literatury, a v empirické části II, kde jsme aplikovali některé z vybraných metod, jsme hledali dílčí odpovědi také na výzkumnou otázku 3.

> RQ3: Jaké metody a postupy by mohly být zavedeny do procesu vývoje testových verzí slovenské maturitní zkoušky z anglického jazyka v rámci stávající legislativy, aby bylo dosahováno ekvivalence skórů a srovnatelnosti používaných testových verzí?

Na základě poznatků z obou částí se níže pokoušíme o výčet oblastí, kterým by měla být věnována pozornost vždy, když poskytovatel testu usiluje o dosažení srovnatelnosti testových verzí, a navrhnout opatření nebo postupy, jejichž zavedení by v těchto oblastech mohlo být zváženo. Doporučení vycházejí z provedeného výzkumu v kontextu slovenské maturitní zkoušky z anglického jazyka na úrovni B1, včetně problematických aspektů, které souvisely s neexperimentálním designem simulujícím ex-post zkoumání míry srovnatelnosti testových verzí. Domníváme se, že zjištění a naznačená řešení mohou být využitelná i v jiných kontextech u jiných zkoušek. Zároveň si uvědomujeme, že existují i jiné, alternativní postupy, metody a jejich provedení, a v žádném případě netvrdíme, že jsou námi uvedená doporučení vyčerpávajícím výčtem.

Považujeme za nezbytné konstatovat, že v žádné z fází výzkumného projektu, tedy ani nyní, nebylo naším cílem kritizovat slovenskou maturitní zkoušku a srovnatelnost či nesrovnatelnost testových verzí, nebo určovat, co by tvůrci a poskytovatelé slovenské maturitní zkoušky měli nebo neměli dělat. O vnitřních procesech a omezeních v NÚCEMu

víme velmi málo, nemůžeme tedy činit žádné konkrétní závěry. Můžeme pouze uvést, jaké možnosti existují, zhodnotit, jak fungovaly v kontextu tohoto projektu a navrhnout jejich vylepšení nebo upozornit na rizika a úskalí, pokud by je chtěli tvůrci zkoušek v podobném situaci využít. Zdůrazňujeme také, že bez laskavého svolení NÚCEMu a poskytnutí dat, s nimiž jsme pracovali v exploratorní faktorové analýze, by výzkum v této podobě nemohl být realizován.

Na úvod výčtu doporučení pro vývoj srovnatelných testových verzí shrnujeme, že cesta ke skórům, které pocházejí z různých testových verzí téže zkoušky, ale mohou být z hlediska jejich interpretace považovány za zaměnitelné, začíná na samém počátku vývoje testu. Plán na vyrovnávání skórů je obvykle dlouhodobý, nezahrnuje jen dvě po sobě následující testové verze, zasahuje do designu pretestací jakožto procesu ověřování kvality úloh, staví na jednoznačném prokázání obsahové a konstruktové ekvivalence a podobnosti psychometrických charakteristik testových verzí, na volbě metody vyrovnávání, na průběžné dokumentaci jednotlivých kroků a výsledků a závěrečném kritickém zhodnocení celého procesu.

6.1 Účel zkoušky, její konstrukt a testové specifikace

Konstrukt slovenské maturitní zkoušky z anglického jazyka na úrovni B1 není explicitně definován, lze na něj pouze usuzovat ze zmínek o požadované výstupní úrovni B1 a z popisu toho, co ověřují subtesty Poslech, Čtení a Gramatika ve specifikacích[45]. Konstrukt zkoušek by měl být vždy formulován velmi jasně a explicitně, měl by zahrnovat účel zkoušky a s ním související interpretaci výsledků, doporučení pro jejich využití a varování před tím, co z výsledků vyčíst nelze a k čemu je tedy nelze využívat. Pokud definice konstruktu deklaruje vztah k jedné z úrovní SERRJ, je třeba, aby proběhl řádný proces přiřazení a aby byl zdokumentován.

Kvalita testových specifikací je nezbytná pro vytváření testových verzí, jež mají být porovnatelné. Měly by být formulovány tak, aby bylo jasné, jak je konstrukt prostřednictvím úloh, případně položek, realizován, dále jaká je váha a zastoupení jednotlivých ověřovaných cílů. Testové specifikace by měly být konkrétní a podrobné a měl by být jasně zdokumentován vztah testových specifikací a podle nich vytvářených

45 http://www.statpedu.sk/files/articles/nove_dokumenty/cielove-poziadavky-pre-mat-skusky/anglicky-jazyk_b1b2.pdf

testových verzí k SERRJ, a to zejména ze dvou důvodů: za prvé, pokud má maturitní zkouška v názvu i v dalších kurikulárních dokumentech označení vztahu k jedné z úrovní SERRJ, je třeba tento vztah teoreticky podložit i empiricky prokázat, aby bylo možné toto označení využívat při interpretaci výsledků; za druhé, podrobné testové specifikace jsou užitečný a důležitý nástroj pro tvůrce úloh a položek a při vytváření přinejmenším obsahově srovnatelných variant; za třetí, specifikace testu mohou být v budoucnu využívány při analýze srovnatelnosti obsahu, a to jako podklad pro popisné modely. Všechny tyto kroky a materiály pak mohou být součástí validačního procesu slovenských maturitních zkoušek z anglického jazyka. Význam podrobnosti a konkrétnosti testových specifikací, provedení studie porovnávající vztah specifikací, potažmo testů, k SERRJ a dokumentování a zdůvodnění lokálních úprav deskriptorů SERRJ pro použití v nových kontextech zdůrazňují např. Alderson a kol. (2006) nebo Weir (2005), který dokonce říká, že podrovná specifikace, validace konstruktu a jeho teoreticky podložené vymezení je stejně důležité jako prokázání statistické ekvivalence testových verzí (s. 283). V obecné rovině o významu kvalitní definice konstruktu, specifikací a popisu výkonových standardů hovoří i Standardy (AERA, APA, & NCME, 2014).

V této souvislosti doporučujeme revidovat či dovysvětlit definici účelu zkoušky. V dokumentu *Cieľové požiadavky na vedomosti a zručnosti maturantov z anglického jazyka úroveň B1*[46] je deklarováno, že cílem maturitní zkoušky z anglického jazyka na úrovni B1 je „zosúladenie požiadaviek na jazykové vedomosti a spôsobilosti žiakov, zjednotenie kritérií hodnotenia komunikačnej jazykovej kompetencie žiakov a dosiahnutie objektivity maturitnej skúšky“. Z takto formulovaného cíle lze odvodit, že jde o tzv. ověřující test, jehož cílem je ověřit dosažení určité požadované úrovně znalostí a dovedností (definované v kurikulárních dokumentech nebo dokumentech popisujících zkoušku – zde Špecifikácia). V dokumentu Špecifikácia testov z cudzích jazykov úroveň B1 pre externú časť a písomnú formu internej časti maturitnej skúšky se ale objevuje informace, že jde o test rozlišující, jehož smyslem je porovnat výkony žáků, nikoli ověřit dosažení určitého standardu (jakým je např. úroveň B1, k níž se test vztahuje). Dále, pokud je stanovena mezní hranice úspěšnosti, jež má být interpretována jako minimální

46 https://www.nucem.sk/dl/749/Specifikacia_testu_B1_MS_2018_web.pdf

hranice pro rozhodnutí, zda žák uspěl nebo neuspěl, případně hranice pro přidělování známek, měl by proběhnout proces stanovení standardu – minimální hranice úspěšnosti (*standard setting*), buď pro výrok *uspěl/neuspěl,* nebo pro výrok *Jazyková způsobilost žáka je na úrovni B1*. Tento proces by měl být zdokumentován. Je tedy nutné definovat význam rozhodnutí *uspěl*: zda lze tento výrok interpretovat jako dosažení minimálních požadavků na maturanta, nebo dosažení minimálních požadavků očekávaných od uživatele jazyka na úrovni B1, případně obojí, a v jakém vztahu je k tomuto výroku současných 33 %.

6.2 Prokázání obsahové a konstruktové srovnatelnosti testových verzí

Po revizi účelu zkoušky je možné dobře definovat konstrukt, od něj odvodit podrobné specifikace, ty operacionalizovat při vytváření testových položek a úloh. Tyto vstupní informace je také možné využít pro posouzení srovnatelnosti obsahů zkoušek pomocí analýzy struktury obsahu napříč testovými verzemi a případně využít výsledky analýzy struktury obsahu při specifikaci modelu pro faktorovou analýzu.

Testové specifikace jako základ popisných modelů

Analýza struktury obsahu pomocí expertního posuzování je jednou z metod, které mohou být využívány při sestavování obsahově srovnatelných testových verzí a vést k naplnění jedné z podmínek pro dosahování srovnatelnosti testových verzí, tedy prokázání, že porovnávané testové verze jsou odvozeny od shodných specifikací. Důležitými podmínkami pro to, aby byla metoda expertního posuzování srovnatelnosti obsahu efektivní a užitečná, je kvalita popisných nástrojů, výběr expertů, jejich proškolení a samotné provedení metody.

Pokud by existovaly dostatečně podrobné testové specifikace, teoreticky podložené a vztažené k definici konstruktu, a pokud by tento vztah byl validován, pak mohou specifikace být základem pro tvorbu popisných modelů. Tyto popisné modely by měly být recenzovány a vyzkoušeny, např. tvůrci úloh, externími i interními spolupracovníky. Práce se specifikacemi a popisnými modely by mohla, v ideálním případě měla být běžným krokem při sestavování testových verzí, případně v určité podobě i součástí procesu tvorby úloh.

Výběr posuzovatelů

Pro analýzu struktury obsahu testových verzí, u nichž se předpokládá uplatnění procesu vyrovnávání testových skórů, je třeba získat posuzovatele. Tito posuzovatelé by měli naplňovat požadavky vztahující se k účelu posuzování, např. by měli být dobře obeznámeni s interpretací SERRJ, měli by mít zkušenost s testováním a znát populaci testovaných atd. Jejich počet by měl být adekvátní závažnosti rozhodnutí, která budou činěna na základě analýzy struktury obsahu, a faktu, že maturitní zkouška je zkouškou vysoké důležitosti.

Školení posuzovatelů

Před každým posuzováním by měli být posuzovatelé vyškoleni. Školení by mělo zahrnovat familiarizaci s kontextem zkoušky, s popisnými modely a jejich teoretickým ukotvením, s účelem metody atd. Na základě zkušeností z tohoto projektu doporučujeme školení provádět dvoukolově. Před prvním kolem posuzování doporučujeme provést familiarizační aktivity směřující k seznámení s nástrojem a metodami a také nácvik na zkušební sadě. Po tomto úvodním kroku by měla následovat diskuse a zpětná vazba. Obě následující kola posuzování by měla probíhat individuálně, mezi koly by však měla proběhnout diskuse, která by umožnila reagovat na případné nejasnosti související s metodou nebo popisnými nástroji a jejich interpretací a aplikací. Druhé kolo by bylo opět individuální posuzování a jeho výstupy by byly vyhodnoceny. Vzhledem k doporučovanému zařazení diskuse se jako ideální jeví posuzování buď zcela prezenční, nebo přinejmenším diskuse a druhé kolo by měly být prováděny v online prostředí.

Vyhodnocení a interpretace dat ze zkoumání srovnatelnosti obsahu a konstruktu

Po provedení vyhodnocení, tj. kvalitativním zpracování dat např. pomocí grafického přehledu, i kvantitativním vyhodnocení např. pomocí koeficientů procentuální shody a AC1, a při uspokojivém nálezu shodné struktury obsahu lze výsledek využít jako vstupní informaci pro specifikaci modelu pro konfirmatorní faktorovou analýzu, nebo jako pomocné interpretační schéma pro exploratorní faktorovou analýzu. Existují však i další metody (např. strukturní modelování – SEM), které lze využít.

Pokud analýza ekvivalence konstruktu prokáže konstruktovou ekvivalenci porovnávaných testových verzí, lze přistoupit k vyrovnávání. V opačném případě vyrovnávání skórů neposkytne smysluplné a validní výsledky. Důležité kroky a rozhodnutí, jež musí být učiněny v souvislosti s vyrovnáváním, diskutujeme v následujících oddílech.

Validace hranice pro výrok uspěl/neuspěl

V současné zkoušce je rozhodnutí *uspěl* vázáno na překonání hranice 33 % z maxima možných bodů. Nevíme, zda existuje studie dokumentující toto rozhodnutí, jeho obsahovou podloženost, funkčnost a důsledky. Pokud by taková zdůvodňující dokumentace neexistovala, doporučujeme zavést mechanismus pro validaci této mezní hranice úspěšnosti a prokázání, že žák, který získá 33 % bodů z testu má skutečně úroveň znalostí a dovedností odpovídající úrovni B1 a dalším požadavkům definovaných v obou výše zmíněných dokumentech k maturitní zkoušce z anglického jazyka B1. Proces stanovení standardu – hranice pro výrok *uspěl/neuspěl* může doložit vztah mezi minimálními požadavky kladenými na testovaného (maturanta) kurikulem a hranicí vyjádřenou na škále se skóry nebo procentuální úspěšností.

Pretestace: ověřování psychometrických vlastností a příprava k vyrovnávání

Pretestace jsou zásadní pro ověřování kvality testového nástroje. Pouze nástroj, který je po obsahové a psychometrické stránce zcela v pořádku, může přinášet validní a spolehlivé informace o testovaných na straně jedné, a na straně druhé může být smysluplně podroben vyrovnávání. Jsou-li pretestace prováděny na dostatečně velkém a reprezentativním vzorku, v dobře promyšleném designu, nebo je-li kromě klasické teorie testů při analýzách výsledků využívána i teorie odpovědi na položku, pak je možné identifikovat položky s neuspokojivými vlastnostmi či další problematické aspekty. Tyto problematické vlastnosti musí být vyhodnoceny, nedostatky odstraněny (formulace distraktorů, kmene položky, nahrazení problematických položek apod.) a úlohy znovu pretestovány. Vyrovnávání je smysluplné u verzí s prokazatelně dobrými psychometrickými parametry položek, s podobnou distribucí obtížnosti položek a s podobnými hodnotami reliability (Dorans, Moses, & Eignor, 2010).

V případě testů vysoké důležitosti jsou pretestace nutným nepominutelným krokem vedoucím ke zvyšování a zajišťování kvality testových nástrojů. V situaci, kdy je poskytovatel zkoušek ze zákona povinen veškeré testové materiály zveřejňovat v plném znění ihned po realizaci testování, jsou však pretestace také ideální a de facto jedinou příležitostí, kde lze uskutečnit sběr dat v designu, který by umožnil vyrovnávání testových skórů (u prokazatelně ekvivalentních testových verzí). Design sběru dat pro přípravu testových verzí k vyrovnávání je nutné předem promyslet, obvykle i na několik let dopředu.

Vycházíme-li ze současné situace slovenské maturitní zkoušky, pak víme, že všechny testové materiály musí obsahovat pouze ty části, které jsou uvedeny ve specifikacích testu, a testy musí být ihned po realizaci testování kompletně zveřejněny. Není tedy možné uvažovat např. o zařazení kotvicích položek do ostré testové verze, což by umožnilo propojit několik testových verzí. To však neplatí pro pretestace a další experimentální situace. Připustíme-li, že na základě obsahových specifikací můžeme jen těžko vytvořit verze srovnatelné i v psychometrických parametrech, je řešením zavést do pretestací používání kotvení a vhodný design (*common-item* nebo *common-person design*) a pretestovat tak, aby bylo možné na základě těchto pretestací přikročit k vyrovnávání skórů.

Platí, že pokud chceme porovnávat dvě a více testových verzí, je nutné, aby tyto dvě verze, kromě toho, že musí být obsahově a konstruktově ekvivalentní, měly něco společného. Pokud např. dva různé testy konají dvě různé skupiny studentů (jako je tomu každý rok v případě maturitní zkoušky), není možné jejich výkony ani obtížnost těchto testů porovnávat a cokoli tvrdit o jejich srovnatelnosti.

Pokud by byl design pretestací založen na využití kotvení (společných položek nebo úloh), pak by kotvicí úlohy měly dobře reprezentovat celý ověřovaný konstrukt, tzn. jejich obsah, konstrukt, obtížnost, diskriminační schopnost a testovací technika by měly být co nejpodobnější celému testu. U interního kotvení by kotvicí úlohy měly tvořit nejméně 20 % testu. Pozornost by měla být věnována umístění položek, neboť kontext, ve kterém se úlohy nacházejí, může mít vliv na výsledky.

Možností, jak realizovat pretestace a provést sběr dat pro realizaci vyrovnávání skórů, je velmi mnoho a volba designu záleží na konkrétní situaci poskytovatele testu, na metodě vyrovnávání a na účelu vyrovnávání, technických možnostech a dostupném vzorku pretestentů.

Můžeme ale shrnout, že kotvení, v podstatě v jakémkoli designu, umožňuje propojení informací o testových verzích v dlouhodobém časovém období, aniž by byli testovaní vystavení týmž úlohám (Michaelides, 2014, s. 3), a jeho zavedení je vhodným krokem na cestě k ekvivalentním skórům z různých testových verzí maturitní zkoušky. U výběru designu záleží i na tom, zda je možné použít IRT, nebo nikoli. U pretestací bez využití IRT je nutné u designu myslet na reprezentativitu vzorku testovaných vzhledem k cílové populaci. Kotvení nemusí probíhat jen přes položky společné dvěma testovým verzím, nýbrž může být realizováno i přes skupinu respondentů, tj. jedné skupině pretestentů lze zadat dvě nebo více testových verzí, (jednoskupinový design, obvykle v kombinaci se zkříženým designem).

Množství žáků testovaných u slovenské maturitní zkoušky z anglického jazyka na úrovni B1 přesahuje 20 000 ročně. Tato skutečnost a dále fakt, že jde o zkoušky vysoké důležitosti, nás vedou k doporučení, aby do standardních procesů zpracování výsledků z pretestací i z ostrého testování byly zavedeny analýzy vycházející z teorie odpovědi na položku (IRT). Tyto analýzy vhodně doplní doposud používané analýzy založené na klasické teorii testů, poskytnou další možnost nahlížet na psychometrickou kvalitu položek i testů a umožní realizovat poměrně snadno mnohá z výše doporučených procesů. Doporučujeme také zavedení IRT analýz do rutinního procesu vývoje testových verzí a přizpůsobení designu pretestací tomu, aby bylo možné pretestované úlohy kalibrovat, odhadovat jejich psychometrické parametry, neboť je to prokazatelný způsob, jak napomoci sestavování testových verzí se srovnatelnými parametry, včetně těch psychometrických. Využití IRT také umožňuje uvažovat o budování banky kalibrovaných úloh a následné sestavení testových verzí s předem známými parametry, což usnadňuje následné vyrovnávání skórů (viz oddíl 3.2.5).

Vyrovnávání testových skórů

Cílem vyrovnávání skórů je upravit skóry z testových verzí, které naplňují podmínky pro vyrovnávání tak, aby studenti se stejnou úrovní měřených schopností (jazykové způsobilosti) dosáhli skórů, které se dají interpretovat stejným způsobem, a to bez ohledu na to, jakou testovou verzi konali. Vyrovnávání skórů upravuje skóry z různých testových verzí tak, aby byly zaměnitelné, a řeší tak problém odlišné obtížnosti různých testových verzí. Pro vyrovnávání musí být naplněny určité předpoklady

(zmiňujeme je v oddíle 3.2.3) a musí být připraven dlouhodobý plán sběru dat (design), který umožní realizovat předem vybranou metodu vyrovnávání skórů.

V závislosti na tom, jaký design pro sběr dat byl zvolen (předpokládáme, že sběr dat proběhl v rámci pretestací), a na tom, jaký je účel vyrovnávání, lze uplatnit některý z mnoha postupů pro vyrovnávání: lineární nebo nelineární vyrovnávání, vyrovnávání pravých nebo pozorovaných (hrubých) skórů, využití metod založených na IRT nebo na klasické teorii testů, metody využívající kotvení přes společné položky nebo přes skupinu testovaných. Na tomto místě není možné doporučit jeden konkrétní postup; ten musí být stanoven na základě důkladné znalosti konkrétních podmínek a možností poskytovatele zkoušek.

Před uskutečněním vyrovnávání je obvykle třeba připravit data pro výpočet vyrovnávací rovnice (Dorans, Moses, & Eignor, 2010). V některých případech, zejména tehdy, pokud např. pretestace probíhá na sice rozsáhlém vzorku, avšak není předem jisté, že odpovídá cílové populaci, je třeba rozhodnout o vyřazení některých testovaných, 1. pokud svými charakteristikami neodpovídají cílové populaci; 2. v případě, že jsou použity kotvicí položky a testovaní už položky v této kotvicí sadě řešili; 3. pokud jejich chování (výsledky) neodpovídají statistickým předpokladům modelu (vykazují např. nízkou reliabilitu v IRT modelech), nebo patří mezi okrajové případy (*outliers*). Jsou-li použity kotvicí položky, je třeba ověřit jejich statistické vlastnosti, a to zda, se chovají v obou porovnávaných verzích podobně. U skórových škál, které nejsou kontinuální, je třeba vyřešit, jak budou vyrovnávané skóry reportovány, stanovit škálu pro reportování, její počátek, maximum a jednotky apod. V některých případech a u některých metod je třeba aplikovat tzv. vyhlazování distribuční křivky skórů obou verzí, pokud je plánováno využít ekvipercentilové vyrovnávání.

Livingston (2004), Kolen a Brennan (2014), Dorans, Moses a Eignor (2010) a další autoři poskytují ucelený přehled všech aspektů, které s vyrovnáváním skórů souvisejí, a shrnují kroky, které by měly být realizovány a které my jsme se zde pokusili v obecné rovině popsat. Zdůrazňují i to, že veškerá rozhodnutí související s vývojem zkoušky, pretestací, realizací, a zejména s vyrovnáváním a reportováním by měla být kriticky na závěr zhodnocena a validována.

Reportování výsledků

Pokud proběhly veškeré procesy popsané v této kapitole, včetně vyrovnávání skórů, pak to znamená, že:

- existují standardizované procesy vývoje testu (dostatečně podrobná a validovaná specifikace testu, standardizovaný proces tvorby úloh, jasně definovaný popis způsobu vyhodnocení, promyšlený design pretestací apod.);
- z pretestací je zřejmé, že testové verze mají uspokojivé psychometrické vlastnosti;
- testové verze jsou obsahově a konstruktově srovnatelné;
- testové verze byly pretestovány s využitím designu, který je propojil přes společné respondenty nebo přes kotvicí položky;
- bylo provedeno vyrovnávání testových skórů, přičemž v souvislosti s tím, jaký design byl v pretestacích uplatněn, byla využita klasická teorie testů nebo teorie odpovědi na položku, a bylo využito lineární vyrovnávání, ekvipercentilové vyrovnávání, nebo na IRT založené vyrovnávání.

Po realizaci ostrého testování je třeba reportovat výsledky testovaných. Vzhledem k tomu, že při vyrovnávání nejsou podstatné nominální hodnoty skórů (pravých nebo pozorovaných), nýbrž stejná interpretace těchto skórů ve smyslu úrovně měřeného rysu, a skóry vyrovnávané testové verze jsou reportovány pomocí jiné než hrubé škály, je obvykle nutné uživatelům výsledků zkoušek umožnit vzájemné porovnávání skórů a připravit způsob reportování vyrovnávaných skórů a poskytnout interpretační nástroj.

Sdělování výsledků – reportování vyrovnávaných skórů by mělo probíhat způsobem srozumitelným uživatelům výsledků zkoušek, zejména samotným testovaným, na jednotné, dobře vysvětlené škále. Výsledky by měly být doprovázeny popisem mechanismu vyrovnávání a vztahu této nové škály ke škále hrubých skórů (škále referenční) a ke škále reportované (odvozené od referenční testové verze). Reportovaná škála může být v procentech udávajících úspěšnost v řešení daného testu, resp. subtestu, nebo může obsahovat skóry po vyrovnání. Sdělování výsledků a informování o procesech vyrovnávání jsou důležitou součástí celého procesu vývoje testu a nedílnou součástí spravedlivého přístupu k testovaným a součástí validačního procesu.

6.3 Závěrečná reflexe

V tomto výzkumném projektu jsme se zaměřili na otázky související s problematikou vytváření srovnatelných testových verzí, a to u zkoušek vysoké důležitosti (jakou je slovenská maturitní zkouška), jejichž legislativní rámec představuje určitá omezení při vývoji srovnatelných verzí.

V úvodní teoretické části (Část I) jsme nejprve provedli teoretické vymezení problematiky, popsali jsme kontext slovenské maturitní zkoušky z anglického jazyka na úrovni B1 a v přehledu jsme nastínili, jaké postupy k dosažení srovnatelnosti testových verzí mohou vést, jaké podmínky je nutné splnit, aby tyto postupy mohly být aplikovány, a jaké metody lze v rámci zmíněných postupů využívat.

V empirické části (Část II) jsme s ohledem na stávající podmínky slovenské maturitní zkoušky vybrali postupy a metody, které by mohly být zavedeny do procesu vývoje testových verzí, aniž by byly nutné legislativní změny v realizaci testování. Tyto postupy a metody jsme aplikovali na testové verze realizované v jarním zkušebním období v letech 2012–2015 a na data z nich získaná. Provedená analýza struktury obsahu, exploratorní faktorová analýza konstruktu, zhodnocení psychometrických vlastností testu a srovnatelnosti populací ukázaly, že testové verze 2012–2015 jsou obsahově i konstruktově podobné, nikoli identické, vykazují odlišnosti v psychometrických vlastnostech a není možné jednoznačně říci, do jaké míry jsou tyto odlišnosti způsobeny rozdíly ve struktuře obsahu a konstruktu, v různé obtížnosti testových verzí nebo v různém rozložení měřeného rysu (jazykové způsobilosti) v populacích.

Použité postupy a metody, respektive jejich vyhodnocení, poskytly velmi komplexní pohled na testové verze. Realizace metod a identifikace problémů, které se v průběhu objevily, naznačily jejich slabá místa. To nám umožnilo navrhnout možná vylepšení (pro realizaci školení posuzovatelů pro analýzu struktury obsahu, pro design pretestací a identifikaci problematických položek apod.). Postupy a metody, které popisujeme v aplikační části (Část III), se ukázaly jako přínosné a mají potenciál přispět ke zvýšení kvality testových nástrojů v podmínkách slovenské maturitní zkoušky, aniž by jejich implementace kladla jakékoli nároky na změnu legislativních podmínek pro konání maturitní zkoušky. Jednoznačně ukázaly, jak důležitá je existence podrobných specifikací a důkladné vymezení konstruktu, jakou roli mohou hrát pretestace při identifikaci problematických položek, které mohou znesnadňovat realizaci procesů vedoucích k vyrovnávání skórů.

V průběhu výzkumu jsme si kladli i otázky, které jsme nemohli přímo řešit, ale mohly by být předmětem následného zkoumání. Například se objevila otázka, zda a případně jak je skutečně možné smysluplně dělit jazykovou kompetenci na jednoznačně oddělitelné konstrukty nazývané dovednostmi (poslech, čtení, gramatika); zda lze řečové dovednosti[47] poslech, čtení popsat pomocí diskrétních kategorií – deskriptorů, a další.

Uvědomujeme si, že prováděný výzkum byl zatížen řadou omezení vyplývajících zejména z neexperimentálního designu výzkumu, se kterými jsme se museli vyrovnávat při jeho realizaci, a že jsme se ne vždy dokázali vyhnout subjektivitě při některých prováděných procesech a u rozhodnutí, která jsme museli činit v průběhu výzkumu.

Neprovedli jsme vyrovnávání skórů u porovnávaných testových verzí, a to hned z několika důvodů. Kromě časového omezení i proto, že si uvědomujeme, že nedisponujeme tolika znalostmi a technicky zaměřenými zkušenostmi, abychom byli schopni bez spolupráce s dalšími odborníky vyrovnávání čtyř testových verzí provést. Také by to znamenalo pravděpodobně provést rozsáhlý sběr dat a připravit datové soubory vhodné k vyrovnávání. To vše bylo již od počátku projektu mimo naše možnosti. Další důvod souvisí s povahou dat. Již po prvotním pohledu na dokumentaci zkoušky, zejména na specifikace zkoušek, a posléze po realizaci analýzy struktury obsahu a konstruktové ekvivalence se objevily zjevné pochybnosti o srovnatelnosti testových verzí, což je zásadní podmínka pro realizaci vyrovnávání.

Domníváme se však, že i přesto náš výzkum, realizovaný v reálných, neexperimentálních podmínkách s využitím kombinace mnoha metod, může nabídnout komplexní a užitečný vhled do široké problematiky ekvivalence testových verzí a skórů, poskytnout návod a informační zdroje při provádění spravedlivého testování a při budování validačního argumentu, což může v důsledku napomoci zvyšování kvality testových nástrojů, k jejich validnímu a spravedlivému využívání a interpretaci.

47 V CEFR Companion Volume (2017) se pracuje s řečovými činnostmi recepce, produkce, interakce a mediace, přičemž např. recepce se dále dělí na řečové činnosti – poslech nahráveka médií apod.

Summary

The research presented in this book focuses on the test versions equivalence of high-stakes tests. We perceive it as the key aspect of validity of test results and a necessary condition for meaningful and fair interpretation and use of the results. Thus, pursuing and achieving test versions equivalence and documenting the processes are two key steps on the way towards fair testing.

The project focused on the Slovak upper-secondary school leaving examination in English at B1 level, specifically on the tests of receptive skills used in the spring sessions between 2012 and 2015. It investigated what methods and procedures could be implemented to develop equivalent test versions and to prove their equivalence.

The book is divided into three parts: theoretical Part I (Chapters 1–3), empirical Part II (Chapters 4 and 5), and Part III with proposals for implementation and with discussion (Chapter 6).

The rationale of the project, its context and three research questions are presented in the Introduction and in Chapter 1. Chapter 2 introduces the context of the Slovak upper-secondary school leaving examination. Chapter 3 attempts to clarify the key concepts, presents the state of the art related to the test versions equivalence investigation, and introduces methods and procedures used in the research related to this topic. The focus of this chapter is on surveying what methods and procedures for the development of equivalent test versions exist and in what phase and under what conditions are implemented in the test development cycle.

Chapters 4 and 5 are chapters dedicated to methodology. They describe practical application of the selected methods, discuss the rationale for their selection, and attempt to find the answer to the research questions 1 and 2 (*To what extent are the test versions of the Slovak upper-secondary school leaving examination in English B1 equivalent*? and *Are the methods used in this project conclusive, reliable and practical*

enough for the use in the real Slovak context?). First, the content analysis method using descriptive models and expert judgement to analyse content equivalence is presented. Next, the use of exploratory factor analysis for construct equivalence analysis is explained. Finally, psychometric characteristics of the four test versions and populations of test takers were compared. The selected methods provided deep insight into the degree of comparability of the test versions used between 2012 and 2015 and proved to be useful tools for this kind of investigation. They also made it possible to identify problem areas in terms of quality of the test versions used in this research.

Chapter 6 answers the research question 3 (*What methods and procedures could be implemented in the test development process in the current context of the Slovak exams to achieve test versions equivalence without any changes in the legislation?)*. The answers and suggestions are based on the results and findings from the theoretical and empirical parts (Part I and Part II), and on thorough discussion of the problematic aspects of the examinations and limitations of the research. We were able to identify weak points in the test development process and also in the methods we used and to propose steps that would lead to improvements. Key areas include test specifications, construct definition, routine implementation of content and construct analyses, well prepared pretesting design and longitudinal plan for pretesting, testing and post--test analyses.

We admit that the research was constrained by several limitations, especially by its strictly non-experimental design and related challenges, subjectivity in human judgements typical for some methods and also by subjectivity involved in the researcher´s decisions. Though, we think this book can provide a complex picture of the real situation of high-stakes exams, point out critical aspects of the test development and test results interpretations, increase awareness about validity, and outline possible ways to achieve test versions equivalence, building validity argument, and especially, to fair testing.

Seznam literatury

AERA, APA, & NCME (1999, 2014). *Standards for Educational and Psychological Testing.*

Alderson, J. C. (1993). Judgements in language testing. In D. Douglas & C. Chapelle (Eds.), *A New Decade of Language Testing: Collaboration and cooperation* (s. 46–57). Ann Arbor, MI: University of Michigan.

Alderson, Ch. J., Figueras, N, Kuijper, H., Nold, G., Takala, & S. Tardieu, C. (2006). Analysing test of reading and listening in relation to the Common European Framework of Reference: The experience of the Dutch CEFR Construct Project. *Assessment Quarterly, 3*(1), 3–30. https://doi.org/10.1207/s15434311laq0301_2

Anýžová, P. (2013). Ekvivalence položek v mezinárodních datech: základní vymezení a možnosti analýzy. *Data a výzkum* – SDA *Info 2013, 7*(1), 29–56. https://doi.org/10.13060/1802-8152.2013.7.1.2

Bachman, L. F., Davidson, F., Ryan, K., & Choi, I. C. (1995). *An Investigation into the Comparability of Two Tests of English as a Foreign Language: The Cambridge-TOEFL Comparability Study.* Cambridge: Cambridge University Press.

Bachman, L. F. (2004). *Statistical Analyses for Language Assessment.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511667350

Bachman, L. F. (2005). Building and supporting a case for test use. *Language Assessment Quarterly*, 2(1), 1–34. https://doi.org/10.1207/s15434311laq0201_1

Bachman, L. F. (1990). *Fundamental Consideration in Language Testing*. Oxford: Oxford University Press.

Bachman, L. F. (2012). Justifying the use of language assessments: linking interpretations with consequences. Conference paper. Dostupné z: http://www.sti.chula.ac.th /conference

Bachman, L. F., & Palmer, A. S. (1996). *Language Testing in Practice: Designing and Developing Useful Language Tests*. Oxford: Oxford University Press.

Bachman, L. F., Davidson. F., & Milanovic, M. (1996). The use of test method characteristics in the content analysis and design of EFL proficiency tests. *Language Testing, 13*(2), 125–150. https://doi.org/10.1177/026553229601300201

Bachman, L. F., & Cohen, A. D. (1998). Language testing-SLA interfaces: An update. In L. F. Bachman & A. D. Cohen (Eds.) *Interfaces between SLA and Language Testing Research* (s. 1–31). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524711.003

Bachman, L. F., & Palmer, A. S. (2010). *Language Assessment in Practice: Developing Language Tests and Justifying Their Use in the Real World.* Oxford: Oxford University Press.

Baghaei, P. (2010). Test score equating and fairness in language assessment. *Journal of English Language Studies, 1*(3), 113–128.

Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. *Journal of English for Academic Purposes, 24*, 1–13. https://doi.org/10.1016/j.jeap.2016.07.004

Bialosiewicz, A., Murphy, K., & Berry, T. (2013). An Introduction to Measurement Invariance Testing: Resource Packet for Participants. CEC. Dostupné z: http://comm.eval.org/HigherLogic/System/DownloadDocumentFile.ashx?DocumentFileKey=63758fed-a490-43f2-8862-2de0217a08b8

Brown, J. D., & Hudson, T. (2002). *Criterion-Referenced Language Testing.* Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139524803

Cicchetti, D. V., & Feinstein, A. R. (1990). High agreement but low kappa: II. Resolving the paradoxes. *Journal of Clinical Epidemiology, 43*(6), 551–558. https://doi.org/10.1016/0895-4356(90)90159-M

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*.

Council of Europe. (2011). *Manual for Language Test Development and Examining.* Dostupné z: http://www.coe.int/t/dg4/linguistic/ManualLanguageTest-Alte2011_EN.pdf

Costa-Santos, C., Bernardes, J., Ayres-de-Campos, D., Costa, A., & Costa, C. (2011). The limits of agreement and the intraclass correlation coefficient may be inconsistent in the interpretation of agreement. *Journal of Clinical Epidemiology, 64*, 264–269. https://doi.org/10.1016/j.jclinepi.2009.11.010

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281–306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J. (2004). Equating, concordance, and expectation. *Applied Psychological Measurement, 28*(4), 227–246. https://doi.org/10.1177/0146621604265031

Dorans, N. J., Moses, T. P., & Eignor, D. R. (2010). Principles and practices of test score equating. *ETS RR-10-29*. Dostupné z: https://www.ets.org/Media/Research/pdf/RR-10-29.pdf. https://doi.org/10.1002/j.2333-8504.2010.tb02236.x

Feuer, M. J., Holland, P. W., Green, B. F., Bertenthal, M. W., & Hemphill, F. C. (Eds.). (1999). *Uncommon measures: Equivalence and linkage among educational tests* (Report of the Committee on Equivalency and Linkage of Educational Tests, National Research Council). Washington DC: National Academy Press.

Field, J. (2009). *Listening in the Language Classroom.* Cambridge: Cambridge University. https://doi.org/10.1017/CBO9780511575945

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76*(5), 378–382. https://doi.org/10.1037/h0031619

Geranpayeh, A. (1994). Are score comparisons across language proficiency test batteries justified? An IELTS-TOEFL comparability study. *Edinburgh Working Paper in Applied Linguistics, 5*, 50–65.

Geranpayeh, A., & Taylor, L. (2013). *Examining Listening: Research and Practice in Assessing Second Language Listening.* Studies in Language Testing 35. Cambridge: UCLES/ CUP

Goh, Ch. C. M., & Aryadoust, V. (2015). Examining the notion of listening subskill divisibility and its implications for second language listening. *International Journal of Listening*, *29*(3), 109–133. https://doi.org/10.1080/10904018.2014.936119

Gwet, K. L. (2002). Inter-rater reliability: Dependency on trait prevalence and marginal homogeneity. *Statistical Methods For Inter-Rater Reliability Assessment, 2*. Dostupné z: www.agreestat.com

Gwet, K. L. (2008a). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psycholgy*, *61*, 29–48. https://doi.org/10.1348/000711006X126600

Gwet, K. L. (2008b). Variance estimation of nominal-scale inter-rater reliability with random selection of raters. *Psychometrika*, *73*, 407–430. https://doi.org/10.1007/s11336-007-9054-8

Gwet, K. L. (2011). On the Krippendorff's alpha coefficient. Dostupné z: www.agreestat.com

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Gaithersburg: Advanced Analytics.

Gwet, K. L. (2015). Standard error of Krippendorff's alpha coefficient. Dostupné z: http://inter-rater-reliability.blogspot.de/2015/08/standard-errorof-krippendorffs-alpha.html

Gwet, K. L. (2016). Testing the difference of correlated agreement coefficients for statistical significance. *Educational and Psychological Measurement, 76,* 609–637. https://doi.org/10.1177/0013164415596420

Harman, H. H. (1976). *Modern Factor Analysis.* (3rd ed.). Chicago: University of Chicago Press.

Haupt, G., Koch, E. (2012). The argument for evaluating language tests for equivalence across language groups. *Southern African Linguistics and Language Studies, 30*(1), 65–76. https://doi.org/10.2989/16073614.2012.693715

Hendl, J. 2009. *Přehled statistických metod: analýza a metaanalýza dat*. Praha: Portál.

Holland, P. W. (2007). A framework and history for score linking. In N. J. Dorans, M. Pommerich, & P. W. Holland (Eds.), *Linking and Aligning Scores and Scales* (s. 5–30). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-0-387-49771-6_2

Holland P. W., & Dorans, N. J. (2006). Linking and equating. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., s. 187–220). Westport, CT: Praeger.

Chen, F., Huang, X., & MacGregor, D. (2009). Equating or linking: basic concepts and a case study. Presentation originally presented at CAL, Washington. Dostupné z: https://fliphtml5.com/xrgx/bfuj/basic

Choi, I., Sung, K, & Boo, J. (2003). Comparability of a paper-based language tests and a computer-based language test. *Language Testing, 20*(3), 295–320. https://doi.org/10.1191/0265532203lt258oa

Chráska, M. (2007). *Metody pedagogického výzkumu. Základy kvantitativního výzkumu*. Praha: Grada

Chvál, M., Straková, J., & Procházková, I. (2015). *Hodnocení výsledků vzdělávání didaktickými testy.* Praha: Česká školní inspekce.

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

Kolen, M. J., & Brennan, R. L. (2004, 2014). *Test equating, scaling, and linking: Methods and practice*. New York: Springer. https://doi.org/10.1007/978-1-4757-4310-4

Kirkebøen, G. (2009). Decision behaviour – Improving expert judgement. Dostupné z: http://www.concept.ntnu.no/attachments/058_Kirkebooen%20%20-%20Expert%20judgement.pdf

Khalifa, H., Weir, C. (2009). *Examining reading.* Cambridge: Cambridge University Press.

Klein, D. (2018). Implementing a general framework for assessing interrater agreement in Stata. *The Stata Journal*, *18*(4), 871–901. https://doi.org/10.1177/1536867X1801800408

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, linking, and scaling: Methods and practices* (2nd ed.). New York, NY: Springer-Verlag. https://doi.org/10.1007/978-1-4757-4310-4

Kottner, J., & Streiner, D. L. (2011). The difference between reliability and agreement. *Journal of Clinical Epidemiology*, *64*(6), 701–702. https://doi.org/10.1016/j.jclinepi.2010.12.001

Krippendorff, K. (2004). *Content analysis. An introduction to its methodology.* Thousand Oaks, CA: Sage Publications, Inc.
Kunnan, A. J., & Carr, N. (2017). A comparability study between the general English proficiency test-advanced and the internet-based test of English as a foreign language. *Language Testing in Asia, 7*(17). https://doi.org/10.1186/s40468-017-0048-x
Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33*, 159–174. https://doi.org/10.2307/2529310
Lavrakas, P. J. (Ed.). (2008). *Encyclopedia of survey research methods,* Thousand Oaks, CA: Sage Publications, Inc. https://doi.org/10.4135/9781412963947
Linn, R. L. (1993). Linking results of distinct assessments. *Applied Measurement in Education, 6(1)*, 83–102. https://doi.org/10.1207/s15324818ame0601_5
Livingston, S. A. (2004). *Test score equating (without IRT)*. Educational Testing Service. Dostupné z: www.ets.org
Lumley, T. (1993). Reading comprehension sub-skills: teachers' perceptions of content in an EAP test. *Melbourne Papers in Language Testing, 2*(1), 24–57.
McCray, G. (2013). Assessing inter-rater agreement for nominal judgement variables. Paper presented at the *Language Testing Forum. Nottingham, November 15-17*, 2013. Dostupné z: www.agreestat.com
Messick, S. (1987). Validity. *ETS Research Report Series*. https://doi.org/10.1002/j.2330-8516.1987.tb00244.x
Messick, S. (1993). Foundations of validity: meaning and consequences in psychological assessment. *ETS Research Report Series.* https://doi.org/10.1002/j.2333-8504.1993.tb01562.x
Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performance as scientific inquiry into score meaning. American Psychologist*, 50*(9). 741–749. https://doi.org/10.1037/0003-066X.50.9.741
Michaelides, M. P., & Haertel, E. H. (2014). Selection of common items as an unrecognized source of variability in test equating: a bootstrap approximation assuming random sampling of common items. *Applied Measurement in Education, 27(1)*, 46–57. https://doi.org/10.1080/08957347.2013.853069
Mislevy, R. J. (1992). *Linking educational assessments: Concepts, issues, methods, and prospects (Policy Information Rep.)*. Princeton, NJ: ETS.
O'Loughlin, K. (1997). *The Comparability of Direct and Semi-direct Speaking Tests: A case study*. Unpublished Ph.D. thesis. Melbourne: University of Melbourne. Dostupné z: https://minerva-access.unimelb.edu.au/handle/11343/38817

Pommerich, M., Hanson, B. A., Harris, D. J., & Sconing, J. A. (2004). Issues in conducting linkages between distinct tests. *Applied Psychological Measurement, 28(4)*, 247–273. https://doi.org/10.1177/0146621604265033

Popham, W. J. (1978). *Criterion-referenced measurement*. Englewood Cliffs, NJ: Prentice Hall.

Purpura, J. E. (2004). *Assessing grammar*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511733086

Purpura, J. E. (2014a). Assessing grammar. In A. J. Kunnan (Ed.). *Companion to language assessment* (s. 100–124). Oxford: Wiley. https://doi.org/10.1002/9781118411360.wbcla147

Purpura, J. E. (2014b). Cognition in language assessment. In A. J. Kunnan (Ed.), *Companion to language assessment* (s. 1452–1476). Oxford: Wiley. https://doi.org/10.1002/9781118411360.wbcla150

Purpura J. E. (2017) Assessing meaning. In E. Shohamy, I. Or., & S. May (Eds.), *Language Testing and Assessment. Encyclopedia of Language and Education* (s. 33–61). (3rd ed.). Cham: Springer. https://doi.org/10.1007/978-3-319-02261-1_1

Sawaki, Y. (2001). Comparability of conventional and computerized tests of reading in a second language. *Language Learning Technology 5,* 38–59.

Sireci, S., &Allalouf, A. (2003). Appraising item equivalence across multiple languages and cultures. *Language Testing 20*(2), 148–166. https://doi.org/10.1191/0265532203lt249oa

Spolsky, B. (1995). *Measured Words: The Development of Objective Language Testing.* Oxford: Oxford University Press.

Thompson, W. D., & Walter, S. D. (1988). A reappraisal of the kappa coefficient. *Journal of Clinical Epidemiology, 41*(10), 949–958. https://doi.org/10.1016/0895-4356(88)90031-5

Urbina. S. (2004). *Essentials of Psychological Testing*. New Jersey: John Wiley & Sons, Inc.

van den Heuvel-Hanhuizen, M., Robitzsch, A., Treffers, A., & Köller, O. (2009). Large-scale assessment of change in student achievement: Dutch primary school students' results on written division in 1997 and 2004 as an example. *Psychometrika, 74*(2), 351–365. https://doi.org/10.1007/s11336-009-9110-7

van de Vijver, F. & Poortinga, Y. (2005). Conceptual and methodological issues in adapting tests. In R. K. Hambleton, P. F. Merenda, & C. D. Spielberger (Eds.), *Adapting Educational and Psychological Tests for Cross-Cultural Assessment* (s. 39–63). Mahwah, NJ: IEA Lawrence Erlbaum Associates, Publishers.

Vandenberg, R. J., & Lance, C. E. (2000). A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. *Organizational Research Methods, 3*, 4–70. https://doi.org/10.1177/109442810031002

Verhelst N. D., & Glas C. A. W. (1995). The one parameter logistic model. In *Fischer G. H., Molenaar I. W. (Eds.), Rasch Models*. New York, NY: Springer. https://doi.org/10.1007/978-1-4612-4230-7_12

von Davier. A. A. (2011) A statistical perspective on equating test scores. In A. A. von Davier (Ed.). *Statistical models for test equating, scaling, and linking* (s. 1–17). New York, NY: Springer. https://doi.org/10.1007/978-0-387-98138-3_1

Watson, J. C. (2017) Establishing evidence for internal structure using exploratory factor analysis. *Measurement and Evaluation in Counseling and Development, 50(4)*, 232–238. https://doi.org/10.1080/07481756.2017.1336931

Weir, C. J. (2005). Limitations of the Common European Framework for developing comparable examinations and tests. *Language Testing*, *22*, 281–300. https://doi.org/10.1191/0265532205lt309oa

Weir, C. (2005). *Language Testing and Validation. An Evidence-based Approach.* Basingstoke: Palgrave, MacMillan. https://doi.org/10.1057/9780230514577

Weir, C, Wu, R. (2006). Establishing test form and individual task comparability: A case study of a semi-direct speaking test. *Language Testing, 23*, 167–197. https://doi.org/10.1191/0265532206lt326oa

Widaman, K. F., Ferrer, E., & Conger, R. D. (2010). Factorial Invariance within Longitudinal Structural Equation Models: Measuring the Same Construct across Time. *Child development perspectives, 4*(1), 10–18. https://doi.org/10.1111/j.1750-8606.2009.00110.x

Wu, R. Y. (2014). *Validating Second Language Reading Examinations Establishing the validity of the GEPT through alignment with the Common European Framework of Reference.* Studies in Language Testing. Cambridge: Cambridge University Press.

Příloha: Popisné modely

Popisný model pro POSLECH

SERRJ deskriptory	Položka ověřuje (původní deskriptory)	Položka ověřuje (sloučené kategorie)
A	zachycení nekomplikované faktografické (konkrétní) informace	A _Zachycení informace
B	porozumění podrobným orientačním pokynům nebo jednoduchým technickým informacím	*x (typ textu a položky ověřující tento deskriptor nebyly zastoupeny)*
C	porozumění hlavním bodům=důležitým informacím textu/nahrávky o známých (běžných) záležitostech	C _Práce s informacemi
E	sledování s porozuměním delší nahrávky a pochopení hlavní linie textu	
D	pochopení smyslu/hlavní myšlenky (z určité části textu) textu	DEF_Interpretace textu, porozumění myšlenkám textu
F	sledování s porozuměním delší nahrávku a pochopení hlavní myšlenku/y textu *(přeformulováno)*	
Kognitivní procesy	Aktivuje se	Aktivuje se
SP	selektivní poslech	selektivní poslech
SG	kombinace selektivního a globálního poslechu	kombinace selektivního a globálního poslechu
GP	globální poslech	globální poslech
RP	responzivní poslech	*Nebyl pozorován*

Popisný model pro GRAMATIKU

Popisné modely pro Gramatiku nebyly měněny, kategorie nebyly slučovány.

SERRJ deskriptory	Položka ověřuje	Položka ověřuje
A	slovotvorba/morfologie	slovotvorba/morfologie
B	lexikum, frazeologie	lexikum, frazeologie
C	morfosyntax (deklinace, konjugace)	morfosyntax (deklinace, konjugace)
E	syntax (včetně prostředků textové návaznosti na úrovni věty a nadvětných celků)	syntax (včetně prostředků textové návaznosti na úrovni věty a nadvětných celků)
D	pragmatické významy (vztahy, vyvozování významů, postoje…), jazykové funkce	pragmatické významy (vztahy, vyvozování významů, postoje…), jazykové funkce
Kognitivní procesy	Aktivuje se	Aktivuje se
H	znalost gramatické formy	znalost gramatické formy
I	znalost gramatického významu	znalost gramatického významu
J	znalost pragmatického významu	znalost pragmatického významu

Popisný model pro ČTENÍ

SERRJ deskriptory	Položka ověřuje (původní deskriptory)	Položka ověřuje (sloučené kategorie)
B	nalezení a porozumění relevantní informaci v každod. materiálech (dopisy, brožury, krátké úřední dokumenty apod.)	B _Vyhledání informace ve standardizovaném textu
C	„rychlé přehlédnutí“ delšího textu a vyhledání požadované informace	
D	„rychlé přehlédnutí“ delšího textu a získání požadované informace z různých částí textu	CDE_Vyhledání informace v delším textu
E	předpokládá „rychlé přehlédnutí“ delšího textu a získání požadované informace z různých textů	
F	rozpoznání hlavních závěrů v jasně signalizovaném argumentativním textu	
G	rozpoznání linie argumentace ve zpracování předkládaného problému (ač ne vždy do detailu)	
H	rozpoznání významných/důležitých bodů v přímočarém textu (např. článku) na známé téma	FGHIA_Globální porozumění textu nebo části textu
I	rozpoznánívýznamných myšlenek v přímočwarém textu (např. článku) na známé téma	
A	porozumění popisu událostí, pocitů a přání v osobních dopisech (přeformulace typu textu)	
J	porozumění jasně napsaným přímočarým instrukcím (k nějakému přístroji)	x (typ textu a položky ověřující tento deskriptor nebyly zastoupeny)
Kognitivní procesy	Aktivuje se	Aktivuje se
P	pozorné čtení – lokální úroveň	P_Pozorné, podrobné čtení
	pozorné čtení – globální úroveň	
R	rychlé čtení – lokální – scanning	R_rychlé čtení – lokální
	rychlé čtení – lokální – skimming	
	rychlé čtení – globální – skimming	
	rychlé čtení – globální – search reading	

Seznam tabulek

Seznam obrázků

Srovnatelnost testových verzí slovenské maturitní zkoušky z anglického jazyka

Mgr. Martina Hulešová, M.A., Ph.D.

Ediční řada: Cizí jazyky a jejich didaktiky: teorie, empirie, praxe
Svazek 10

Vydala Masarykova univerzita, Žerotínovo nám. 617/9, 601 77 Brno
Jazykové korektury Ondřej Pechník
Grafický návrh obálky Jana Nedomová
Sazba Tereza Češková
1., elektronické vydání, 2021

ISBN 978-80-210-9950-0

Publikace prezentuje výsledky výzkumu, který se zabýval otázkou, jakými metodami je možné dosahovat srovnatelnosti testových verzí ve zkouškách vysoké důležitosti, a tím i srovnatelnosti a spravedlnosti při interpretaci výsledků těchto zkoušek. Výzkum byl realizován na didaktických testech použitých ve slovenské maturitní zkoušce z anglického jazyka na úrovni B1 v testech receptivních dovedností realizovaných v jarních termínech 2012–2015. Nejprve je vymezen kontext slovenské maturitní zkoušky a jsou představeny postupy a metody používané při studiích srovnatelnosti testových verzí v různých oblastech testování a hodnocení. Dále publikace popisuje aplikaci vybraných postupů a metod na testové verze maturitní zkoušky z anglického jazyka na úrovni B1 a na reálná data z nich získaná. Zjištění poukazují na potenciál použitých postupů a na to, že jejich využití by mohlo přispět ke zvýšení kvality testových nástrojů v podmínkách slovenské maturitní zkoušky, aniž by jejich implementace kladla nároky na změnu legislativních podmínek pro konání maturitní zkoušky. Z výsledků dále vyplývá, že tyto postupy nejenže mohou upozornit na problematické oblasti při vývoji testových verzí, nýbrž také vedou k dosažení srovnatelnosti testových verzí, a tím ke srovnatelné, spravedlivé a validní interpretaci výsledků žáků konajících různé testové verze maturitní zkoušky.

MUNI PRESS
MUNI PED