KONSTRUKCE TESTŮ A EMPIRICKÉ VÝSLEDKY TESTOVÁNÍ

Hana Romová
Policejní akademie ČR v Praze, Fakulta bezpečnostního managementu, Katedra jazyků,
Lhotecká 559/7, 143 00 Praha 4, Česká republika
e-mail: romova@polac.cz

Abstrakt

Hodnocení v oblasti osvojování neboli učení se cizím jazykům má ve srovnání s jinými oblastmi vzdělávání specifické rysy. Jazykové testování je kromě jiného unikátní tím, že jazyk je nejen obsahem, ale i nástrojem měření. Pro potřeby testování byla řešitelským týmem vytvořena banka testových úloh obsahující svazky položek, ze kterých lze vytvořit v subtestech velký počet testů obsahujících termíny a profesní lexiku korespondující s reálnými požadavky kladenými na pracovníky v praxi. Byly sestaveny achievement testy a kontrolní testy na základě položkové banky, takže z funkčního hlediska převládá hodnotící funkce. Achievement testy nejsou zaměřeny jen na testování úspěšnosti jednotlivých studentů, ale také na úspěšnost vyučujících. Z hlediska interpretace výsledků je pozornost věnována achievement testům, které umožňují součinnost reálných počtů položek a testovaných studentů.

Keywords

Test construction; Item bank; Items; Reliability; Distractors; Terms; Terminology; Practice.

Úvod

V humanitních oborech se v oblasti testování, měření a vyhodnocování výsledků práce respondentů potýkáme s naprosto zásadním problémem, a to jak objektivizovat výstupní informace získané v procesu testování. Vysoká míra subjektivity vlastní těmto oborům a velké množství dat, které je nutné vyhodnotit, vedou k implementaci a aplikaci postupů, které digitalizují získaná data a dále pak umožňují jejich elektronické vyhodnocování na základě statistických počítačových modelů.

Při testování je třeba vycházet ze tří základních otázek: co bude hodnoceno, jak, tj. jaké nástroje a způsob „měření“ je třeba zvolit, a v jakém kontextu, tj. mimo jiné k jakému účelu budou výsledky testování sloužit.

Jazykové testy mají nepochybně dlouhou historii. Za první odbornou monografii k problematice testování anglického jazyka bývá považováno Language Testing, které v roce 1961 publikoval R. Lado. Koncem 70. a počátkem 80. let 20. století se začíná rozvíjet, komunikační přístup k jazykovému testování, který je považován za jeden z nevlivnějších trendů.

Počátky komunikačního přístupu byly výrazně zatíženy negací předchozích priorit, zejména důrazu na reliabilitu testů. Rané komunikační testování považovalo reliabilitu a validitu testů, resp. interpretace jejich výsledků, za kvality, které jsou vzájemně negativně závislé. Došlo k posunu a reliabilita spolu s validitou dnes patří k základním vlastnostem každého jazykového testu. Kvalita (nejen) jazykového testu přímo souvisí s účelem, k němuž je využit. Na základě účelů testování pak můžeme hovořit o typech testů. Klasifikace testů v zásadě vycházejí především ze způsobů interpretací skórů.

Věnujeme-li se testování odborného a akademického jazyka v současnosti u nás, jsou inspirativní poznatky a doporučení, které poskytují publikace a semináře v rámci projektu COMPACT: Kompetence v jazykovém vzdělávání na Masarykově univerzitě v Brně, který je spolufinancován Evropským sociálním fondem a státním rozpočtem České republiky.

1 Předmět výzkumu

Předmětem našeho výzkumného zájmu byla vybraná problematika testů jazykových, a to didaktických. Zajímají nás především testy hodnotící stupeň osvojení jazykových prostředků a řečových dovedností. Zaměřujeme se v souladu s cílem osvojování cizího jazyka na vysoké škole na testování odborné slovní zásoby, na poslech odborných textů a jejich čtení s porozuměním.

Součástí řízení pedagogického procesu je kontrola a hodnocení výsledků učebních činností studentů, a to při respektování pedagogických a didaktických principů a specifik studijního předmětu.

„Společný evropský referenční rámec pro jazyky“ hodnocení věnuje adekvátní pozornost. Podrobně jsou vyloženy typy hodnocení, jejich výhody a limitace s ohledem na účel hodnocení, a diskutovány otázky praktičnosti či proveditelnosti při hodnocení. Samostatnou část tvoří testování jako jedna z forem hodnocení výsledků cizojazyčného vzdělávání [3]. Navazující publikace pod záštitou Rady Evropy dále problematiku testování rozvíjejí.

Předmětem našeho výzkumného zájmu byla vybraná problematika testů jazykových, a to didaktických. Zajímají nás především testy hodnotící stupeň osvojení jazykových prostředků a řečových dovedností. Zaměřujeme se v souladu s cílem osvojování cizího jazyka na vysoké škole na testování odborné slovní zásoby, na poslech odborných textů a jejich čtení s porozuměním.

Část týmu – s odborností anglický jazyk – postupně zpracovávala zejména problematiku tvorby položkové banky (ITEM BANK), realizovala pilotní fázi zpracovaných testů a interpretaci získaných dat z empirického šetření. Druhá část týmu – s odborností německý jazyk – vytvářela odborné materiály, rozpracovávala je didakticky a tvořila k nim testy s klíčem [2].

2 Položková banka, položky, svazky položek, úložiště testových položek

V návaznosti na rozbor skutečných potřeb testování byly především pro položkovou banku pro anglický jazyk vypracovány zásady jejího vytváření, konstruktu a specifikací testů a jejich obsahové náplně vycházející z učebnic doporučených pro studium.

Položková banka, terminologicky též nazývaná banka testových úloh (BTÚ) nebo banka testových položek (BTP), anglicky Item Bank, v našem pojetí operuje s jednotlivými kompletními svazky položek, které jsou výsledkem promyšleného sestavování tvůrcem testu, nikoli náhodného generování počítačem či náhodného výběru tvůrcem testu ze sumy položek. Vzhledem k našim technickým možnostem vytváříme tuto banku ve smyslu úložiště testových položek, nikoli ve smyslu celého informačního systému se všemi procesy až po vygenerování testu počítačem.

Tvorba kvalitních testových položek je didakticky a časově velmi náročná činnost, na níž se v podmínkách jazykové katedry podílí testovací tým vyučujících, kteří oproti profesionálním testovacím pracovištím implementují všechny funkce v procesu tvorby testů, tzn. tvůrců testů, moderátorů, administrátorů a testerů.

2.1 Struktura položkové banky, svazky položek

Na základě postupného vytváření položkové banky lze následně pomocí různých kombinací v jednotlivých subtestech vytvořit poměrně velký počet různých testů se stejnou specifikací a konstrukcí. Vytvoření kvalitních a reliabilních testů závisí na tom, zda položky skutečně měří to, co autoři zamýšlejí měřit, zda jsou reliabilní (tzn. zda poskytují stejné výsledky za stejných podmínek při opakovaném zadávání) a zda jsou objektivní. Vzhledem k požadavku jazykových didaktických testů pokrývat rovnoměrně celý testovaný úsek učiva jsme se rozhodli vytvářet již zmíněné svazky položek.

Jednotlivé svazky položek obsahují stejný počet položek (deset položek, poslech a čtení s porozuměním pět položek) stejného typu a formy podle stanoveného paradigmatu, konstrukce a specifikace. Každá položka ve svazku je typově multiple choice se čtyřmi distraktory a správná je vždy jen jedna možnost. Na úrovni B1, B1+, B2– nevolíme možnost „není uvedeno“, ani zřetelně nesprávné nebo neexistující struktury (zejména v lexikálně-gramatické části). Každý svazek má svoji specifikaci, informaci o vývoji a vlastnostech položek, řekněme jakýsi svůj „rodný list“.

2.2 Příklad specifikace

- téma (aby nedocházelo k dublování témat v rámci jednoho testu)
- autor daného svazku položek
- moderátor (odborník v daném jazyku, v některých případech odborník z dané oblasti konzultující obsahovou přesnost testovaných termínů)
- obtížnost (předpokládané užití jazykových prostředků minimálně na úrovni B1 podle SERR („Společného evropského rámce pro jazyky“), odborná terminologie přesahuje do úrovně B2)
- datum vložení (vybrané svazky/položky by se neměly použít pro zadání do opravných testů ani v následujícím akademickém roce)
- status položky (např. aktivní, ve kterém testu, kdy a kým byla využita; svazky/položky po delší dobu neaktivní vyřadit nebo zrekonstruovat)
- reliabilita (vyjádřena koeficientem r s hodnotami od 0 do 1, závislá na počtu položek *p*, mediánu *m* a rozptylu s^2 a vypočítaná podle vzorce $r = m\,(p - m) / ps^2$, přičemž očekávaná spolehlivost je $r = 0{,}6$ a vyšší. Položky s nízkou reliabilitou je nutno přepracovat. Dále je třeba přepracovat velmi snadné (splněné 85 % a více testovaných) a velmi obtížné položky (splněné pouze 25 % a méně testovaných.)
- správná odpověď (zatmavěná odpověď v části odpověďníku subtestu Řešení)

Důležitá je úloha moderátora či moderátorů, jejichž úkolem je mimo jiné kriticky revidovat položky, aby se vyloučila analogie mezi nimi a aby byly položky koherentní a neodchylovaly se od požadavků kladených na danou jazykovou úroveň.

Vzhledem k poměrně velkému objemu slovní zásoby, která má být testována, je žádoucí, aby selekce termínů a profesní lexiky pro úlohy byla velmi pečlivá a korespondovala s reálnými požadavky kladenými na pracovníky v praxi. Tvorba položek ve svazcích vychází z klasického testování, kdy otázkami pokrýváme celý rozsah předpokládané úrovně znalostí.

3 Základní zásady tvorby a použití distraktorů

Tvůrce testu by se měl vyvarovat nesprávného použití distraktorů, to znamená, že by neměl užívat neexistující tvary, absurdní či banální odpovědi, chytáky, nereálný obsah či působit sugestivně. Účelem testu je uchopit lexiku a gramatické prostředky jako nástroj komunikace a pochopení vnitřních i vnějších souvislostí, nikoli jako samoúčelný nástroj prověřování znalostí jednotlivých slovíček či gramatických jevů.

Platí obecné pravidlo, že test prověřuje reálně existující cizí jazyk a ne to, zda je student schopen rozlišit, které slovo nebo tvar existuje a které nikoliv.

Položky by měly být formulovány tak, aby studenti neměli při vyplňování testu možnost taktizovat. Taktizováním myslíme situaci, kdy odpovědi, zejména v gramatické části obsahují nezamýšlené narážky či nápovědy, které bystrým studentům pomohou zvolit správnou odpověď, aniž by ji měli podloženou patřičnou znalostí. Zařazujeme-li testování tzv. soft skills, máme v jazykovém testování na mysli prověření dovednosti používat zavedený úzus ve společenských a pracovních situacích, jejichž neznalost by mohla vést k neporozumění nebo komunikačním nepříjemnostem. Jako příklad lze uvést následující otázky:

1. Jak byste reagovali na žádost vašeho anglického instruktora?
   Will you stay here overtime this afternoon, please?
   a) No, I'm busy.
   b) Oh, I'm sorry, but I must leave now.
   c) No, I won't.
   d) Oh, no, again?

správná odpověď b)

2. Odpovězte zdvořile na následující sdělení:
   May I introduce you to sergeant Diane Brown? Her husband is your instructor.
   a) Nice to meet you, Mrs.

   b) Nice to meet you, lady.
   c) Nice to meet you, Mrs Brown.
   d) Nice to meet you, Miss Brown.

správná odpověď c)

3.1 Druhy a formy položek: uzavřené položky, multiple choice questions, stem, distraktory

Pro účel této banky vytváříme uzavřené položky formátu úloh s výběrem odpovědí (Multiple Choice Questions, MCQ) se čtyřmi distraktory (správná 1 možnost). Otázkou (question) je míněno jakékoliv zadání v testové položce. Nejčastějším typem úloh s MCQ jsou v podobných testech užívány 2–4 distraktory. Naší volbou jsou 4 distraktory, protože dostatečně pokryjí možnosti výběru, grafické zpracování je kompaktní a po vizuální stránce působí přehledně a přátelsky vůči testovaným.

Položka je tvořena z kmene (stemu), kterým je otázka, neúplné tvrzení nebo úloha (např. Odpovězte zdvořile na následující sdělení. Následuje výběr 4 možností, z nichž 1 vyhovuje společenskému úzu.). Stemy jsou formulovány tak, aby byly jasné, stručné a aby uváděly všechny [pouze] relevantní údaje. Jsou obvykle delší než distraktory a slova, která by se opakovala v distraktorech, zpravidla uvádíme ve stemu.

Každá položka by měla testovat pouze jednu oblast, aby nedocházelo k tomu, že by zvolená položka, i když skrytě, měla dva významy. Stejně tak musíme dbát na jednoznačnost položky. Například:

She stopped …….. the suspect. at four o'clock.
a) interrogate b) to interrogate c) interrogation d) interrogating
1. Možnost: b) Zastavila se, aby vyslechla podezřelého.
2. Možnost: d) Skončila s výslechem podezřelého.

Aby položka byla jednoznačná třeba ve smyslu d), upravíme stem doplněním relevantní informace:
She stopped ……. the suspect because he had a solid alibi.

Dbáme na to, aby distraktory byly stručné, jednoznačné, přibližně stejně dlouhé a stejně atraktivní či funkční. Distraktor, který je zjevně nesprávný nebo není koherentní s testovaným učivem, nemá pro testování smysl a po moderaci položek je nahrazován. Pokud jsou tvůrci testů učitelé příslušného cizího jazyka, znají z praxe typické chyby studentů a je užitečné je zahrnovat do položek.

3.2 Shrnutí doporučení k tvorbě položek a distraktorů

1. Každá testovaná položka koresponduje se vzdělávacím cílem a učivem probíraným v kurzu.
2. Výběr odpovědí má jednu prokazatelně správnou odpověď.
3. V každé položce testujeme jeden jazykový prostředek, v případě dovedností jednu informaci.
4. Při výběru testové položky zohledňujeme praktičnost testovaného.
5. Distraktory jsou relevantní možnosti odpovědi i v jiných kontextech.
6. Distraktory zahrnují často se opakující chyby studentů.
7. Měníme náhodně pozici distraktorů mezi variantami a, b, c, d.
8. Délka nabízených možností je přibližně stejná.
9. Distraktory jsou samy o sobě možné, neodpovídají však zadání.
10. Distraktory, které si nikdo nevybírá, nahraďme.

3.3 Shrnutí doporučení k formě a procesu tvorby položek a distraktorů

1. Používejme formát správná odpověď a nejlepší odpověď. Tuto skutečnost uveďme v zadání.
2. Vyhraďme si dostatečný čas na přípravu a moderaci položek a distraktorů vícero učiteli.
3. Délka zadání položky by měla být přibližně stejná.
4. Střídejme pravidelně správné odpovědi mezi pozicemi a, b, c, d.
5. Pokud zadání položky obsahuje mezeru, umístěme ji ke konci zadání položky.
6. Používáme-li negativní možnosti, zdůrazněme zápor, např. NOT.
7. Hlavní informace položky se nachází v zadání.
8. Informace pro určení správné odpovědi je v zadání, netestujeme všeobecnou znalost.
9. Nepoužívejme špatné nebo neexistující gramatické struktury.
10. Položky testu jsou na sobě nezávislé, správná odpověď na jednu položku neovlivní další položky.

Poznámky:

Ad 1. Důležitost a obtížnost testovaného učiva (základní frekventované jevy úrovně B1 – B2)

Ad 2. Určení počtu a druhu položek (uzavřené multiple choice a True/False – po 1 svazku v poslechu a čtení s porozuměním)

Ad 3. Vytvoření svazku, příprava nahrávek, korektura, specifikace, přiřazení bodových hodnot, řešení

Ad 4. Stanovení časové dotace na svazek (poslechová cvičení se přehrávají 3x)

Ad 5. Analýza svazků položek (statistické měření – průměrný skór, obtížnost, reliabilita) a interpretace výsledků

Ad 6. Moderace distraktorů v položkách a rekonstrukce svazku

Při praktické realizaci tvorby položkové banky byly řešitelským týmem empiricky identifikovány problematické oblasti, ke kterým je nutno přihlížet při koncipování multiple-choice testu a výběru vhodnosti jejich použití:

1. Typově je tento druh testů neefektivní pro řešení problémově-orientovaných otázek vyžadujících schopnost organizace a vyjadřování myšlenek.
2. V reálných jazykových situacích respondenti nemají možnost vybírat správnou odpověď z předem dané nabídky.
3. Testy vykazují menší prostor pro zpětnou vazbu respondentů, protože není zaznamenán proces, jakým se respondent dopracoval k volbě chybné odpovědi.
4. Testy nejsou vhodné pro testování produktivních jazykových dovedností.
5. Testy mají velkou náročnost na examinátory při sestavování kvalitních položek a distraktorů a z toho plynoucí časová náročnost sestavování a moderování testů.
6. Umožňuje náhodný výběr správné odpovědi respondentem.

4 Konstruování testu na základě položkové banky

Vzhledem k tomu, že terminologie klasifikace testů není zcela jednotná, vymezujeme si pro naše potřeby jejich dělení takto:

4.1 Test průběžný („Achievement Test“)

Tento test aplikujeme na výuce, objektivní testové položky s vícenásobnou volbou odpovědí, z nichž pouze jedna je správná, jsou orientovány na učivo daného semestru. Testujeme, co se student/studentka skutečně naučili, ve srovnání s tím, čemu se podle sylabu a studijních materiálů naučit měli.

4.2 Test kontrolní („Progress Test“)

Tento nestandardizovaný test aplikujeme ve zkouškovém období druhého semestru, zjišťujeme, do jaké míry student/studentka probíraná odborná témata zvládli. Úspěšný/á student/studentka (70% správných odpovědí) za něj získávají zápočet.

4.3 Test závěrečný („Proficiency Test“)

Tento nestandardizovaný test je součástí zkoušky. Z hlediska učiva má průřezový charakter a zjišťuje, co student/ka skutečně v přítomné době znají a jak se rozvinuly jejich receptivní řečové dovednosti.

- Formy úloh: multiple choice questions (3 distraktory – 1 správná odpověď, za nesprávnou odpověď se body nestrhávají)
- Každá položka testuje pouze jednu oblast, aby nedocházelo k tomu, že by zvolená položka, i když skrytě, měla dva významy

5 Cílové skupiny, analýza učiva

Cílovou skupinou jsou studenti 1. a 2. ročníku kombinované a prezenční formy bakalářského studia programů bezpečnostně právního a bezpečnostního managementu. Každý z programů má zadaný doporučený studijní materiál, ze kterého tvůrci testů vybrali na základě analýzy učiva podstatné jevy nezbytné pro zvládnutí profesní terminologie a řečových dovedností na úrovni B1 SERR pro jazyky (Threshold). Nutno dodat, že při testování profesní slovní zásoby je požadovaná úroveň vyšší – B1+ až B2.

Aktuálně používané výukové materiály využívají nejpravděpodobnější témata a situace, které se mohou vyskytnout při vykonávání policejních činností a práce v oblasti veřejné správy. Obsahují lexiku, kterou lze podle kontextu obecně zařadit do okruhů uvedených v jednotlivých tématech položkové banky, jimiž jsou subbanky SECURITY AND LAW a PUBLIC ADMINISTRATION. V každé variantě testu vytvořeného na základě dané banky by měla být vyváženě zastoupena základní a typická slovní zásoba z aktuálně testovaných témat a gramatické jevy odpovídající aktuálně požadovanému stupni znalostí.

Na základě položkové banky jsme sestavili testy průběžné (typu „achievement“) a kontrolní, tak aby vyhovovaly zamýšlenému výstupu, aby měřily, co změřit mají a aby z funkčního hlediska převažovala funkce diagnostická. V případě průběžných testů můžeme změřit nejen úspěšnost jednotlivých studentů, ale též i úspěšnost pedagoga, tzn., jak studenty motivoval, kolik toho studenty naučil a (v našem případě) jak dobře test konstruoval jako jeho autor.

S využitím banky testových položek byly vypracovány dva druhy testů:

- **sumativní** – polytematický, použitý jako test u zkoušky B 71 po 4 semestrech studia a byl součástí hodnocení studenta,
- **průběžné** – mono nebo polytematické, v 1. a 3. semestru kombinované a prezenční formy bakalářského studia programů bezpečnostně právního a bezpečnostního managementu; tyto testy byly výchozím materiálem pro počítačové vyhodnocení. Slouží k hodnocení průběhu výuky, neslouží k hodnocení testantů.

Z hlediska interpretace výsledků se v následující části zabýváme testy pouze průběžnými, protože můžeme v tomto případě pracovat s relevantním množstvím položek a probandů.

6 Požadavky a cíle testování, základní metodická pravidla

Základními požadavky kladenými obecně na tyto průběžné testy jako na způsob měření jazykových dovedností studentů [1] jsou především:

- objektivita (didaktický test dovedností, které vyžadují od studentů řešit stejná zadání za stejných podmínek a stejné časové dotace a které jsou vyhodnocovány počítačovým programem nezávisle na osobě hodnotitele);
- reliabilita (homogennost a reprodukovatelnost)
- validita (zda test měří, co potřebujeme nebo chceme změřit – u jazykových testů jde zejména o porozumění poslechu či textu, pochopení i složitějších gramatických struktur)
- variabilita

Při vývoji testů i jejich jednotlivých úloh a vypracování zadání jsou dodržována základní metodická pravidla:

- metodická, didaktická a obsahová stránka testu musí být relevantní vůči zadání a účelu testu (studenti jsou již na začátku kurzu informováni o obsahové náplni testu, jeho formě, rozsahu pokrývajícím učební látku, typu úloh, časové dotaci, způsobu vyhodnocení, sdělení výsledků a zpětné vazbě)
- výstavba/konstrukt testu je validní a reliabilní (vychází ze specifikace položek);
- parametry zadání splňují požadavky na míru úspěšnosti testu [5]

7 Forma testů, zvolený druh testů, typy testů

Z praktických důvodů, zejména z důvodů finančních nároků na technické vybavení, pracujeme pro potřeby testování s kombinací PBT (papírově vyplňovaných testů – paper-based testing) a CBT (počítačově vyplňovaných testů – computer-based-testing) [4] s následující charakteristikou:

- PBT test s úlohami vybranými testerem z položkové banky
- záznamový list – odpověďník pro zaznamenání odpovědí
- oskenování, měření a vyhodnocení odpovědí počítačem

Týmová spolupráce založená na sdílení zkušeností při vytváření položek a úloh pro banku tříbí jejich kvalitu a navíc přináší jistou nadhodnotu důvěry studentů ve spravedlivou a objektivní evaluaci spíše než při konkurenčním přístupu jednotlivých testerů. Postupným vytvářením položkové banky lze následně pomocí různých kombinací v jednotlivých subtestech vytvořit poměrně velký počet různých testů se stejnou specifikací a konstruktem.

Podle stupně důkladnosti přípravy a ověřování testu a jeho vybavenosti je tento průběžný test na rozhraní standardizovaného a nestandardizovaného testu. Takový test je některými autory (např. Hambleton, Eignor, Rovinelli, 1986) označován jako kvazistandardizovaný, protože sice obsahuje standardy pro hodnocení výsledků, ale jeho standardizace není úplná [8].

Navíc úplná standardizace je projekt časově i finančně náročný, a proto v našich podmínkách vyhovuje tento „kvazistandardizovaný“ test, jenž ověřuje úroveň vědomostí studentů v několika paralelních skupinách na jedné škole, které se na test připravovali ve srovnatelných podmínkách ze stejných doporučených výukových materiálů. Za srovnatelných podmínek je test implementován a také evaluován.

Průběžné testy byly typově koncipovány jako achievement test ve sféře odborného jazyka i v gramatice. Achievement test (test jazykových schopností) je tvořený úlohami s výběrem odpovědí (Multiple Choice Test) a True/False a je navržen tak, aby měřil rozsah určitých aktuálních jazykových schopností studenta [6].

7.1 Subtesty, fáze přípravy testu

Průběžný test má tři subtesty – poslech s porozuměním (10 položek), čtení s porozuměním (10 položek) a lexikálně-gramatický (60 položek).

Subtesty poslech a čtení s porozuměním se skládají každý ze dvou svazků, z nichž každý obsahuje 5 položek typu true/false a 5 položek multiple choice, a subtest lexikálně-gramatický zahrnuje 6 svazků o 10 položkách. 4 svazky lexikálního subtestu se zaměřují na doplňování vhodných slov do vzájemně nesouvisejících vět, doplňování vhodných slov do souvislého textu, přiřazení správného termínu k dané definici a výběru správného překladu výrazu z Č do

A nebo obráceně. Gramatický subtest se skládá z jednoho svazku doplňování správných gramatických tvarů do vět a z jednoho svazku položek testujících výběr správné předložky (zejména vazebných nebo u nejfrekventovanějších frázových sloves).

Poslechová část tvoří 10 % testu, protože je zohledněna skutečnost, že studenti kombinovaného studia mají omezenou možnost řízeného nácviku poslechových dovedností a vyšší bodová dotace poslechu by mohla znehodnotit celkové bodové hodnocení.

Dovednosti čtení s porozuměním bylo též přisouzeno 10 %, a to z toho důvodu, že i v gramaticko-lexikálních testových cvičeních jsou kontextové úlohy vyžadující zvládnutí této dovednosti. Přesto na základě analýzy subtestu čtení s porozuměním v průběžných testech (viz část zprávy týkající se této problematiky) bude pravděpodobně váha tohoto subtestu přehodnocena.

Vzhledem k poměrně velkému objemu slovní zásoby, která má být testována, je žádoucí, aby selekce termínů a profesní lexiky pro úlohy byla velmi pečlivá a korespondovala s reálnými požadavky kladenými na policisty a pracovníky veřejné správy v praxi.

Standardizace testů a vytvoření položkové banky efektivně napomáhá flexibilnímu vytváření reliabilních a validních testů díky elektronickému formátu, který umožňuje průběžnou aktualizaci a dodatečné úpravy. V procesu tvorby testů je nezbytné prostřednictvím moderátora dbát na analýzu potřeb a zpětnou vazbu testantů, testerů a tvůrců testových úloh, která vede k eventuálním revizím. Personální zázemí testování nemůže být záležitostí jednoho člověka – tvorby testů se v našich podmínkách účastní nejen tvůrci testů, ale také moderátoři, administrátoři a hodnotitelé, jimiž jsou vyučující katedry jazyků. Moderátory mohou být též přizvaní kolegové ze zainteresovaných kateder pro posouzení odborné terminologie. Při vytváření položek pro všechny jazykové úrovně A2 – C1 by měli mít tvůrci vždy na zřeteli testování dovedností porozumění textu a poslechu s porozuměním, přičemž na vyšších úrovních by gramatické jevy měly být testovány z hlediska jejich vlivu na správnost obsahu sdělení.

8 Jakou procentuální úspěšnost vykazují studenti obou programů kombinované i prezenční formy v průběžném testu?

Procentuální úspěšnost je jedno ze základních kritérií, které používáme pro hodnocení výsledků testových výstupů respondentů. V rámci druhé otázky VÚ se blíže zabýváme porovnáním procentuálního hodnocení úspěšnosti pilotní sady testovaných skupin a dále interpretujeme naměřená data. Základní přehledová tabulka s procentuálním hodnocením sumarizuje výstupy provedené na pilotním vzorku pěti testů u studentů prezenčního a kombinovaného studia obou studijních programů. Výsledný datový výstup je průřezovým vzorkem testů administrovaných vždy na závěr jednoho ze semestrů čtyřsemestrového jazykového kurzu. Tabulka 1 sumarizuje výstupy závěrečných testů.

Tab. 1: *Základní přehledová tabulka s procentuálním hodnocením respondentů*

TEST	Celková průměrná úspěšnost (%)	Průměrná úspěšnost: Prezenčním studium (%) Průměrná úspěšnost: Kombinované studium (%)	Rozdíl průměrných procentuálních hodnot úspěšnosti (%)	Cohenovo *d*	Počet respondentů	Počet respondentů: Prezenční studium Počet respondentů: Kombinované studium
B60 1.sem.						
Listening	90	96 85	11	1,02	200	92
Reading	87	91 83	8	0,61		
Vocabulary	83	90 77	13	1,06		108
Grammar	82	89 76	13	0,85		
B60 3.sem.						
Listening	65	92 47	45	2,79	134	52
Reading	67	97 49	48	2,83		
Vocabulary	68	92 53	39	2,99		82
Grammar	52	80 34	46	2,67		
B71 1.sem.						
Listening	63	73 53	20	1,44	87	44
Reading	86	90 81	9	0,73		
Vocabulary	83	87 78	9	0,96		43
Grammar	71	77 64	13	0,86		
B71 3.sem.						
Listening	70	72 68	4	0,20	76	44
Reading	77	80 73	7	0,38		
Vocabulary	81	83 78	5	0,47		32
Grammar	69	72 64	8	0,55		
B71 4.sem.						
Listening	73	77 70	7	0,40	72	25
Reading	51	38 58	- 20	0,95		
Vocabulary	72	76 70	6	0,36		47
Grammar	65	63 66	-3	0,18		

Zdroj: Vlastní

9 Komparace procentuální úspěšnosti studentů prezenční a kombinované formy studia

Při sběru vzorků dat bylo našim primárním záměrem zajistit, aby každý z testů byl administrován v obou typech studia, a to s cílem porovnat výstupní procentuální úspěšnost respondentů, zjistit případnou korelační tendenci těchto dvou skupin a následně formulovat možné závislosti mezi procentní úspěšnosti sledovaných skupin.

Po analýze získaných dat můžeme konstatovat skutečnost, že studenti prezenčního studia dosahují lepších výsledků než studenti kombinovaného studia. Tato skutečnost může být vysvětlena několika faktory. Zaprvé je to rozdíl ve formě výuky jazyka v kurzech pro prezenční a kombinované studenty. Absence pravidelnosti a přímého kontaktu s vyučujícím zřejmě působí negativně na celkové testové procentuální skóre respondentů. Další faktory nejsou podchyceny daty z empirického šetření, ale na základě zkušeností členů řešitelského týmu zde může jistou roli hrát také profil průměrného studenta kombinovaného studia, který determinují jeho vstupní jazykovou připravenost a následnou jazykovou akviziční kapacitu v průběhu studia, tj. odstup od předchozí systematické jazykové přípravy, aktuální časové možnosti, prostor pro pravidelnost a systematičnost při jazykové samostatné přípravě.

Při všem výše řečeném data ukazují, že rozdíly v hodnotě procentní úspěšnosti se pohybují okolo 10 % s výjimkou skupiny subtestů B60 3. semestru. Zde zaznamenáváme markantní rozdíl, který neodpovídá měřením u zbývajících testů. Průměrná hodnota rozdílu zde činí alarmujících 44,5 % ve prospěch studentů prezenčního studia. Tato skutečnost je ještě podtržena faktem, že skupina stejného programu B60 1.sem. dosahuje standardního průměrného rozdílu 11,25 %. Jedním z vysvětlení může být relativně strmý nárůst probírané látky mezi 1. a 3. semestrem. Můžeme pak usuzovat na tendenci, že při zvýšeném množství a obtížnosti probíraného učiva dochází k většímu rozptylu skórů průměrné úspěšnosti mezi prezenčními a kombinovanými studenty. Pro větší objektivnost posuzování rozdílů mezi skupinami jsme dále aplikovali statistický princip měření velikosti účinku (effect size) vyjádřený korelačním koeficientem Cohenovo *d*, jehož velikost měří míru významnosti rozdílu mezi dvěma komparovanými skupinami. Z této stupnice je jasně patrné, že rozptyl procentní úspěšnosti pro B60 3.sem. podstatně převyšuje hodnotu 2, což je charakterizováno na Cohenově stupnici jako velký rozptyl [7].

Tuto skutečnost jsme dále chtěli ověřit na následující komparaci. Pokud je výše zmíněné pravda a vycházíme-li z předpokladu, že testy s vyšší průměrnou procentuální hodnotou úspěšnosti jsou lehčí než testy s nižší mírou procentuální úspěšnosti, můžeme předpokládat, že u testu, jehož průměrná hodnota procentuální úspěšnosti je vyšší (lehčí test), bude průměrný rozdíl mezi skupinou prezenčních a kombinovaných studentů menší a naopak. Jinými slovy vyšší obtížnost testů zvyšuje rozptyl hodnot průměrné procentuální úspěšnosti mezi skupinami.

Tab. 2: *Komparační tabulka průměrné procentuální úspěšnosti*

Test	Průměrná procentuální úspěšnost (%)	Rozdíl průměrných procentuálních hodnot úspěšnosti mezi prezenčním a kombinovaným studiem (%)	Počet respondentů
B60 1. semestr	85,50	11,25	200
B60 3. semestr	63,00	44,50	134
B71 1. semestr	75,75	12,75	87
B71 3. semestr	74,25	6,00	76
B71 4. semestr	65,25	- *	72

* Hodnotu nelze generovat, protože v tomto případě některé sub-testy vykazují lepší výsledky pro skupiny prezenčního studia a jiné pro skupiny kombinovaného studia. Celkově tento test pracuje s malým vzorkem respondentů.

Zdroj: Vlastní

Z tabulky 2 vidíme, že náš předpoklad se tímto potvrdil pouze částečně, a to na testech s největší a nejmenší mírou úspěšnosti (extrémy škály procentuální úspěšnosti). Na testech uprostřed této škály obtížnosti se tento předpoklad nepotvrdil. Je zde nutné ještě ale podotknout, že testy, které předpoklad potvrzují, byly administrovány na větším vzorku respondentů (200 resp. 134 respondentů), a tudíž by se mohlo očekávat, že jejich data budou mít větší vypovídací hodnotu než u testů s nižším vzorkem respondentů. Bylo by tedy zajímavé tuto skutečnost dále výzkumně prověřit při dalším experimentu.

9.1 Komparace procentuální úspěšnosti studentů programů bezpečnostně-právního (B60) a programu veřejná správa (B71)

Během empirického šetření a jeho interpretace se řešitelský tým také zabýval porovnáním úspěšnosti studentů programu bezpečnostně právního a programu veřejná správa, a to za účelem zjištění, zda výsledky jsou porovnatelné, a dále také z důvodu prozkoumání možnosti případného transferu dobré praxe mezi studijními programy [8].

Z výše uvedené tabulky 2 vyplývá, že průměrná úspěšnost testů studijního programu B71 tenduje k větší kompaktnosti než u studijního programu B60, který se naopak vyznačuje tendencí k polarizaci skórů procentuální úspěšnosti.

Po porovnání průměrné úspěšnosti obou studijních programů můžeme konstatovat, že

a) studijní program B71 dosahuje kompaktnějších průměrných hodnot úspěšnosti a celkově se úspěšnost ve všech testech obou studijních programů pohybuje v koridoru se středovou hodnotou 75 % s rozptylem cca ±10 %,

b) generované testy dosahují ve většině případů vyšších hodnot procentní úspěšnosti, než bylo zamýšleno.

Vzhledem k této skutečnosti je třeba aplikovat výše uvedené příklady dobré praxe při tvorbě nových položek a následně statisticky vyhodnotit a prověřit očekávaný posun v hodnotách procentní úspěšnosti do předem stanoveného pásma.

Závěr

Pokud jde o celkové průměrné procentuální hodnoty úspěšnosti testů, cílíme na rozpětí mezi 60 a 65%, které považujeme na základě naší praktické zkušenosti a studia odborné literatury za optimální. Zároveň také tato výsledková hladina odpovídá střední hodnotě bodové hodnotící škály aplikované pro atestaci výsledků zkoušek na PA ČR.

Z dat však vyplývá, že většina testů se nepohybuje v tomto hodnotovém pásmu vyjma testu B60 3. sem. Problematickým se jeví z tohoto pohledu příliš vysoké průměrné skóre úspěšnosti respondentů u většiny testů. Díky tomuto výzkumnému úkolu a provedené komparaci procentuální úspěšnosti testů musí řešitelský tým zkvalitnit sestavování budoucích testů, a to tak, že testy budou náročnější a také budou detailněji mapovat probrané učivo. Je také nasnadě se zamyslet nad rozsahem základního učiva a po analýze přistoupit k případnému navýšení či doplnění probíraných tematických celků.

Navrhovaná doporučení optimalizující přípravy testů vychází ze získaných analytických dat a praktické zkušenosti členů týmu. Tyto zásady pro tvorbu kvalitních položek testové banky, tzv. dobrá praxe, budou dále aplikovány v následujících testovacích obdobích obou studijních programů.

Poděkování

Prezentovaný výzkum byl realizován v rámci projektu "Optimalizace testování cizojazyčných kompetencí studentů kombinované formy studia v podmínkách PA ČR a možnosti IT při jeho řešení" na Pocejní akademii ČR v Praze pod vedením paní doc. PhDr. Věry Poláčkové, CSc.

Literatura

[1] ALDERSON J. Ch.; CLAPHAM C.; WALL D.: *Language Test Construction and Evaluation*. Cambridge University Press, 1995, 324 s. ISBN 978-0521478298.

[2] ALTE: *Handbuch zur Entwicklung und Durchführung von Sprachtests*. Europarat/Abteilung für Sprachenpolitik, Frankfurt am Main, 2012. ISBN: 978-3-86375-093-0.

[3] ALTE: *Manual for Language Test. Development and Examining. For use with the CEFR*. [online]. 2011. Available from WWW: http://www.coe.int/T/DG4/Linguistic/ManualLanguageTest-Alte2011_EN.pdf

[4] BACHMAN, L. F.; PALMER A. S.: *Language Testing in Practice: Designing and Developing Useful Language Tests*. 1. vyd. Oxford University Press, 1996, 377 s. ISBN 978-0-19-437148-3.

[5] BROWN, J. D.; HUDSON, T.: *Criterion-Referenced Language Testing*. The Cambridge Applied Linguistics Series. Cambridge University Press, 2008, 310 s, ISBN 978-0-521-80628-2.

[6] FLUCHER, G.: *Practical Language Testing*. Routledge, 2013, 343 s. ISBN13 978-340-98448-2.

[7] GREEN, A.: *Exploring Language Assessment and Testing*. Routledge Taylor Francis Group, 2014, 260 s. ISBN 978-1-415-88962-7.

[8] HAMBLETON, R. K.; ROVINELLI, R. J.: Assessing the Dimensionality of a Set of Test Items. *Applied Psychological Measurement*. 1986, Vol. 10, Issue 3, pp. 287–302. DOI: 10.1177/014662168601000307

[9] McNAMARA, T.: *Language Testing*. Oxford University Press, 2000, 156 s., ISBN 978-0-19-437222-0.

TEST CONSTRUCTION AND EMPIRIC RESULTS OF TESTING

In comparison to other domains of education, assessment in that of foreign language teaching has some specific features. Testing language skills is unique because language is not only the content but also the tool of measurement. The research team created a bank of items used for testing. This item bank comprises files of items which make it possible to create a great number of subtests. These include terms and professional vocabulary corresponding to real requirements put on employees in practice. Using the item bank, achievement and check tests were created. Thus, from the functional point of view, the diagnostic function prevails. Achievement tests are not only aimed at measuring the success of individual students but also at the success of the teacher. From the point of view of the interpretation of the results, achievement test are predominantly dealt with. These enable cooperation with a real number of items and tested students.

DIE ERSTELLUNG VON TESTEN UND DIE EMPIRISCHEN ERGEBNISSE DES TESTENS

Die Bewertung im Bereich der Aneignung, d. h. des Lernens einer Fremdsprache hat im Vergleich mit anderen Bereichen der Bildung ganz spezifische Züge. Das Testen von Sprachkenntnissen ist unter anderem deshalb einzigartig, weil die Sprache nicht nur Inhalt, sondern gleichzeitig auch Messwerkzeug ist. Für die Bedürfnisse des Testens wurde vom Team eine Bank mit Testaufgaben erstellt, welche Termini und professionelles Lexikon enthält. Diese korrespondieren mit den realen, in der Praxis an die Arbeitenden gestellten Anforderungen. Es wurden Achievement-Tests und Kontrolltests auf der Grundlage der Listenbank erstellt, so dass aus funktioneller Sicht die Bewertungsfunktion überwiegt. Die Achievement-Tests sind nicht nur auf das Testen des Erfolgs der einzelnen Studenten orientiert, sondern auch auf die Erfolgsquote der Lehrenden. Aus der Sicht der Interpretation der Ergebnisse gilt die Aufmerksamkeit den Achievement-Tests, welche das Zusammenwirken der realen Anzahl der Posten und der getesteten Studenten ermöglichen.

KONSTRUKCJA TESTÓW A EMPIRYCZNE WYNIKI EGZAMINÓW

Ocena w zakresie przyswajania lub uczenia się języków obcych ma, w porównaniu z innymi dziedzinami kształcenia, specyficzne cechy. Sprawdzanie znajomości języka jest unikatowe między innymi z tego względu, że język nie stanowi tylko treści, ale jest także narzędziem pomiaru. Zespół badawczy w celu sprawdzenia znajomości języka opracował bazę zadań egzaminacyjnych obejmującą pakiety pozycji, na bazie których można opracować dużą liczbę testów zawierających terminologię oraz leksykę specjalistyczną uwzględniającą rzeczywiste wymagania stawiane pracownikom w praktyce. Na podstawie bazy pozycji opracowano testy osiągnięć oraz testy kontrolne, a więc pod względem funkcjonalnym przeważa funkcja oceniająca. Celem testów osiągnięć nie jest tylko sprawdzenie wyników poszczególnych studentów, ale także dydaktyków. W ramach interpretacji wyników uwagę poświęcono testom osiągnięć, które umożliwiają współdziałanie realnych liczb pozycji a egzaminowanych studentów.