Monitoring kvality Certifikované zkoušky z češtiny pro cizince

Quality Monitoring of the Czech Language Certificated Exam

Martina Hulešová

Abstrakt: Certifikovaná zkouška z češtiny pro cizince (CCE) je připravována v souladu s principy dobré praxe v testování a se standardy asociace Association of Language Testers in Europe (ALTE), jejímž je ÚJOP UK jako organizace poskytující certifikované zkoušky z češtiny plnoprávným členem. Z tenze mezi reálnou situací a snahou o naplňování principů spravedlivosti, zodpovědnosti a dobré praxe v testování vznikl dlouhodobý projekt interního monitoringu kvality zkoušek, jenž usiluje o nalezení cesty, jak v rámci existujících podmínek a možností poskytovat zkoušky odpovídající standardům pro mezinárodní testování. V textu nastíním, v jakých mantinelech zkouška vzniká, jakými způsoby její tvůrci zajišťují kvalitu a jak tvůrci zkoušky řeší průběžné stanovování hraničního skóre modifikovanou metodou Direct Consensus a sestavování ekvivalentních testových verzí k předem danému hraničnímu skóru. V závěru budou naznačeny další cesty, kudy se bude monitoring kvality zkoušek CCE ubírat.

Klíčová slova: stanovení standardu/hraničního skóru, přímá shoda, ekvivalentní verze, expertní posouzení

Abstract: The Czech Language Certificate Exam has been prepared in line with the ALTE (Association of Language Testers in Europe) principles of good practice. A long-term project of monitoring the quality of the exams has its origin in the effort to meet the principles of fairness and best practice in language testing. This article presents the situation and constraints of the exam development and the ways in which the test developers pursue its quality using methods that enable test authors to assemble equivalent test versions.

Key words: standard setting, direct consensus, equivalent forms, expert judgement

1 O zkoušce CCE

Certifikovaná zkouška z češtiny pro cizince (CCE)[1] byla vyvinuta Výzkumným a testovacím centrem Ústavu jazykové a odborné přípravy Univerzity Karlovy v Praze (VTC ÚJOP UK) mezi léty 2005–2006. Jedná se o mezinárodně uznávanou zkoušku z českého jazyka pro nerodilé mluvčíurčenou kandidátům starším 16 let, kteří si chtějí nebo potřebují ověřit úroveň komunikační kompetence v češtině. Zkoušku uznává mnoho českých vysokých škol a někteří zaměstnavatelé u nás i v zahraničí.

Certifikované zkoušky z češtiny pro cizince se vztahují celkem k pěti úrovním definovaným Společným evropským referenčním rámcem pro jazyky (dále SERRJ),

[1] Podrobně jsou zkoušky popsány na http://ujop.cuni.cz/cce.

a sice A1, A2, B1, B2 a C1. Zkouška prochází pravidelným auditem kvality ALTE (v pětiletých obdobích), který je podmínkou pro plnoprávné členství v ALTE.

Vývoj zkoušky a testových verzí reflektuje principy dobré praxe a standardy; jak ty, které jsou nutné pro úspěšný audit ALTE, tak obecně akceptované standardy v testování, např. Standardy pro pedagogické a psychologické testování (AERA, APA & NCME, 1999). Zároveň je však třeba brát v úvahu lokální kontext a relativní „velikost" zkoušky, zejména počet kandidátů a finanční a lidské zdroje. Z tohoto pohledu má CCE poměrně obtížnou situaci, protože škála možností, jak zkoušky validovat a jak naplnit standardy, je oproti jiným organizacím typu Cambridge ESOL, Goethe Institut, CIEP apod. nesrovnatelně menší. Na druhou stranu to přináší i motivaci nacházet způsoby, jak v daném kontextu a podmínkách dospět k řešení, které by umožnilo vytváření a poskytování kvalitního testovacího nástroje naplňujícího mezinárodní standardy.

Z výše zmíněné potřeby vznikl i dlouhodobý projekt interního monitoringu kvality zkoušek, jehož součástí je i otázka stanovení hraničního skóru a s tím související srovnatelnost testových verzí.

1.1 Srovnatelnost – (nejen) terminologický problém

Existují přinejmenším čtyři termíny, které se vztahují k míře porovnatelnosti nebo zaměnitelnosti verzí téhož testu. Ve snaze používat terminologii systematicky zde pracujeme s vymezením terminologie tak, jak ji poskytují standardy pro pedagogické a psychologické testování (AERA, APA & NCME, 1999). Ty rozlišují verze alternativní, srovnatelné, ekvivalentní a paralelní.

Alternativní verze (alternate forms) je jakýsi nadřazený termín zahrnující i tři pojmy uvedené níže. Jde o označení verzí zaměnitelných, které měří stejné konstrukty stejným způsobem a jsou administrovány za stejných podmínek. **Srovnatelné verze** jsou kategorie s nejnižšími nároky na míru shody, kdy si jsou verze vzájemně podobné obsahem, neobsahují stejné, tj. kotvicí položky (anchor items) a není u nich prokázaná shoda v psychometrických parametrech. **Ekvivalentní verze** měří stejný konstrukt za použití stejných testovacích technik, nevykazují však číselnou shodu v hrubých skórech. **Paralelní verze** měří stejný konstrukt stejnými testovacími technikami, mají shodný průměrný hrubý skór, směrodatnou odchylku, shodnou strukturu chyby měření a stejnou korelaci s jinými měřeními. Vzhledem k tomu, že je v praxi velmi obtížné je vytvořit, je možné je sestavit / připravit z ekvivalentních verzí ex-post tzv. *vyrovnáváním* (equating), tj. převedením hrubých skórů ze dvou či více verzí testu na společnou škálu nebo díky využití postupů založených na teorii odpovědi na položku a banky úloh s kalibrovanými úlohami; to vše za podmínky, že jsou verze prokazatelně strukturně ekvivalentní, vytvořené podle shodných specifikací, jsou podobně obtížné, administrují se za stejných podmínek.

2 Stanovení hraničního skóru v rámci projektu *Monitoring kvality Certifikované zkoušky z češtiny pro cizince*

Hraniční skór pro CCE byl původně stanoven na základě tradice českých vysokých škol, jiných zkoušek uznávaných MŠMT ČR a na základě doporučení a/nebo požadavků uživatelů zkoušky. Kandidáti museli z každého subtestu získat nejméně 60 %, přičemž nejvýše v jednom subtestu písemné části mohli získat i jen 50 %. Snaha o zvyšování kvality CCE vedla mimo jiné i k přehodnocení pohledu na cut-off skór a ke snaze o jeho stanovení teoreticky i empiricky podloženým způsobem. Řádné a zdokumentované stanovení hraničního skóru (dále standard setting), v případě CCE též související s přiřazením zkoušek k referenčním úrovním SERRJ, totiž představuje důležitý krok z pohledu validace zkoušky.

Standard setting, realizovaný v červnu 2014 a považovaný za pilotáž postupů a metody, poukázal jak na pozitivní aspekty (vhodnost zvolené metody pro daný kontext, vhodně zvolený způsob vyhodnocení a interpretace dat, užitečnost školicích aktivit i panelisty[2] definovaného konceptu *minimálně kompetentního kandidáta* – MKK – dané úrovně dle SERRJ), tak i na problematičnost některých aspektů původně plánovaného postupu (časová, personální i finanční náročnost stanovování standardu pro celé sestavené testové varianty, obtíže se způsobem, jak vysvětlit uživatelům výsledků zkoušek (stakeholders) to, proč by měl být aplikován tzv. pohyblivý standard – tedy takový, který by se mohl lišit jak pro různé zkoušky i jejich jednotlivé části, tak pro různé testové verze zkoušky stejné úrovně).

Z uvedených důvodů a také proto, že v budoucnu lze předpokládat revize specifikací zkoušek, jež by mohly ovlivnit podobu testů, bylo rozhodnuto zachovat cut-off skór ve stávající podobě, tedy 60 % pro každý subtest[3], a vydat se opačným směrem, tedy **najít teoreticky i empiricky podložený způsob, jak sestavovat testy tak, aby 60 % bodů odpovídalo tomu, co by měl zvládnout MKK**.

2.1 Výchozí stav

Zkoušky CCE vycházejí z jedné koncepce, testové verze zkoušky v každé z pěti úrovní jsou připravovány podle identických specifikací a specifikační tabulky. Ověřovaný konstrukt je tedy operacionalizován standardizovaným způsobem, zkoušky jsou stejně administrovány a vyhodnocovány. Prozatím však neexistuje studie[4], která by prokazovala, že shodné specifikace a standardizace v oblasti tvorby položek, administrace a vyhodnocení jsou dostatečnou zárukou toho, že lze testové verze považovat za paralelní či ekvivalentní (viz výše).

[2] Člen panelu – skupiny účastníků procesu stanovení hraničního skóru

[3] Byla však zrušena výjimka dosažení alespoň 50 % cut-off skóru pro nejvýše jeden ze subtestů.

[4] Přinejmenším v českém kontextu testování o takové studii nevíme.

Zároveň byl dán cut-off skór, shodný pro všechny úrovně, resp. testové verze, a to na 60 % z maxima pro každý subtest, navíc s povolenou výjimkou minima 50 % v jednom ze subtestů písemné zkoušky.

V roce 2013/2014 VTC ÚJOP UK vyvolalo interní diskusi o revizi a validaci stanoveného cut-off skóru. Úvahy vedoucí k tomuto rozhodnutí by se daly shrnout do otázky fiktivního kandidáta: *Poprvé jsem to neudělal jen o bod. Opravný termín byl daleko těžší, chybělo mi 6 bodů. Jak je to možné?* nebo do otázky, kterou si ÚJOP UK, resp. VTC položilo: *Jak víme a jak prokážeme, že jsou všechny testové verze stejně obtížné, zejména v situaci, kdy nemáme možnost rozsáhlých (a ve speciálním designu prováděných) pretestací, a v situaci bez užití kotvicích úloh?*

2.2 Otazníky a omezení

Vzhledem k tomu, že zkouška CCE již existuje, není možné s ohledem na spravedlivost testování a možné důsledky pro kandidáty a testované změnit zkoušky ze dne na den, a i při předem ohlašovaných změnách je nutné zajistit porovnatelnost zkoušek.

VTC si bylo vědomo těchto vstupních omezení:

- Přechod na tzv. proměnlivý cut-off skór, tedy stanovovaný zvlášť pro každou testovou verzi není ze strany uživatelů výsledků zkoušek akceptovatelný, je obtížně komunikovatelný a jeho realizace by byla náročná na finanční a lidské zdroje.
- Interpretace výsledků musí být snadno vysvětlitelná, zejména testovaným.
- Cut-off skór musí číselně odpovídat 60 % z max. počtu bodů a musí být interpretován jako minimum, které musí zvládnout kandidát interně definovaný jako MKK dané úrovně obtížnosti.
- Pretesty se konají na malém počtu kandidátů a ne vždy je lehké zajistit složení pretestovaného vzorku tak, aby odpovídalo tomu, jak vypadá skupina při ostrém testování.
- Dosud není možné pretestovat tak, aby dlouhodobý model pretestací zahrnoval možnost kalibrovat testové úlohy.
- K ostrému testování se dostavují i mnozí kandidáti, kteří jsou výrazně pod nebo nad úrovní, na kterou cílí test, není proto možné se spolehnout na výsledky z ostrého testování.

2.3 Cíle

1. Vytvoření metody/postupu, jak sestavovat ekvivalentní (později případně i paralelní) testové verze/mutace z úloh, u nichž panelisté odhadli dílčí cut-off skór pro MKK dané úrovně.

2. Vytvoření banky úloh s odhadnutými parametry vzhledem k očekávanému výkonu MKK a s dalšími parametry položek.
3. Vytvoření způsobu, jak reportovat výsledky v případě, že cut-off skór některé testové verze nebude odpovídat (přesně) hodnotě 60 %.
4. Průběžná revize či doplňování konceptu MKK.

2.4 Řešení a volba metody

Na základě výše uvedených podmínek, omezení a stanovených cílů dospěli tvůrci testu k rozhodnutí, že nominální hodnota formálně stanoveného cut-off skóru se nezmění, ale bude zajištěno, aby „obtížnost" testových verzí odpovídala v hodnotě 60 % právě profilu MKK.

Nebude určován cut-off skór pro sestavený test, nýbrž budou sestavovány testy z jednotlivých úloh tak, aby 60 % cut-off odpovídal profilu MKK. Jednotlivé úlohy budou mít známou hodnotu „dílčího" cut-off skóru, která bude výsledkem odhadu skupiny panelistů v takzvaných průběžných setkáních ke standard settingu. Úlohy se budou do sestavovaných testových verzí vybírat tak, aby celkový součet dílčích cut-off skórů úloh odpovídal 60 % a aby se hodnoty dílčích cut-off skórů úloh na stejných pozicích (úlohy = pozice 1–12) příliš nelišily – viz ukázka modelování na obrázku 4.

Při plánování postupu a volbě metody byla zohledněna nejen výše uvedená výchozí situace a cíl, nýbrž také možnosti instituce a závažnost rozhodnutí o případném přehodnocení cut-off skórů – jak pro testované, tak pro poskytovatele zkoušek i uživatele výsledků zkoušek. Bylo tedy třeba zvážit, jaké jsou finanční a lidské zdroje instituce, jak časově náročný proces si může dovolit, s jakým množstvím materiálu bude třeba pracovat apod.

Na základě těchto úvah i očekávaného výstupu byla jako nejvhodnější pro provedení standard settingu v daném kontextu vyhodnocena metoda přímé shody (Pitoniak, Hambleton a Sireci, 2004; Cizek, 2001). Tato metoda je zaměřená na posouzení úloh, avšak na rozdíl od např. metody Angoffovy či jejích modifikací nepracuje s pravděpodobností, na rozdíl od také relativně často využívané košíkové metody (Basket method) nepřiřazuje položkám úroveň dle SERRJ (což by bylo hrubé rozdělení), nevyžaduje velké množství dat ani využití analýz založených na teorii odpovědi na položku (IRT). Pracuje s částmi testu, zde úlohami, a posuzuje je jako celek. Je tedy vhodná pro jazykové testování, kde jsou obvykle položky svázány společným stimulem (textem). Odhady panelistů jsou nezávislé na výsledcích kandidátů. Tato metoda je relativně rychlá, snadno pochopitelná a nenáročná na výpočty. Základní otázka, na kterou odpovídají panelisty zní: *Kolik bodů by celkem v této úloze získal MKK dané úrovně obtížnosti?*

V roce 2014, kdy byla metoda přímé shody využita poprvé, bylo posuzováno pět celých testů receptivních dovedností (A1–C1), sestandard settingu zúčastnilo 16 osob, online školicí část trvala 8 týdnů a prezenční část téměř 3 dny. Byly ověřeny aktivity směřující k familiarizaci s úrovněmi SERRJ, vytvořeny popisy MKK, ověřena vhodnost dvoukolového postupu i metody jako takové.

V tabulce 1 pro přehled uvádíme výstupy – hodnoty cut-off skóru tak, jak byly panelisty stanoveny pro jednotlivé subtesty v pilotní fázi 2014:

Tab. 1: *Výsledky standard settingu 2014*

	C-off skór %	C-off skór body	SD	SE	C-off skór %	C-off skór body	SD	SE	C-off skór %	C-off skór body	SD	SE				
A1	**Čtení**				**Poslech**				**Celý test**							
Kolo 1	57,88	14,47	1,92	0,48	58,25	14,53	2,35	0,59	58,00	29,00	3,84	0,96				
Kolo 2	**47,25**	11,81	2,44	0,61	**54,50**	13,63	1,84	0,46	50,88	25,44	3,83	0,96				
A2	**Čtení**				**Poslech**				**Celý test**							
Kolo 1	61,38	15,34	2,54	0,63	59,19	14,80	1,95	0,49	60,28	30,14	4,04	1,01				
Kolo 2	**57,81**	14,45	2,3	0,58	**57,62**	14,41	1,41	0,35	57,72	28,86	3,48	0,87				
B1	**Čtení**				**Poslech**				**Celý test**							
Kolo 1	69,38	17,34	3,18	0,79	68,69	17,17	1,88	0,47	69,03	34,52	4,83	1,21				
Kolo 2	**66,94**	16,73	1,69	0,42	**68,19**	17,05	1,08	0,27	67,56	33,78	2,51	0,63				
B2	**Čtení**				**Poslech**				**Gramlex**				**Celý test**			
Kolo 1	73,75	14,75	1,80	0,45	74,69	14,94	1,72	0,43	69,30	13,86	1,71	0,43	72,58	43,55	4,33	1,08
Kolo 2	**73,44**	14,69	1,21	0,3	**76,25**	15,25	1,16	0,29	**66,25**	13,25	1,14	0,29	71,98	43,19	3,05	0,76
C1	**Čtení**				**Poslech**				**Gramlex**				**Celý test**			
Kolo 1	66,35	19,91	2,37	0,59	68,65	20,59	1,94	0,48	59,58	17,88	2,66	0,67	64,86	58,38	5,27	1,32
Kolo 2	**65,1**	19,53	1,40	0,36	**65,94**	19,78	0,87	0,22	**59,38**	17,81	2,10	0,53	63,47	57,13	3,32	0,83

Jak je z tabulky 1 patrné, aplikace této metody v nezměněné formě tak, jak byla popsána v literatuře, by mohla vést a) k velmi odlišným cut-off skórům pro každý subtest; b) k tomu, že by bylo nutné každou další vytvářenou testovou verzi, i kdyby vznikla pouhou výměnou jedné z 8 úloh, znovu posuzovat, a to včetně již odhadnutých položek; c) k nutnosti cut-off skór měnit a oznamovat před realizací každého zkušebního termínu; d) k nutnosti provádět vyvažování skórů z jednotlivých testových verzí. Proto jsme se rozhodli změnit způsob, jak sestavovat testové verze, a předřadit standard setting procesu sestavování testů. Tím se zároveň přiblížíme i cíli vybudovat banku úloh se známými parametry.

3 Stanovování hraničního skóru od roku 2015

3.1 Rozhodnutí

Jako nevyhovující byla identifikována výjimka 50 % úspěšnosti v jednom ze subtestů a jako taková byla s platností od roku 2016 zrušena. Cut-off skór 60 % nebude změněn, ale je považován za nepřenositelný z jedné testové verze na jinou. Proto byl v roce 2015 ověřen a následně zaveden tzv. *průběžný standard setting* a nový způsob sestavování testových verzí pomocí tzv. *modelování*. Tento postup bude případně doplněn o *lineární transformaci skórů*.

3.2 Realizace

Průběžný standard setting znamená aplikaci metody přímé shody (a všech ostatních doprovodných aktivit standard settingu) na úrovni úloh. Panelisté tedy budou tak jako v klasické metodě přímé shody odhadovat dílčí cut-off skór úloh pro jednotlivé úrovně dle SERRJ, avšak tyto úlohy nebudou patřit do konkrétní testové verze. Úlohy se po provedení odhadu dílčího cut-off budou vracet do banky úloh spolu s informací o dílčím cut-off skóru. Z banky úloh budou na základě předem stanovených kritérií (např. téma, typ textu, cut-off skór pro MKK) vybírány úlohy tak, aby z nich bylo možné sestavit obsahově vyváženou testovou verzi (resp. verzi subtestu), jejíž celkový cut-off skór 60 % (součet dílčích cut-off skórů) bude odpovídat nebo se co nejvíce blížit profilu MKK (= modelování cut-off skóru). V případě nemožnosti sestavit (tj. namodelovat) testové verze s cut-off skórem 60 % bude proces doplněn o lineární transformaci výsledků kandidátů (viz níže).

Takto sestavené testové verze budou moci být považovány za ekvivalentní testové verze.

3.3 Provedení průběžného standard settingu 2015: postup a ukázky výsledků

Při jakémkoli způsobu odhadu cut-off skóru je vždy nezbytné provádět důkladnou familiarizaci panelisty s obsahem a s metodou, a to včetně opakování stejné metody po určitém čase. Familiarizační fázi jsme v roce 2015 prováděli kombinovanou formou, tedy část aktivit online a část aktivit prezenčně. Vždy šlo o práci s deskriptory referenčních úrovní SERRJ a s popisy MKK. Cílem bylo upevnit interpretaci úrovní i relevantních deskriptorů a vymezit profil MKK, který je stěžejním konceptem při posuzování vztahu obsahu a cílů posuzovaných úloh a toho, co se očekává od MKK dané úrovně.

Prezenčně byli panelisté proškoleni v metodě posuzování a byli seznámeni s celým procesem posuzování, se způsobem zápisu odhadů a s tím, jak jim budou prezentovány výsledky jejich odhadů. Důležitou součástí byla i prezentace výstupů ze standard settingu z roku 2014, zdůvodnění změn, ke kterým VTC přistoupilo, a informace o plánovaném využívání průběžného standard settingu a banky úloh.

Po familiarizační fázi následovalo dvoukolové odhadování úloh. V prvním kole odhadovali panelisté cut-off skór samostatně, a to jako odpověď na otázku *Jaký průměrný skór BY ZÍSKAL minimálně kompetentní kandidát (MKK) v každé z x částí testu?* Své odpovědi – odhady zapsali do online sběrné aplikace, která poskytovala možnost okamžité zpětné vazby. Tato zpětná vazba byla podkladem pro skupinovou diskusi nad jednotlivými odhady. Diskuse byla moderovaná tak, aby argumenty diskutujících reflektovaly zejména koncept MKK. Po ukončení diskuse následovalo druhé kolo individuálních odhadů a prezentace výsledků druhého kola. V závěrečném vyhodnocení byly prezentovány výsledky před tzv. modelováním a za využití modelování. V prvním případě byly subtesty sestaveny tak, jak byly původně navrženy bez informací o odhadnutém cut-off skóru; byla tedy zohledněna pouze obsahová kritéria (viz tabulka 2 a 3). Následovalo modelování, které při respektování obsahových kritérií umožnilo kombinací úloh z různých subtestů sestavit testové varianty tak, že kromě obsahových kritérií zohledňovaly i požadavek na celkový cut-off skór na 60 % (viz tabulka 4).

Tab. 2: *Výsledky posuzování subtestu Čtení B1 (Kolo 2)*

	úloha 1		úloha 2		úloha 3		úloha 4		% cut-off subtestu bez modelování
	Average K2		Average K2		Average K2		Average K2		
	SD	SE	SD	SE	SD	SE	SD	SE	
V1	**73,44**		**58,59**		**59,38**		**50,00**		**60,64**
	0,56	0,20	0,32	0,11	0,32	0,22	0,31	0,11	
V2	**76,04**		**64,45**		**57,29**		**57,75**		**63,38**
	0,37	0,13	0,82	0,29	0,30	0,11	0,35	0,12	
V3	**71,35**		**51,56**		**61,98**		**63,13**		**61,13**
	0,44	0,16	0,57	0,20	0,32	0,11	0,60	0,21	

Tab. 3: *Výsledky posuzování subtestu Poslech B2 (Kolo 2)*

	úloha 1		úloha 2		úloha 3		úloha 4		% cut-off subtestu bez modelování
	Average K2		Average K2		Average K2		Average K2		
	SD	SE	SD	SE	SD	SE	SD	SE	
V1	**63,57**		**70,00**		**68,43**		**63,57**		**66,39**
	0,22	0,08	0,33	0,12	0,48	0,18	0,39	0,15	
V2	**60,00**		**64,29**		**52,86**		**42,86**		**55,00**
	0,27	0,10	0,34	0,13	0,26	0,10	0,40	0,15	

(V1 – verze 1; Average K2 – průměr odhadů panelistů ve druhém kole v %; SD a SE – směrodatná odchylka a směrodatná chyba průměru v bodech)

Jak lze vidět v tabulkách 2 a 3, cut-off skór z náhodně sestavených úloh se u subtestů liší, vnitřní variabilita dílčích cut-off skórů je různá v každé verzi každého subtestu – subtest obsahuje úlohy (a položky) s širším rozptylem obtížnosti. Je žádoucí a tato tendence se zde projevuje, aby byla variabilita co nejnižší mezi verzemi jedné úlohy, variabilita napříč úlohami jedné testové verze by měla být naopak vyšší, neboť jde o test ověřující dosažení dané úrovně (zde B1 a B2), kterou lze považovat za kontinuum z hlediska obtížnosti.

Na příkladu v tabulce 3 (poslech B2) vidíme, že průměrné odhady úloh (ve sloupcích) ve verzích V1 a V2 se liší, rozdíl se pohybuje mezi 3,57 procentními body (úloha 1 verze 1 a 2) a 20,71 procentními body (úloha 4 verze 1 a 2). Lze však také pozorovat, že úlohy zařazené do verze 2 jsou všechny pro MKK těžší (odhadovaný počet bodů získaný MKK je nižší) než u verze 1. Proto je také výsledný cut-off skór verze 1 a verze 2 velmi odlišný. Rozdíl činí 11,39 procentního bodu, což v bodech činí cca 2,28 bodu (maximum pro subtest u B2 je 20 bodů). V tomto případě bude nutné modelování (viz tabulka 4) a případně i transformace skórů.

Tab. 4: *Ukázka cut-off skórů po tzv. modelování*

Odhady	Poslech					Čtení									
	1	2	3	4	%	5	6	7	8	%					
B1	3,86	3,96	4,04	4,36	64,86	4,28	5,16	3,56	2,50	62,00					
	3,25	4,46	3,93	4,46	64,43	4,56	4,69	3,44	2,69	61,50					
	3,61	4,82	3,79	3,82	64,14	4,41	4,13	3,72	3,16	61,63					
	Poslech					Čtení					Gramlex				
	1	2	3	4	%	5	6	7	8	%	9	10	11	12	%
B2	3,18	3,50	3,42	2,14	61,21	3,63	2,34	4,00	3,69	68,28	3,53	1,39	3,92	4,03	64,31
	3,00	3,21	2,64	3,18	60,18	3,06	3,25	4,22	3,03	67,81	3,50	1,69	3,78	3,44	62,08
								4,16			3,42	1,75	3,53	3,72	62,08

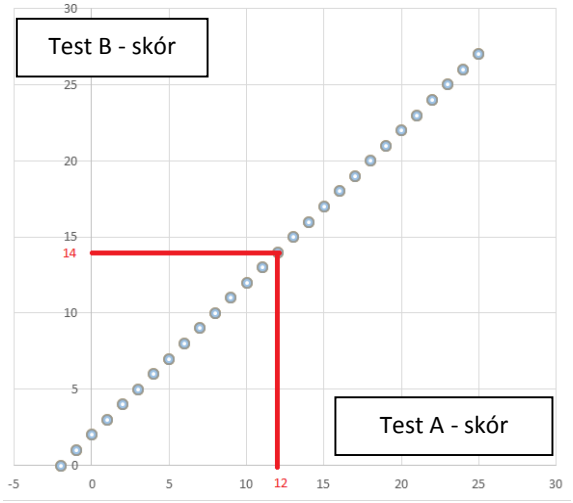
3.4 Vyvažování skórů

Zatímco modelováním jsme získali testové verze, které jsou srovnatelné z hlediska cut-off skóru, zbývá ještě vyřešit problém s odlišností cut-off skóru od stanovené hranice 60 %, ke které musí být základní rozhodnutí *uspěl/neuspěl* vztaženo. Jako řešení byla navržena lineární transformace. Jsme si vědomi, že skóry na obou koncích škály jsou zkresleny, nicméně vzhledem k tomu, že se skóry kandidátů dále nepřevádějí na známky, je lineární transformace postačující.

V tabulce 5 a na obrázku 1 vidíme ukázku transformace pro skóry ve čtení úrovně B2, kde byla odlišnost stanoveného cut-off skóru od 60 % hranice cca 7 procentních bodů, tedy asi 1,5 až 2 body. Test A je testem modelovým, referenčním, test B je nově sestavená modelovaná verze. Ve verzi A je 60 = hranice rovna 12 bodům, ve verzi B 14 bodům. Tyto dva body tedy stanovíme jako rovnocenné a převedeme lineárně 14 bodů na 12, 13 bodů převedeme na 11 apod.

4 Zhodnocení procesu

Vzhledem k tomu, že jsme pracovali s metodou stanovování hraničního skóru založenou na subjektivním posuzování úloh a jejich vztahu k MKK, bylo třeba sledovat



Obr. 1: *Příklad transformace skórů*

Tab. 5: *Příklad transformace skórů*

suma bodů Test B	skór v %	nový skór v %	suma bodů test A
0	0	-10	-2
1	5	-5	-1
2	10	0	0
3	15	5	1
4	20	10	2
5	25	15	3
6	30	20	4
7	35	25	5
8	40	30	6
9	45	35	7
10	50	40	8
11	55	45	9
12	60	50	10
13	65	55	11
14	70	60	12
15	75	65	13
16	80	70	14
17	85	75	15
18	90	80	16
19	95	85	17
20	100	90	18
21	105	95	19
22	110	100	20

také spolehlivost panelistů a získaných výsledků, a to i přes důraz kladený na obsah a rozsah familiarizační fáze. Zkoumali jsme spolehlivost metody, význam intervence na hodnocení panelistů, spolehlivost a konzistentnost panelistů jako skupiny a také jednotlivých panelistů. V některých případech bylo třeba ještě kvalitativní interpretace na úrovni subtestů či úrovní B1 a B2, aby bylo možné

identifikovat pravděpodobnou příčinu odchylek či nižší míry konzistentnosti ve srovnání s ostatními panelisty.

4.1 Spolehlivost metody (Test-retest reliabilita)

Bylo třeba zjistit, zda je výsledek – tedy hodnoty odhadů – spolehlivý. Nejprve jsme porovnávali průměry úloh (dílčí cut-off skóry) za první a druhé kolo odhadů. Pro odhad spolehlivosti výsledků metody (analogicky k test-retest reliabilitě) jsme použili Spearmanův pořadový korelační koeficient, neboť nás nezajímalo porovnání hodnot – dílčích cut-off skórů úloh, nýbrž to, zda skupina panelistů bude přistupovat ke stejným úlohám stejně v prvním i druhém kole. Důležité tedy bylo to, do jaké míry pořadí úloh dle odhadované obtížnosti pro MKK v prvním kole koreluje s výsledky kola druhého, respektive to, zda panelisté posuzují stejné úlohy v subtestu a v úrovni (B1 a B2) stejným způsobem. Naší hypotézou bylo, že pokud panelisté aplikují metodu konsistentně, pak jimi odhadnuté úlohy v prvním i druhém kole budou vykazovat stejné pořadí dle hodnoty dílčího cut-off skóru. Poté bylo hodnotám přiřazeno pořadí. Spearmanův korelační koeficient byl vypočten pro úlohy v každém z pěti subtestů a dále pro úlohy v obou úrovních B1 a B2. Výsledky jsou sumarizovány v tabulce 6.

Tab. 6: *Korelace výsledků Kola 1 a Kola 2*

subtest	B1 Čtení	B1 Poslech	B1 Celý test	B2 Čtení	B2 Poslech	B2 Gramlex	B2 Celý test
Spearmanův koeficient	**0,979**	**0,979**	**0,973**	**1**	**0,790**	**0,658**	**0,866**

Byla zjištěna silná až velmi silná korelace mezi oběma koly posuzování. Hodnoty korelace jsou vyšší než kritické hodnoty na hladině významnosti 0,95. Lze tedy konstatovat, že postup odhadování, resp. jeho výsledky jsou spolehlivé a panelisté posuzovali obtížnost úloh v prvním i druhém kole konzistentně.

4.2 Zkoumání vlivu intervence – diskuse mezi koly a prezentace výsledků prvního kola

V dalším kroku jsme zkoumali, jak intervence ovlivnila hodnocení panelistů, resp. jak hodně a jakým směrem se jejich odhad dílčího cut-off skóru úloh změnil. Předpokládáme, že v prvním kole odhadů je individuální hodnota odhadu dána tím, jak panelista pochopil koncept MKK a jak jej aplikoval na úlohu. Po intervenci pak v kole druhém vstupuje do odhadu dílčího cut-off skóru i obsah a závěry skupinové diskuse a normativních dat. Naší hypotézou bylo, že na základě intervence, tedy prezentace výstupů prvního kola a řízené diskuse panelistů o důvodech hodnocení s důrazem na argumentaci opírající se o koncept MKK, dojde ke zpřesnění interpretace a tím ke snížení rozptylu odchylky od průměru (odhadu dílčího cut-off skóru). Proto jsme porovnali průměry odchylek v kole 1 a 2.

Spočítali jsme pro každého panelistu, každý subtest a každé kolo celkovou odchylku od průměru a sečetli je. Takto jsme získali celkovou míru variability (rozptylu) ve skupině panelistů v prvním a druhém kole. V tabulce 7 vidíme, že došlo ke snížení rozptylu zhruba o jednu třetinu. Lze tedy říci, že intervence pravděpodobně měla vliv na variabilitu hodnocení panelistů (variabilita poklesla) a skupina se stala soudržnější.

Tab. 7: *Variabilita skupiny: Kolo 1 versus Kolo 2*

	Suma odchylek (v bodech)
Kolo 1	166,052
Kolo 2	117,761

O odchylkách hodnocení jednotlivých panelistů pojednáváme v následujícím oddíle, v souvislosti s konzistentností panelistů. Toto zkoumání považujeme za důležité také proto, že s panelisty budeme pracovat i v budoucnu a chceme jim poskytovat individualizovanou zpětnou vazbu o jejich přístupu k posuzování, jejich tendencích a míře spolehlivosti a dlouhodobě tak pracovat na zkvalitňování skupinového i individuálního hodnocení.

4.3 Zkoumání konzistentnosti/spolehlivosti panelistů jako skupiny

Zabývali jsme se otázkou, do jaké míry je odhad panelistů spolehlivý. Za ukazatel jsme určili míru přísnosti a mírnosti[5] vůči průměru celé skupiny. Zajímalo nás, zda panelista posuzuje ve shodě s průměrem skupiny a na něj navázaným intervalem spolehlivosti. Zvolili jsme hladinu významnosti 0,01. U každého panelisty (N=8), u každé úlohy (N=53) a kola (K1 a K2) jsme určili, kolikrát a kterým směrem (pod, nebo nad interval spolehlivosti) panelista vybočí. Celkem bylo provedeno 808[6] posouzení úloh. Panelista byl označen jako mírný, pokud bylo jeho hodnocení úlohy POD dolní hranicí intervalu spolehlivosti; jako přísný byl označen tehdy, jestliže jeho hodnocení bylo NAD horní hranicí intervalu spolehlivosti. V obou případech dostal za odchylku od intervalu spolehlivosti označení 1. Pokud se jeho hodnocení pohybovalo UVNITŘ intervalu spolehlivosti, dostal 0. Tabulka 8 udává celkový počet odchylek skupiny v každém kole.

Z tabulky 8 vyplývá, že četnost odchylek celkově poklesla ve druhém kole, pravděpodobně vlivem intervence. Největší posun je patrný v kategorii „mírnost“, kde se četnost odhadů POD dolní hranicí intervalu spolehlivosti snížila. Z tabulky 8 však

[5] Označení mírnost a přísnost je užito i přesto, že nejde o totéž, jako když se posuzuje spolehlivost hodnotitelů např. písemné práce, obvykle vůči etalonu. Nicméně princip posuzování je analogický a jako etalon zde funguje průměr a k němu zvolený interval spolehlivosti.

[6] Panelista 3 se nezúčastnil všech posuzování, proto není možné pracovat s násobkem 53x2x8.

Tab. 8: *Celkový počet odchylek v Kole 1 a Kole 2*

	Mírnost	Přísnost	Suma
K1	75	78	153
K2	55	78	133
Suma	130	156	286

nelze vyčíst, zda se četnosti odchylek týkají stále stejných úloh a hodnotitelů, nebo zda došlo ke změnám uvnitř těchto skupin, i přesto, že výsledná čísla jsou stejná.

4.4 Zkoumání spolehlivosti/konzistentnosti panelistů jako jednotlivců

Zajímalo nás, zda je možné panelisty jako jednotlivce považovat za spolehlivé a konzistentní, zda je některý příliš přísný nebo příliš mírný a zda, případně jak je možné mírnost, přísnost či nekonzistentnost některého panelisty dát do souvislosti s charakteristikami posuzovaných úloh, neboli zda existuje souvislost mezi hodnocením úlohy[7] a jejími vlastnostmi. Tyto informace chceme využít v budoucnu jako podklad při úvahách o případném přeškolení nebo dokonce i vyřazení panelisty. Údaje o celkovém počtu odchylek jsou uvedeny v tabulce 9. Bylo zjištěno, že panelisté 6 a 7 jsou nejméně v konsensu se skupinou, mají nejvíce odchylek od intervalu spolehlivosti, a to jak v mírnosti, tak v přísnosti. Panelista 4 vykazuje tendenci k přísnějšímu posuzování úloh ve vztahu k očekáváním na MKK, částečně by se toto dalo tvrdit i o panelistovi 5.

Tab. 9: *Počty odchylek panelistů*

ID	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**Suma**
Mírnost	10	10	2	8	17	31	32	20	130
Přísnost	10	12	4	26	28	34	22	20	156
Suma	20	22	6	34	45	65	54	40	286*

* celkem 808 posouzení (úloha × posuzovatel × kolo)

Při podrobném pohledu na odchylky v jednotlivých subtestech a kolech zjistíme následující: panelista 7 se výrazně lišil od průměru (směrem k mírnosti) v Gramaticko-lexikálním subtestu na úrovni B2. V obou kolech byl 9x, resp. 8x pod dolní hranicí intervalu spolehlivosti, nad horní hranici intervalu spolehlivosti se ale jeho hodnocení nedostalo ani jednou. Jako přísný se jeví zejména v subtestu Čtení s porozuměním na úrovni B2; jeho hodnocení v obou kolech je konzistentní. Pokud bychom chtěli teoreticky zvážit do budoucna to, jak k panelistům tohoto profilu přistupovat, pak se zdá vhodné spíše doškolení zaměřené na interpretaci úrovně B2 a MKK B2, než vyřazení takovéhoto panelisty.

[7] V této fázi jsme v hodnocení spolehlivosti panelistů zkoumali pouze souvislost se subtesty a úrovněmi.

U panelisty 4, který v celkovém součtu odchylek jevil tendenci k přísnosti, můžeme vidět posun v přísnost ve druhém kole, zřejmě pod vlivem intervence, a to ve Čtení s porozuměním a Poslechu s porozuměním na úrovni B1 a v Gramaticko-lexikálním subtestu na úrovni B2. V případě úrovně B1 můžeme považovat za možnou příčinu posunu k přísnosti vliv argumentace skupiny a přehodnocení interpretace MKK B1 tímto panelistou. U Gramaticko-lexikálního subtestu by bylo vhodné podívat se ještě na to, ve kterých konkrétních úlohách k odchylkám docházelo, k jak velkým a zda panelista měl k úlohám nějaké výhrady.

Panelista 6 byl konzistentní sám se sebou, intervence, tedy diskuse a prezentace výsledků kola prvního na jeho hodnocení ve druhém kole měla spíše menší vliv. Zajímavý je jeho přístup k poslechu: na úrovni B1 je přísný, na úrovni B2 naopak mírnější. S tímto typem panelistů (pokud vyloučíme vliv testologické vady úlohy) by bylo vhodné pracovat na úrovni dovednosti a proškolit jej opakovaně v konceptu MKK pro poslech a v kontrastu očekávání pro úroveň B1 a B2.

Před případným rozhodnutím o vyřazení některého panelisty z panelu se však vždy budeme zabývat nejen četností a rozložením odchylek, nýbrž také jejich velikostí, případně v budoucnu i tím, jak byl panelista konzistentní v čase: plánujeme mít tzv. incomplete design průběžného standard settingu, tzn. pracovat s kotvicími úlohami. Bude tak možné sledovat vývoj panelisty v čase a rozhodování o setrvání v panelu, nebo o vyřazení bude postaveno na více zdrojích informací.

Budeme také sledovat, do jaké míry je rozhodování panelistů ovlivňováno jejich vnímáním testologických kvalit úloh (na základě poznámek v testových sešitech nebo v diskusi) a jak se vztahuje k výsledkům položkových analýz. Informace o chování panelistů (jejich mírnosti a přísnosti) budou využívány jako zpětná vazba pro panelisty samotné, jako doložení procedurální validity procesu stanovování standardu a jako ukazatel vlivu intervence.

5 Závěr

Na základě ověřování v roce 2014 a 2015 byl zaveden nový postup sestavování srovnatelných testových variant. Průběžnými standard settingy za užití modifikované metody Direct Consensus se buduje banka úloh s odhadem cut-off skóru pro minimálně kompetentního kandidáta. Kombinací úloh z této banky se modelují testové varianty, které sečtením dílčích cut-off skórů jednotlivých úloh, dosahují celkového cut-off skóru 60 %. V případech, kdy má sestavená testová verze celkový cut-off skór pro MKK vyšší nebo nižší než 60 %, se využívá lineární transformace skórů, což je další opatření k dosažení srovnatelnosti testových verzí.

V souvislosti s touto změnou byla zrušena dřívější možnost získat v jednom z písemných subtestů pouze 50 % cut-off skóru. Tato výjimka byla vyhodnocena jako

nekompatibilní se zavedeným způsobem sestavování testových verzí a s úsilím o zkvalitňování zkoušek.

Aplikací nového postupu se změnily také procesy, které s vývojem a konstrukcí testu souvisejí. Došlo ke změně ve způsobu práce s tvůrci, v posuzování kvality úloh a jejich vztahu k referenčním úrovním SERRJ, zejména ve smyslu standardizace posuzovacích postupů.

Byl zaveden nový způsob poskytování zpětné vazby tvůrcům úloh, který využívá dat z průběžného standard settingu.

Průběžný standard setting zároveň umožňuje získávat přesnější informace o chování panelistů a činit závěry o jejich konzistentnosti a tendencích k mírnému či přísnému posuzování a využívat tyto informace v další práci s panelisty.

Do konstrukce testových verzí se také dostává nový prvek, a sice využívání zpětné vazby ze standard settingu a propojení s normativními daty z ostrých testování těchto. Účelem tohoto kroku je validace konceptu minimálně kompetentního kandidáta, který je stěžejním jak pro proces standard settingu, tak pro přiřazení zkoušek k referenčním úrovním SERRJ.

Literatura

AERA, APA & NCME. (1999). *Standards for educational and psychological testing*. Washington D.C. Český překlad: Standardy pro pedagogické a psychologické testování (2001). Praha: Testcentrum (přel. Klimusová, H.)

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A Generalized Examinee-Centered Method for Setting Standards on Achievement Tests. *Applied Measurement in Education*, 12:4, 343–366, retrieved April, 27, 2015.

Cizek, G. J. (Ed.) (2001). *Setting performance standards: Concepts, methods, and perspectives.* Mahwah, NJ: Lawrence Erlbaum Associates.

Hambleton, R. K., & Pitoniak, M. J. (2005). Setting performance standards. In Brennan, R L. (Ed.) *Educational measurement.* 4th Ed. Westport, CT: American Council on Education and Praeger Publishers. (Technical), pp. 433–470.

Livingston, S. A. (2004). *Equating test scores (without IRT).* Princeton, NJ: ETS.

Messick, S. (1996). Validity and washback in language testing. *Language Testing* 13(3): 241–256.

Pitoniak, M. J., Hambleton, R. K., & Sireci, S. G. *Advances in Standard Setting for Professional Licensure Examinations.* Amherst: University of Massachusetts. Retrieved April 2014 from http://www.performancetest.org/uploads/resources/Advances_in_Standard_Setting_for_Professional_Licensure_Examinations.pdf

Relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR). A Manual. (2009). Strasbourg: Council of Europe.

Autorka

Mgr. Martina Hulešová, MA, e-mail: mhulesova@volny.cz, Výzkumné a testovací centrum Ústavu jazykové a odborné přípravy UK Praha.

Autorka je absolventkou FF UK, oboru hispánská filologie, a MA in Language Testing na Lancaster University. Vyučovala španělský jazyk, později pracovala v CERMATu v Sekci evaluačních nástrojů. V současné době působí ve Výzkumném a testovacím centru ÚJOP UK. Věnuje se dalším projektům souvisejícím s testováním a hodnocením v ČR i v zahraničí, především v rámci Asociace jazykových testerů ČR.