PRAKTICKÉ POSTŘEHY K VÝZNAMU STATISTICKÉ ANALÝZY PŘI TVORBĚ JAZYKOVÝCH TESTŮ

Jiřina Hrbáčková
*** Milan Boháček**

Masarykova univerzita
Ekonomicko-správní fakulta
Centrum jazykového vzdělávání
Lipová 41a, 602 00 Brno, Česká republika
jirina@econ.muni.cz
* bohacek@econ.muni.cz

Abstrakt

Tento příspěvek zkoumá informační přínos statistické analýzy jazykových testů Ekonomicko-správní fakulty Masarykovy univerzity, vytvořených pro úroveň C1 dle Evropského referenčního rámce, pro zvýšení validity a reliability těchto testů. Pro zpracování výstupů testů a získání hodnot průměru, obtížnosti testových položek, indexu diskriminace a Cronbachovo alfa s vyloučením dané položky bylo pro daný účel použito programu SPSS. Byla provedena analýza funkčnosti distraktorů pro úlohy typu výběr ze 4 možností. Testové položky nevyhovující stanoveným kritériím byly před novým kolem moderací změněny nebo upraveny. Očekáváme, že se statistická analýza, představující přínosný element při zkvalitňování testů a poskytující cennou zpětnou vazbu jejich tvůrcům, stane napříště nezbytnou součástí tvorby testů daného pracoviště.

Úvod

V průběhu akademického roku 2012/2013 bylo v rámci klíčové aktivity 4 – testování projektu IMPACT vytvořeno na oddělení jazyků Ekonomicko-správní fakulty Masarykovy univerzity (ESF) celkem 5 výstupních testů s předpokládanou výstupní úrovní studentů C1 dle SERR. Testy byly specifikovány jako tzv. pro-achievement testy, jedná se tedy o typ testů měřící jak celkovou úroveň jazykových kompetencí studentů, tak míru osvojení učiva probíraného v kurzech, kde složku achievement tvoří do velké míry odborná slovní zásoba a gramatické jevy typické pro obchodní angličtinu. V souvislosti s projektem, který podpořil snahy vyučujících jazykového centra Masarykovy univerzity (MU) o implementaci všech základních částí procesu vývoje testů a o jeho zkvalitnění, vyvstala řada otázek týkajících se validity, spolehlivosti (reliability) a zároveň praktičnosti těchto testů, tedy parametrů, které jsou pro objektivní a spolehlivé měření schopností studentů na základě daných testů naprosto nezbytné. [3] Pomíjíme otázku, do jaké míry tyto testy skutečně odpovídají dané úrovni rámce, protože odpověď na ni je otázkou samostatného výzkumu, a chceme vědět, zda testy objektivně a spolehlivě hodnotí ty schopnosti, které by podle specifikací daného testu testovat měly. V tomto příspěvku budeme hledat odpověď na otázku, jaké informace poskytla statistická analýza provedená na výstupech pretestace všech pěti testů vytvořených a administrovaných na ESF a jak lze některé z hodnot, které z ní vzešly, interpretovat a využít při revizi stávajících testů tak, aby se validita a reliabilita testů posílila a aby testy zároveň zůstaly praktické.

Statistická analýza výstupních jazykových testů vytvořených na jednotlivých fakultách Centra jazykového vzdělávání MU (CJV) je fází vývoje testů, která nemá dlouhou historii a do povědomí tvůrců testů, a zároveň vyučujících, se v hlubším pojetí dostala až díky projektu Impact, v jehož rámci proběhl na jaře 2013 týdenní seminář Rity Green, primárně věnovaný

statistické analýze. Dr. Rita Green působí na univerzitě v Lancasteru, kde se kromě jiného specializuje právě na statistickou analýzu jazykových testů. Pod jejím vedením se lektoři seznámili s hlavními parametry statistické analýzy v jazykovém testování, tzn., obtížností testových položek – tzv. facility value, diskriminačním koeficientem, tzv. Corrected Item Total Correlation, vnitřní konsistencí testů založené na hodnotách Cronbach's Alpha (CA) a Cronbach's Aphfa if Item Deleted (CAID). Dále byla nastíněna položková analýza distraktorů a deskriptivní statistika neboli grafické zobrazení výstupů testů. Při výběru náplně kurzu byly brány na zřetel potřeby oddělení CJV a praktičnost prováděné analýzy. Pro zpracování statistických dat souborů bylo použito programu SPSS. Dalším důležitým parametrům, jak např. korelaci, standardní odchylce nebo t-testu byla věnována marginální pozornost.

V této práci zkoumáme základní veličiny statistické analýzy a jejich hodnoty, které vzešly z pretestace na jaře 2013, tzn. průměru, spolehlivosti měřené pomocí ukazatelů Cronbach's Aplha a Cronbach's Aplha if Item Deleted, obtížnosti vybraných položek a položkovou analýzou konkrétního poslechového cvičení. Můžeme tvrdit, že test je statisticky spolehlivý? Pokud ne, jaké kroky můžeme podniknout ke zvýšení spolehlivosti a jak dané hodnoty interpretovat pro funkční revizi jednotlivých testových položek? Zároveň nás zajímá odpověď na otázku, jakou roli by měla statistická analýza při vývoji jazykových testů hodnotících jazykové kompetence studentů MU v budoucnosti hrát. V této práci nezohledňujeme samotný proces tvorby testů, který je pro spolehlivost testu zásadní, a věnujeme se spolehlivosti z hlediska výpovědní hodnoty statistické analýzy.

1 Parametry testů

Podle Douglase [1, str. 104] jsou jazykové testy v zásadě nespolehlivé nástroje měření, nicméně některé testy jsou spolehlivější než jiné. V případě CJV MU usilujeme o to, aby testy vytvořené za účelem hodnocení studentů MU byly spolehlivé a aby studentům různých fakult byly předkládány testy, o kterých lze tvrdit, že jsou validní a reliabilní. Ty části testu anglického jazyka, vytvořené pro studenty ekonomických oborů s očekávanou výstupní úrovní C1, které podléhají statistické analýze, se skládají celkem ze 60 položek, z nichž 20 tvoří poslechové položky – listening comprehension, 20 položek představuje gramaticko-lexikální část a zbývajících 20 porozumění textu, tzn. reading comprehension. V gramaticko-lexikální části bylo nutno pro účely statistické analýzy rozdělit 4. testovou úlohu, tzn. překlad dvouslovných odborných termínů, na dvě části 21a/21b – 25a/25b tak, aby bylo možno hodnotit každý výraz hodnotami 0-1 (dobře/špatně) zvlášť a zároveň aby analýza dobře zohledňovala schopnosti studentů. Testová úloha č. 5, tzn. výběr správného slovesa a jeho vložení do správného tvaru v daných větách, se naopak pro potřeby statistické analýzy musela zredukovat pouze na to, zda studenti vybrali správné sloveso či ne. Pro rozhodování dobře/špatně není možné spolehlivě a objektivně hodnotit dva jevy současně. Produktivní části testování jazykových kompetencí, psaný a mluvený projev, ze své podstaty statistické analýze nepodléhají.

Všemi 5 testy prošlo celkem 411 studentů. Počty studentů pro jednotlivé varianty testu zachycuje tabulka 1:

Tab. 1: *Počty studentů a průměrné dosažené hodnoty*

verze testu	**01**	**02**	**03**	**04**	**05**
počet studentů	84	89	87	73	78
průměrný počet bodů poslech	10,75	10,74	12,74	11,82	10,85
průměr gram.- lex. část	14,63	14,80	11,99	12,86	12,32
průměr čtení	12,44	12,02	12,55	12,14	14,27
Průměr test celkem	37,81	37,67	37,28	36,82	37,44

Zdroj: autoři

V tabulce 1 je rovněž zachycen průměrný počet bodů, kterého studenti dosáhli u každého z pěti testů a zároveň průměrný počet bodů, kterého dosáhli v jednotlivých kompetencích – tedy poslechu, gramaticko-lexikální části a v části čtení. Vzhledem k tomu, že varianty testu byly pretestovány na různých skupinách studentů, nelze z těchto hodnot jednoznačně usuzovat na paralelní validitu testů. Průměrný počet bodů všech tří částí jednotlivých testů, kterého studenti u všech testů dosáhli, nicméně vykazuje jen malé odchylky, z čehož lze usuzovat na podobnou obtížnost všech pěti verzí. Výsledek autoři považují za pozitivní indikátor snahy o jednotnost obtížnosti variant testu.

2 Měření spolehlivosti testů

Pokud bychom však chtěli mluvit o skutečné spolehlivosti testu, mohli bychom ji jednoznačně posoudit pouze tak, že stejné skupině studentů předložíme stejný test dvakrát [1]. Tím ale vyvstává logický problém, kdy při druhém průchodu testem lze očekávat lepší výsledky, neboť studenti jsou již s testovými úlohami do určité míry obeznámeni. Další možnost představuje metoda rozdělení testu na polovinu (split-half method) a jiné, které jsme však při analýze nepoužili. Pro zjednodušené posouzení spolehlivosti testu jsme zkoumali hodnoty Cronbachova alfa, veličiny, kterou jsme pro každý test získali zpracováním výstupů testů programem SPSS. Jedná se o veličinu, která nabývá hodnot od +1 do -1, přičemž záporné hodnoty, jichž je dosahováno zřídka [2, str. 38], poukazují na položky, které jsou špatně konstruované nebo z jiných důvodů chybné. Hodnota Cronbachova alfa je indikátorem vnitřní konzistence testu, vypovídá tedy o tom, zda test funguje jako celek a zda jeho jednotlivé položky hodnotí ten konstrukt, který hodnotit chceme. Podle Rity Green [2, str. 38] jsou položky nabývající hodnot mírně převyšujících 0,70 již akceptovatelné, ale žádoucí jsou hodnoty vyšší. Hodnoty 1 Cronbachovo alfa v podstatě nenabývá. [4] Vzhledem k charakteristice daných testových úloh a jejich rozmanitosti (pravdivá/nepravdivá tvrzení, výběr správné položky ze 4 možností, krátké volné odpovědi, překlad sousloví, doplňování správných tvarů) nepředpokládáme u jazykových testů vytvořených v reálném akademickém prostředí hodnoty běžně se blížící maximální hodnotě 1. Při analýze spolehlivosti testu je důležité porovnat hodnoty spolehlivosti na základě celkového Cronbachova alfa (CA) s jeho hodnotami pro jednotlivé položky testu vyjma dané položky samotné (Cronbach's Alpha If Item Deleted). Pokud tato hodnota (CAID) nabývá hodnoty nižší než celkové CA, pak tato položka přispívá ke spolehlivosti testu a při vynechání takové položky z testu by se celková hodnota CA snížila na danou hodnotu CAID. Zároveň tedy platí, že pokud je hodnota CAID vyšší, položka nepřispívá dostatečně ke spolehlivosti testu. Položky je však nutno zkoumat i na základě jejich obtížnosti a indexu diskriminace (viz dále). Tabulka 2 pak shrnuje hodnoty celkové spolehlivosti Cronbachova alfa pro všechny dané testy.

Tab. 2: *Cronbachovo alfa pro jednotlivé verze testů*

verze testu	**01**	**02**	**03**	**04**	**05**
Cronbachovo alfa	0,875	0,823	0,838	0,821	0,827

Zdroj: autoři

Z tabulky je patrné, že CA nabývá u daných testů hodnot od 0,821 do 0,875, přičemž podle výše uvedeného lze tvrdit, že se jedná o hodnoty akceptovatelné až pozitivní. U všech položek daných testů jsme porovnali hodnoty CAID (pro subtest poslechu viz Tab. 3) a CA. Položky, které by zjevně nepřispívaly ke spolehlivosti testu a zároveň nerozlišovaly mezi dobrými a horšími studenty nebo jejich obtížnost by byla diskutabilní, by byly přepracovány.

Tab. 3: *CAID pro jednotlivé položky části poslechu všech testových variant*

varianta testu	**01**	**02**	**03**	**04**	**05**
Item1	0,872	0,824	0,840	0,819	0,822
Item2	0,874	0,825	0,839	0,818	0,828
Item3	0,874	0,821	0,841	0,821	0,830
Item4	0,871	0,823	0,839	0,820	0,828
Item5	0,873	0,823	0,838	0,823	0,826
Item6	0,875	0,819	0,832	0,813	0,822
Item7	0,871	0,815	0,833	0,819	0,822
Item8	0,869	0,815	0,835	0,818	0,821
Item9	0,873	0,819	0,836	0,815	0,821
Item10	0,868	0,818	0,833	0,819	0,824
Item11	0,872	0,819	0,833	0,818	0,820
Item12	0,872	0,819	0,833	0,815	0,821
Item13	0,872	0,822	0,833	0,817	0,820
Item14	0,869	0,817	0,836	0,817	0,824
Item15	0,870	0,819	0,833	0,814	0,822
Item16	0,874	0,820	0,840	0,819	0,826
Item17	0,877	0,820	0,838	0,818	0,828
Item18	0,874	0,825	0,834	0,817	0,827
Item19	0,874	0,822	0,837	0,820	0,827
Item20	0,870	0,822	0,836	0,816	0,826

Zdroj: autoři

Z tabulky 3 vyplývá, že většina hodnot CAID poslechových položek se pohybuje kolem hodnoty celkové CA a ty, které nabývají hodnoty vyšší než celková CA, se liší v řádu setin. Jelikož se jedná o zanedbatelně malé hodnoty, nebyly na základě analýzy samotné spolehlivosti položky o vyšší hodnotě CAID než celková CA v této fázi eliminovány či změněny.

3 Obtížnost testových položek

Další důležitou veličinou statistické analýzy, kterou jsme podrobně prozkoumali, je obtížnost testových úloh neboli tzv. facility value (FV). Obtížnost nabývá hodnot od 0 do 1, kdy 0 znamená, že položku správně nikdo nevyřešil, zatímco 1 představuje 100 procentní úspěšnost. FV se dá jednoduše vyjádřit procentuálně. Pro testy zkoumající celkovou jazykovou schopnost testovaných (tzv. proficiency) považují testeři hodnoty FV v rozpětí 0,3 – 0,7 za takové, podle nichž lze posuzovat funkčnost testové položky (Green, 26). V našich podmínkách, kdy testujeme jak celkovou jazykovou úroveň studentů – proficiency (úrovně

dané Společným evropským referenčním rámcem), tak osvojení vyučované látky (achievement), očekáváme, že hodnoty FV budou dosahovat úrovně vyšší, nikoliv však pro testy administrované v řádných termínech – maximální. Tvůrci testů na ESF MU se shodli na hodnotách FV od 0,15 do 0,90 jako hodnotách pro dané účely akceptovatelné. V tabulce 4 je zachycen počet položek s extrémními hodnotami FV, tedy hodnotami FV 0 – 0,15 a 0,90 – 1 pro každý jednotlivý test. V těchto případech bylo třeba znovu projít testové otázky a zareagovat na daný výstup úpravou nebo úplným přepracováním položek. Při revizi byly brány v potaz i další ukazatele analýzy.

Tab. 4: *Počet položek s obtížností nižší než 0,15 a vyšší než 0,9*

varianta testu	**01**	**02**	**03**	**04**	**05**
Počet položek s FV 0,15 a nižší	2	2	2	2	1
Počet položek s FV 0,90 a vyšší	3	2	7	3	2

Zdroj: autoři

Všechny položky, které vykazovaly extrémní hodnoty FV byly podrobeny přezkoumání. U těch položek, u kterých byla hodnota FV 1 (jednalo se celkem o 2 položky) došlo k přepracování položky. Žádná položka nenabývala hodnoty FV 0. Ostatní položky s extrémní FV byly interpretovány společně s indexem diskriminace (tabulka 5), případně s analýzou distraktorů, pokud se jednalo o cvičení typu výběr ze 4 možností. Brali jsme ovšem v potaz i fakt, že test nebyl nasazen tzv. naostro a ty položky, které spadaly do kategorie achievement a měly velmi nízkou hodnotu FV, byly prozatím ponechány v původním znění. Po nasazení testu naostro budou tyto položky znovu podrobeny zkoumání. Předpokládáme, že administrace testu v reálné situaci povede ke zlepšení výsledků.

4 Index diskriminace

Index diskriminace (měřen jako Corrected Item-Total Correlation – CITC) poskytuje tvůrcům testu důležité a zajímavé informace. Zkoumá, jak si zkoušený vede u testu jako celku a porovnává, jak si vede u jednotlivých položek [2, str. 29]. Pokud je celkový výkon zkoušeného u testu dobrý, očekává se, že jednoduché, průměrné i některé těžké položky zodpoví správně. Je pravděpodobné, že některé těžké položky správně nezodpoví. Při splnění těchto podmínek nabývá hodnota diskriminačního indexu vyšších hodnot a lze tvrdit, že položka dobře diskriminuje, tedy rozlišuje mezi studenty s vyšší a nižší dosaženou jazykovou úrovní. Pokud zkoušený s celkově dobrým výsledkem nezvládne položky, o kterých jsme předpokládali, že jsou jednoduché, hodnoty diskriminačního indexu budou nízké a můžeme tvrdit, že položka málo nebo špatně diskriminuje. Podobně jako hodnoty Cronbachova alfa, také index diskriminace nabývá hodnot od -1 do +1. Pokud hodnoty diskriminačního koeficientu nabývají hodnot od +0,3 a vyšších, lze považovat položky za velmi dobře diskriminující. Hodnoty nižší nebo přímo záporné nejsou žádoucí a je třeba se na ně zaměřit.

Jak ukazuje tabulka 5, vyskytují se v testech z hlediska diskriminace problematické položky. Např. pro položku 06 verze testu č. 01 je hodnota diskriminačního koeficientu 0,000. Při dalším zkoumání položky zjistíme, že to je položka, u které FV nabývá hodnoty 1, tzn., všichni studenti ji odpověděli správně, a tudíž nerozlišuje mezi studenty s lepší a horší dosaženou jazykovou úrovní. Hodnota CAID je rovna hodnotě CA, z čehož lze usuzovat na to, že položka se nijak výrazně na vnitřní konsistenci testu nepodílí. Je to příklad položky, která by měla být úplně eliminována, pokud je naším úmyslem aplikovat výsledky statistické analýzy v praxi a posílit spolehlivost testu. Podobným příkladem jsou mimo jiné i položky č. 16, a č. 17 v testu č. 01, kde CICT nabývá hodnot 0,198, respektive 0,131. Rozdíl však vyvstává u hodnot FV, kdy pro položku č. 16 je tato proměnná rovna 0,93, ale pro položku č.

17 je to 0,51. V prvním případě položka nediskriminuje dostatečně, jelikož je velmi lehká. V druhém případě je naopak spíše těžší a dá se usuzovat, že studenti s lepší i horší dosaženou jazykovou úrovní s ní měli problém. Položku č. 16 upravíme, pokud budeme chtít zvýšit její náročnost nebo ji ponecháme v původním stavu, pokud přezkoumáním usoudíme, že položka je dobře konstruovaná a hodnota FV 0,93 je pro nás v tomto případě akceptovatelná. Jelikož jsou ale položky č. 16 a č. 17 zároveň položkami poslechového úkolu typu výběr správné varianty ze čtyř nabízených (MCQ), provedeme položkovou analýzu testové úlohy a zjistíme, zda a jak u této položky fungují správná odpověď a distraktory.

Tab. 5: *Hodnoty CITC pro jednotlivé položky poslechové části analyzovaných testů*

varianta testu	test 01	test 02	test 03	test 04	test 05
Item1	0,381	0,015	0,098	0,121	0,300
Item2	0,149	0,039	0,014	0,340	0,303
Item3	0,175	0,219	0,089	0,282	0,171
Item4	0,480	0,065	0,102	0,152	0,079
Item5	0,265	0,035	0,284	0,099	0,161
Item6	0,000	0,297	0,423	0,423	0,309
Item7	0,426	0,465	0,411	0,187	0,440
Item8	0,627	0,568	0,371	0,287	0,463
Item9	0,322	0,398	0,223	0,478	0,378
Item10	0,523	0,394	0,265	0,186	0,385
Item11	0,513	0,371	0,413	0,334	0,550
Item12	0,387	0,383	0,339	0,308	0,481
Item13	0,436	0,133	0,408	0,370	0,498
Item14	0,564	0,258	0,213	0,373	0,160
Item15	0,528	0,359	0,354	0,574	0,286
Item16	0,198	0,351	0,030	0,202	0,185
Item17	0,131	0,276	0,137	0,467	0,089
Item18	0,155	0,131	0,413	0,384	0,171
Item19	0,341	0,135	0,222	0,305	0,077
Item20	0,577	0,214	0,312	0,375	0,279

Zdroj: autoři

Pro položky č. 21 až 40 bylo konzultacemi se všemi tvůrci testů dohodnuto, že statistická analýza bude mít pouze informační hodnotu. Tato část je tvořena gramaticko-lexikálními položkami, které testují gramatické jevy a odbornou slovní zásobu probíranou v seminárních hodinách, jedná se tedy o tu část testu, kterou považujeme za achievement, neboť zvládnutí těchto jevů je jedním z cílů výuky jazykových kurzů na fakultách MU. Předpokládáme, že studenti přistupovali k testu s nadhledem, který se odrazil v nízké diskriminaci nebo vysoké obtížnosti některých položek.

5 Analýza distraktorů

Poslední, ale rozhodně ne nejméně zajímavou částí analýzy, kterou považujeme z hlediska informační hodnoty za velmi přínosnou, je položková analýza úloh typu výběru správné odpovědi ze čtyř nabízených možností. V daných testech se jedná o tři cvičení o celkem 19 položkách.

Tabulka č. 6 udává hodnoty položkové analýzy poslechového cvičení (položky 16 – 20) testu verze č. 01. V jednotlivých sloupcích je uveden počet studentů, kteří danou variantu (A, B, C nebo D) vybrali. Klíč označuje správnou odpověď. Z tabulky je patrné, že pro položku č. 16

jsou funkční pouze distraktory B a C. Můžeme usuzovat, že distraktor A je buď špatně konstruovaný, nebo příliš jednoduchý, a proto jej studenti vůbec nevolili. Z tohoto důvodu je nutné podrobit jej revizi. Pro položku č. 17 se naopak distraktor B zdá být nevhodně konstruovaný nebo se příliš blíží správné odpovědi, jelikož jej vybralo 32 studentů. Přihlédneme-li k indexu diskriminace, který je 0,24, můžeme tvrdit, že to pravděpodobně byli spíše studenti s nižší dosaženou jazykovou úrovní, kteří distraktor B volili. S hodnotami FV 0,51 a CAID 0,877 (CA=0,875) lze považovat položku za přijatelnou a ponechat ji v původním stavu, nicméně tomuto rozhodnutí musí předcházet zvážení formulace distraktoru B a snadnost distraktoru A. Výsledky pro položku č. 18 považujeme za uspokojující. Většina studentů zvolila položku správně a zároveň se všechny distraktory projevily jako funkční (CITC = 0,219, obtížnost je 0,81). Podobně lze interpretovat i výsledky položky č. 20. Zde navíc index diskriminace (0,524) potvrzuje, že studenti s vyšší dosaženou jazykovou úrovní vybírali správně, a tím prokazuje správnost konstruované položky. Naopak položka č. 19, kdy více studentů vybralo jako správnou odpověď distraktor B, se jeví jako problematičtější, navíc hodnota diskriminace je pouze 0,245. Tyto dva faktory společně poukazují na nutnost znovu věnovat položce pozornost a pokusit se o její přepracování.

Tab. 6: *Hodnoty analýzy distraktorů pro položky č. 16 až 20 testu č. 01*

test verze 01	**počet studentů, kteří vybrali jednu z možností**			
	A	**B**	**C**	**D**
položka č. 16	0	4	2	78 (klíč)
položka č. 17	2	32	8	42 (klíč)
položka č. 18	6	2	68 (klíč)	8
položka č. 19[1]	26 (klíč)	37	11	9
položka č. 20	11	51 (klíč)	15	7

Zdroj: autoři

Jak tedy z výše uvedeného vyplývá, poskytuje analýza distraktorů velmi užitečné informace a napomáhá k lepší konstrukci jak správných odpovědí, tak distraktorů, a tím i ke zvýšení celkové spolehlivosti testu. Položkovou analýzu jsme provedli na všech cvičeních daného typu u všech testových verzí a výsledky podrobně analyzovali. Jsme přesvědčeni, že analýza distraktorů významně přispěla k vylepšení kvality a spolehlivosti testu.

Otázkou zůstává, jakou roli bude mít statistická analýza testů v budoucnosti. Jedná se především o praktické hledisko získávání dat a jejich zpracování. Získávání dat je činnost ne příliš časově náročná, nicméně nad rámec povinností vyučujících na CJV. Řešením by bylo využití externích sil, pak ale vyvstává otázka nutnosti zabezpečení testů proti nechtěnému šíření a otázka finanční. Pokud se zamyslíme nad cíli celé koncepce testování v projektu IMPACT, bude záležet na tom, jak se na jednotlivých pracovištích podaří prosadit a udržet inovovaný proces tvorby testů. Pokud bude naším cílem tvořit testy s kvalitní výpovědní hodnotou, pak autoři článku vidí statistickou analýzu jako krok, který nelze ignorovat.

Závěr

V tomto příspěvku jsme se zabývali statistickou analýzou testů a její výpovědní hodnotou pro tvůrce jazykových testů na pracovišti poskytujícím výuku angličtiny vysokoškolským studentům se zaměřením na obchodní angličtinu. Kladli jsme si otázku, zda v kontextu výuky

[1] Pro položku č. 19 jeden student nezatrhl ani jednu variantu.

ESP poskytuje statistická analýza informace, které se dají využít pro posílení validity a reliability testu a zda by zpracování výstupů statistické analýzy mělo být nezbytnou součástí procesu tvorby testu na pracovišti daného charakteru.

Na první otázku lze po prozkoumání údajů, které statistická analýza přinesla, odpovědět kladně. V některých případech analýza poukázala na nesprávnost konstrukce testové položky, jindy zase buď potvrdila nebo vyvrátila naše očekávání o obtížnosti či snadnosti položky, a tím dala hmatatelnou zpětnou vazbu našim domněnkám. Jako jednoznačně přínosné a jednoduše interpretovatelné se ukázaly hodnoty reflektující funkci distraktorů ve cvičeních typu výběr správné položky z nabídky 4 možností. Kromě toho, že se tvůrci dozvěděli, jak jimi vytvořená nabídka odpovědí funguje, se zároveň domníváme, že pečlivé prozkoumání funkce distraktorů může vést ke schopnosti lépe konstruovat položky budoucích testů. Situace, kdy se úloha typu nabídka ze 4 možností zredukovala na dichotomický výběr mezi dvěma odpověďmi, nebyla výjimečná. Předpokládáme, že přepracování položek, jejich následná moderace, nová pretestace a opětovné statistické zpracování pomůže autorům získat mnohem hlubší vhled do fungování disktraktorů a povede ke kvalitněji tvořeným položkám.

Odpověď na druhou otázku je poněkud komplexnější a je třeba vidět celý kontext, při kterém testy na jazykovém centru vznikají. Pro jednotlivé týmy je práce na testech druhořadou záležitostí a její časová náročnost vede ke snaze eliminovat ty části procesu tvorby testu, které nejsou na první pohled naprosto nezbytné. Autoři tohoto příspěvku však mají za to, že pečlivé sledování dat, která přinášejí velmi zajímavou informaci o kvalitě testu, je důležité a nemělo by se opomíjet. Tým tvůrců testů anglického jazyka na ESF se shodl na začlenění statistické analýzy do tvorby testů jako na nezbytnosti. Závazkem a posláním jazykového centra MU je zodpovědný přístup k tvorbě testů jako validních a reliabilních nástrojů pro měření schopností studentů, ve kterých má statistická analýza, poskytující faktickou zpětnou vazbu, své nezastupitelné místo. Dodržením tohoto závazku se máme možnost posunout dále jako vyučující, jako testeři, i jako kvalitní a spolehlivé vzdělávací pracoviště.

Literatura

[1] DOUGLAS, Dan. *Understanding language testing*. London: Hodder Education, 2010, ix, 156 s. ISBN 9780340983430.

[2] GREEN, Rita: *Statistical Analyses for Language Testers*. Palgrave Macmillan, 2013, 309 s. ISBN: 978-1-137-01828

[3] *Manual for relating Language Examinations to the Common European Framework of Reference for Languages (CEFR).* [online]. Strasbourg: Council of Europe, Language Policy Division, January 2009, 200 s. [citováno 10. srpna 2013]. Dostupné na WWW: <http://www.coe.int/t/dg4/linguistic/Source/ManualRevision-proofread-FINAL_en.pdf>.

[4] *Analýza úloh.* [online]. www.SCIO.cz s.r.o., © 2008 – 2011. [citováno 15. srpna 2013]. Dostupné na WWW: <http://www.scio.cz/vyzkum/tvorba_testu/thodnoceni/analyza.asp>.

PRACTICAL OBSERVATIONS ON THE ROLE OF STATISTICAL ANALYSIS APPLIED TO LANGUAGE TEST DEVELOPMENT

This contribution is about the statistical analysis of tests developed at individual faculties of the Language Centre at the Masaryk University Brno. After conducting statistical analysis on sets of tests developed by the Faculty of Economics and Administration, the Faculty of Social Studies and the Faculty of Pedagogy, the argument for checking assumptions originally attributed to the tests, which had been liable to close scrutiny of teams of language teachers and pretested on students of the target population, turned out to be not only important but also hugely fascinating and worth in-depth investigation. While the analysis confirmed some of the original assumptions about individual test items as well as complete parts of tests, its evidence also shed light on those items calling for further revision and improvement. This paper will look at ways of how statistical analysis can help test developers make tests more reliable and at the same time alert us of pitfalls that we would not be aware of be it not for its insight.

PRAKTISCHE BEOBACHTUNGEN ZUR BEDEUTUNG DER STATISTISCHEN ANALYSE BEI DER ERSTELLUNG VON TESTS

Dieser Beitrag untersucht den informativen Nutzen der statistischen Analyse von Sprachtests der ökonomisch-juristischen Fakultät der Masaryk-Universität, die für das Niveau C1 nach dem europäischen Referenzrahmen erstellt wurden, um deren Validität und Reliabilität zu erhöhen. Zur Festlegung der Testgestaltung, der Erlangung eines Bewertungsdurchschnitts, des Schwierigkeitsgrades der Testaufgaben, des Diskriminationsindexes und des Cronbach-Alfa unter Ausschluss der gegebenen Position wurde zu diesem Zweck das Programm SPSS verwendet. Es wurde eine Analyse der Funktionalität der Distraktoren für Aufgaben des Typs „Auswahl aus vier Möglichkeiten“ durchgeführt. Die Testaufgaben, welche den festgelegten Kriterien nicht entsprechen, wurden vor der neuen Moderationsrunde geändert oder angepasst. Wir erwarten, dass die statistische Analyse, welche ein gewichtiges Element bei der Verbesserung der Teste darstellt und ihren Autoren eine wichtige Rückmeldung liefert, demnächst zu einem unerlässlichen Bestandteil der Testgestaltung am genannten Lehrstuhl wird.

PRAKTYCZNE SPOSTRZEŻENIA DOTYCZĄCE ZNACZENIA ANALIZY STATYSTYCZNEJ PODCZAS OPRACOWYWANIA TESTÓW JĘZYKOWYCH

W niniejszym artykule zbadano korzyści informacyjne wynikające z analizy statystycznej testów językowych Wydziału Ekonomiczno-Administracyjnego Uniwersytetu Masaryka, opracowanych dla poziomu C1 według CEFR, w celu zwiększenia stosowności i rzetelności tych testów. W celu opracowania wyników testów i pozyskania wartości średnich, ustalenia trudności zadań testowych, wskaźnika dyskryminacji oraz współczynnika rzetelności alfa (Alfa-Cronbacha) z wykluczeniem danego zadania zastosowano program SPSS. Przeprowadzono analizę funkcjonalności dystraktorów w zadaniach zamkniętych z wyborem z 4 odpowiedzi. Zadania testowe, które nie odpowiadały przyjętym kryteriom, zostały zmienione lub skorygowane przed ich kolejnym zastosowaniem. Oczekujemy, że analiza statystyczna, która jest przydatna w doskonaleniu testów i daje cenną informację zwrotną dla ich autorów, będzie następnym razem niezbędnym elementem w procesie tworzenia testów w danej placówce.