SCATE (Scio Computer Adaptive Test)
Nová role testu

LENKA FIŘTOVÁ[1,2], MARTIN HOLÍK[1]

Abstrakt: *Adaptivní testování na počítačích podporuje individualizaci, zajímavé a interaktivní typy úloh a výsledky mohou být použity k vytvoření učebního plánu na míru. Scio přistoupilo k vývoji adaptivních testů v roce 2011. Adaptivní test z anglického jazyka byl jedním z prvních. Úkolem bylo vytvořit množství úloh, které pokrývají všechny úrovně CEFR. Navíc jsme vytvářeli i úlohy pro úplné začátečníky, jež jsme označili úrovní A0. Úlohy byly zařazeny do kategorií a byly pilotovány na několika stovkách lidí, u kterých byla prokázána úroveň angličtiny (čerství držitelé certifikátů např. FCE, CAE, TOEFL). Test pak předkládá každému testovanému úlohy, které co nejvíce korespondují s jeho úrovní angličtiny. Základní princip je dát testovanému co nejtěžší úlohu, kterou je ještě schopen vyřešit. Na začátku dostane několik úloh různé obtížnosti pro přibližné zjištění úrovně. Další otázky jsou pak již vybírány na základě předchozích odpovědí. Pro vyhodnocení byla zvolena metoda klasifikátorů (MDT). Adaptivní testování umožňuje zadávat studentům pouze otázky, které jsou pro ně zajímavé. Test se stává nejen konečným hodnocením, ale i pomocníkem v učebním procesu.*

Klíčová slova: *adaptivní testování, Společný evropský referenční rámec pro jazyky, Measurement Decision Theory*

Adaptivní testování je ve společnosti Scio rozvíjeno od roku 2011, kdy byl vytvořen první adaptivní test s názvem SCATE (Scio Adaptive Computer Test) zaměřený na určení jazykové úrovně respondentů vzhledem ke škále definované Společným evropským referenčním rámcem[3]. Hlavní cílovou skupinou testu jsou žáci základních škol a studenti středních škol, byť jej mohou skládat i dospělí. Test byl uveden do škol v roce 2012 a byl následován dalšími adaptivními testy, které stejně jako SCATE využívaly principů Measurement Decision Theory, konkrétně testem informační gramotnosti (Gepard), testem čtenářské gramotnosti (Čtenář), testem Dovednosti pro život a v neposlední řadě německou variantou anglického SCATE.

[1] Scio, s.r.o., Pobřežní 34, 186 00 Praha 8
[2] Katedra ekonometrie, Fakulta informatiky a statistiky VŠE, nám. W. Churchilla 4, 130 67 Praha 3
[3] Společný referenční rámec (Common European Framework of Reference for Languages) definuje šest úrovní znalosti cizího jazyka: A1, A2, B1, B2, C1, C2. U každé úrovně je popsáno, jaké projevy by měl jedinec na dané úrovni vykazovat (například schopnost číst odborné články, schopnost porozumět jednoduchým nápisům apod.)

Obliba adaptivního testování a s ní související rychlý rozvoj tohoto typu testů je důsledkem nesporné řady výhod vyplývajících z adaptivního testování, které budou podrobněji rozebrány dále. Oproti těmto výhodám však stojí jednak určitá rizika adaptivních testů, jednak náročnost, kterou s sebou jejich vývoj přináší, zejména z hlediska know-how a programátorských dovedností.

Cílem následujícího textu je proto ve stručnosti přiblížit, co vývoj adaptivních testů obnáší, a seznámit čtenáře s přínosy a riziky jejich použití.

Vývoj testu

Vývoj v zásadě každého testu, SCATE nevyjímaje, sestává ze dvou hlavních fází, a to zaprvé ze samotné tvorby úloh a jejich úpravy do takové podoby, aby byly pokud možno co nejjednoznačnější a co nejlépe měřily zamýšlený konstrukt, zadruhé z odhadu jejich parametrů prostřednictvím jejich „odzkoušení“ na cílové skupině. Druhá fáze je nezbytná, neboť reálné fungování úloh se nezřídka výrazně liší od odhadu expertů, kteří úlohy tvoří.

Tvorba úloh

Již samotný vývoj úloh pro test SCATE probíhal v návaznosti na Evropský referenční rámec pro jazyky (dále CEFR z anglického Common European Framework of Reference). Ten definuje u každého jazyka šest úrovní znalostí, které označuje, od nejnižší po nejvyšší, A1, A2, B1, B2, C1 a C2, přičemž u maturantů se předpokládá úroveň B1. U každé úrovně je zároveň referenčním rámcem definováno, co by měl student již ovládat a jaké projevy by u něj mělo být možné vysledovat (Figueras, North, Takala, Verhelst & Van Avermaet, 2005).

Již během vývoje úloh byla tedy každá úloha na základě expertního posudku zařazena podle své obtížnosti do některé z CEFR úrovní, přičemž k výše zmíněným šesti úrovním byla přidána ještě úroveň A0, kterou byli označeni úplní začátečníci a která představuje nejnižší možnou úroveň, jíž je možno v testu dosáhnout.

Kromě samotné obtížnosti se jednotlivé úlohy v testu SCATE liší formální stránkou a zaměřením. Po formální stránce bylo vyvinuto několik různých typů úloh, mezi nimi například klasické multiple choice úlohy, seřazovací úlohy testující schopnost skládání konverzací či částí věty do správného pořadí, přiřazovací úlohy testující schopnost přiřadit související položky jako například popis a obrázek, doplňovačky určené pro otevřené úlohy a několik dalších. Cílem bylo jednak co nejlépe pokrýt širší spektrum dovedností, neboť například pro testování pravopisu nejsou klasické multiple choice úlohy příliš vhodné, jednak vytvořit test, který bude pro respondenty zajímavý a který udrží jejich pozornost.

Z hlediska zaměření úloh byly vytvořeny dvě hlavní kategorie, čtení a poslech, a u každé z nich pak tatáž množina podkategorií, a to slovní zásoba (vocabulary), gramatika

a pravopis (grammar/spelling), konverzace (conversation) a v neposlední řadě úlohy na porozumění (gist/detail).

Vývoj každé úlohy začal jejím vytvořením autorem, který úzce spolupracoval s garantem. Následovalo její posouzení několika oponenty a nakonec byla předložena tzv. pretestantům, tedy dětem z cílové skupiny s dovednostmi přibližně na takové úrovni, aby byly úlohu schopny vyřešit. Cílem tohoto procesu bylo odstranit z úloh veškeré možné nejasnosti a nejednoznačnosti. Je nutno poznamenat, že v některých případech to byly skutečně až samotné děti, které odhalily, že úloha není zcela jednoznačná a že si ji lze vyložit, a potažmo i vyřešit, různými způsoby. Je proto vhodné při vývoji testu skutečně nepodceňovat význam pretestace.

Výsledkem popsaného procesu byla banka obsahově kvalitních úloh, u nichž však stále chyběla informace o tom, jak se reálně chovají v jednotlivých CEFR úrovních. U úloh zaměřených na gramatiku může být relativně snadné posoudit, na které úrovni lze znalost testované dovednosti očekávat, avšak u úloh zaměřených na porozumění či konverzační dovednosti představuje expertní zařazení skutečně pouze prvotní odhad obtížnosti úlohy, a odhad jejich parametrů prostřednictvím jejich předložení cílové skupině je zde tedy zcela nezbytným krokem.

Kalibrace úloh

Druhou fázi vývoje představovala pilotáž úloh, kdy za pilotanty byly vybrány děti, u nichž byla CEFR úroveň již známá z předchozího testování. Šlo tedy o děti, které v krátké době před pilotováním složily některou z mezinárodně uznávaných jazykových zkoušek, jako jsou například TOEFL či FCE.

Tím bylo možné u každé úlohy určit její obtížnost vzhledem k jednotlivým CEFR úrovním. Obtížností úlohy se v tomto kontextu myslí hrubá úspěšnost, tedy podíl respondentů z dané kategorie, který úlohu vyřešil správně. Dále bylo možné u každé úlohy určit její diskriminační schopnost vzhledem ke každé dvojici sousedních CEFR úrovní, tedy její schopnost rozlišit mezi respondenty sousedních úrovní.

Výstupem druhé fáze byla množina úloh se známými parametry, které již tedy bylo možné použít v ostrém testování za účelem zařazení respondentů do odpovídající CEFR úrovně. Algoritmus tohoto zařazování je popsán dále.

Popis algoritmu testu SCATE

Přestože většina adaptivních testů vychází z teorie odpovědi na položku (IRT), pro test SCATE to neplatí. Algoritmus testu SCATE vychází z Measurement Decision Theory (dále MDT), kterou uvedl Rudner (2009) a která je vhodná právě v případě, kdy je cílem testu rozdělit respondenty do několika předem definovaných kategorií. V oblasti psychometrie se této teorii nedostává takové pozornosti jako IRT, přesto má však několik nesporných výhod, mezi něž patří zejména ve srovnání s IRT výpočetní jednoduchost a také menší náročnost na počet respondentů. Jediným klíčovým

předpokladem této teorie je vzájemná nezávislost úloh. Tato teorie bude nejprve stručně vysvětlena, načež bude podrobněji popsána její konkrétní aplikace v testu SCATE.

Measurement Decision Theory

MDT vychází z Bayesovy věty, která vyjadřuje pravděpodobnost, že nastane jev *A*, za podmínky, že nastal jev *B*, následujícím vztahem:

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

kde $P(A)$ resp. $P(B)$ jsou pravděpodobnosti jevů *A* resp. *B*, $P(A|B)$ je pravděpodobnost jevu *A* podmíněná výskytem jevu *B* a $P(B|A)$ je pravděpodobnost jevu *B* podmíněná výskytem jevu *A*. V našem případě se jevem *A* rozumí fakt, že respondent patří do určité CEFR úrovně, jevem *B* pak určitý vektor odpovědí na předložené úlohy.

Označme **M** množinu *k* různých kategorií, do jedné z nichž má být respondent zařazen. V našem případě se jedná o sedmiprvkovou množinu, kde každý prvek množiny představuje jednu z CEFR úrovní použitých v testu SCATE (A0, A1, A2, B1, B2, C1, C2). Pro každou kategorii $m_j \in \mathbf{M}$ označme $P(m_j)$ pravděpodobnost, že respondent patří do kategorie m_j.

Dále označme **z** *n*-složkový vektor odpovědí, představující odpovědi na každou z *n* úloh, na základě nichž je respondent zařazen do jedné z kategorií, přičemž *i*-tá složka vektoru **z**, představující odpověď na *i*-tou úlohu, nabývá hodnoty 1, zodpověděl-li ji respondent správně, a hodnoty 0, zodpověděl-li ji nesprávně.

Označme $P(u_i \mid m_j)$ pravděpodobnost správné odpovědi na *i*-tou úlohu za předpokladu, že respondent patří do kategorie m_j.

Pak lze pravděpodobnost, že u respondenta z kategorie m_j vypozorujeme vektor odpovědí **z**, vyjádřit jako

$$P(\mathbf{z}\,|m_j) = \prod_{i=1}^{n} P(u_i\,|m_j)^{z_i}\left[1 - P(u_i\,|m_j)^{(1-z_i)}\right].$$

S využitím právě zavedeného značení lze obecné vyjádření Baysovy věty výše přepsat jako

$$P(m_j|\mathbf{z}) = \frac{P(\mathbf{z}\,|m_j) \cdot P(m_j)}{\sum_{j=1}^{k} P(\mathbf{z}\,|m_j) \cdot P(m_j)}.$$

Tento vztah vyjadřuje pravděpodobnost, že testovaný patří do kategorie m_j za předpokladu, že jsme u něj pozorovali vektor odpovědí **z**. V čitateli je součin $P(\mathbf{z}\,|m_j)$, tedy pravděpodobnosti, že u respondenta napozorujeme vektor odpovědí **z** za podmínky, že patří do kategorie m_j, a $P(m_j)$, tedy pravděpodobnosti, že náhodně

vybraný respondent patří do kategorie m_j. Jmenovatele tvoří tzv. normalizační konstanta, která zajistí, že součet pravděpodobností, že respondent spadá do *j*-té kategorie, přes všech *k* kategorií bude roven jedné. Tato normalizační konstanta zároveň představuje pravděpodobnost, že u náhodně vybraného respondenta napozorujeme vektor odpovědí **z**.

Ze vztahu výše je zřejmé, že k výpočtu jsou potřeba tři vstupní informace. Zaprvé jsou to pravděpodobnosti, že respondenti jednotlivých kategorií zodpovědí správně příslušné úlohy. Tyto pravděpodobnosti lze získat z pilotáže tak, jak bylo popsáno výše. Zadruhé je nutné znát, jaká je pravděpodobnost, že náhodně vybraný respondent patří do jednotlivých kategorií. Tuto informaci je možno získat rovněž z pilotáže, tvoří-li ji reprezentativní vzorek cílové populace, nebo na základě externích statistik, případně lze předpokládat rovnoměrné rozložení kategorií v populaci. Třetí vstupní informací je vektor odpovědí **z**, který získáme při samotném testování.

Následně spočítáme pravděpodobnost $P(m_j|\mathbf{z})$ pro všechny kategorie a respondenta zařadíme do kategorie, pro niž je tato pravděpodobnost nejvyšší.

Pro ilustraci uvažujme situaci, kdy máme pouze dvě kategorie, *X* (respondent zvládá učivo dostatečně) a *Y* (respondent nezvládá učivo dostatečně) a tři různé úlohy. Předpokládejme, že respondent z kategorie *X* zodpoví úlohu 1 správně s pravděpodobností 0,8, úlohu 2 s pravděpodobností 0,6 a úlohu 3 s pravděpodobností 0,7. Dále předpokládejme, že respondent z kategorie *Y* zodpoví úlohu 1 správně s pravděpodobností 0,4, úlohu 2 s pravděpodobností 0,5 a úlohu 3 s pravděpodobností 0,1. Dále víme, že respondenti kategorie *X* tvoří 25 % populace. Nyní uvažujme respondenta, u kterého jsme pozorovali vektor odpovědí **z** = (1, 1, 0). Pravděpodobnost, že tento respondent patří do kategorie *X* resp. *Y*, je pak rovna

$$P(X|\mathbf{z}) = \frac{(0{,}8 \cdot 0{,}6 \cdot 0{,}3) \cdot 0{,}25}{(0{,}8 \cdot 0{,}6 \cdot 0{,}3) \cdot 0{,}25 + (0{,}4 \cdot 0{,}5 \cdot 0{,}9) \cdot 0{,}75} = 0{,}21,$$

resp.

$$P(Y|\mathbf{z}) = \frac{(0{,}4 \cdot 0{,}5 \cdot 0{,}9) \cdot 0{,}75}{(0{,}8 \cdot 0{,}6 \cdot 0{,}3) \cdot 0{,}25 + (0{,}4 \cdot 0{,}5 \cdot 0{,}9) \cdot 0{,}75} = 0{,}79.$$

Respondent by tedy byl zařazen do kategorie *Y*.

Aplikace MDT v testu SCATE

Algoritmus testu SCATE se skládá ze tří základních fází: fáze inicializace (předběžného odhadu CEFR úrovně respondenta), fáze adaptivního předkládání úloh a fáze finálního zařazení respondenta do některé z CEFR úrovní.

Fáze inicializace

Fáze inicializace spočívá v tom, že je respondentovi předloženo 8 úloh, a to v takovém složení, aby byly v této fázi obsaženy úlohy na čtení i poslech a zároveň pokrývající různé dovednosti (slovní zásobu, gramatiku, konverzaci i porozumění).

Poté by pro každou ze sedmi CEFR úrovní (tedy pro každou z kategorií m_j, kde j = 1, …, 7) mohla být spočítána pravděpodobnost

$$P(m_j|\mathbf{z}) = \frac{P(\mathbf{z}\,|m_j) \cdot P(m_j)}{\sum_{j=1}^{7} P(\mathbf{z}\,|m_j) \cdot P(m_j)}.$$

U testu SCATE se ovšem předpokládá rovnoměrné rozložení CEFR úrovní v populaci (není totiž známo, z jaké konkrétní skupiny jedinec pochází, zda je to žák základní školy, maturant či dospělý člověk, a proto je vhodnější apriorní předpoklad o rozložení kategorií nepoužívat). Proto lze vztah zjednodušit na

$$P(m_j|\mathbf{z}) = \frac{P(\mathbf{z}\,|m_j)}{\sum_{j=1}^{7} P(\mathbf{z}\,|m_j)}.$$

Jmenovatel zlomku je pouze normalizační konstanta, zajišťující, že součet pravděpodobností přes všechny CEFR úrovně bude roven 1. Při rozhodování o tom, jaké úlohy respondentovi předkládat dále, však postačí pracovat s čitatelem zlomku. Spočítá se tedy pouze

$$P(\mathbf{z}\,|m_j) = \prod_{i=1}^{8} P(u_i\,|m_j)^{z_i}\left[1 - P(u_i\,|m_j)^{(1-z_i)}\right].$$

Následně se vybere ta CEFR úroveň, pro niž je tato hodnota nejvyšší, tedy úroveň, kam spadá respondent s největší pravděpodobností (ačkoli bez provedení normalizace nelze hovořit přímo o pravděpodobnosti jako takové), a ta ze dvou sousedních CEFR úrovní, pro niž je tato hodnota vyšší.

Fáze adaptivního předkládání dalších úloh

Vstupní informací druhé fáze jsou hodnoty $P(\mathbf{z}\,|m_j)$ pro každou z j CEFR úrovní, které označme P_j^{init} a s tím související určení toho, do které z úrovní respondent nejpravděpodobněji spadá a do které ze dvou sousedních úrovní spadá spíše. Nyní je respondentovi předloženo postupně 24 úloh, tato fáze tedy sestává z t = 1, 2, … , 24 kroků. Přitom je ošetřeno, aby úlohy opět odpovídaly předem definované specifikaci testu, která určuje podíl úloh na slovní zásobu, konverzaci, gramatiku a porozumění (u čtení i poslechu) tak, aby byly veškeré tyto dovednosti testem pokryty.

Veškeré úlohy, které jsou v bance úloh k dispozici, s sebou nesou informaci o tom, jak dobře diskriminují mezi každými dvěma sousedními CEFR úrovněmi. Pro každou dvojici sousedních CEFR úrovní tak existuje množina úloh na slovní zásobu, konverzaci,

gramatiku i porozumění (u čtení i poslechu), které dostatečně dobře diskriminují mezi těmito dvěma úrovněmi.

V každém kroku je pak jedna z těchto úloh náhodně vybrána, a to na základě cílové specifikace testu a své diskriminační schopnosti, přičemž se bere v potaz diskriminační schopnost mezi těmi dvěma sousedními CEFR úrovněmi, které jsou v daném kroku relevantní (vyhodnoceny jako nejpravděpodobnější). Tato úloha je předložena respondentovi. Označme pravděpodobnost, že respondent *j*-té kategorie odpoví na úlohu předloženou v kroku t správně, jako $P(u_t|m_j)$. Odpověď respondenta na tuto úlohu označme z_t, kde z_t = 1, zodpověděl-li ji respondent správně, jinak 0. Pak

$$P_j^t = P_j^{init} \cdot P(u_t\,|m_j)^{z_t}\left[1 - P(u_t|m_j)^{(1-z_t)}\right], \text{kde } t = 1,$$

přičemž P_j^t označuje pravděpodobnost, že respondent patří v kroku t do kategorie *j*. V každém z dalších t = 2, 3, … 24 kroků postupujeme analogicky, tedy:

$$P_j^t = P_j^{t-1} \cdot P(u_t\,|m_j)^{z_t}\left[1 - P(u_t|m_j)^{(1-z_t)}\right], \text{kde } t = 2, 3, \ldots, 24.$$

Je zřejmé, že dvojice sousedních CEFR úrovní, do nichž respondent nejpravděpodobněji patří, se může v průběhu fáze adaptivního předkládání úloh postupně měnit.

Poté, co bylo respondentovi předloženo 32 úloh (8 v inicializační fázi a 24 v adaptivní fázi), přichází fáze poslední.

Fáze finálního zařazení respondenta

Poté, co respondent vyřešil 32 úloh, spočítáme, s jakou pravděpodobností patří do jednotlivých CEFR úrovní. Pro t = 24 označme tedy P_j^t jako P_j^{final}. Pak pravděpodobnost, že respondent patří do kategorie *j* (kde *j* = 1, …, 7, neboť pracujeme se sedmi CEFR úrovněmi) spočítáme jako

$$\frac{P_j^{final}}{\sum_{j=1}^{7} P_j^{final}}.$$

Ověříme, jestli existuje CEFR úroveň, do níž respondent patří alespoň s 96% pravděpodobností, tedy jestli pro některou CEFR úroveň dosahuje výše uvedený podíl hodnoty alespoň 0,96. Pokud ano, je do této CEFR úrovně respondent zařazen. Pokud ne, je mu předložena další úloha a postupuje se analogicky jako doposud, a to dokud nenastane jedna ze dvou možností: buď po předložení některé úlohy dosáhne pro jednu z CEFR úrovní pravděpodobnost, že do ní respondent patří, hodnoty alespoň 0,96, nebo respondent zodpoví 40 úloh, aniž by takto vysoké hodnoty bylo v některé z CEFR úrovní dosaženo. Ve druhém případě je algoritmus ukončen a respondent je zařazen do té CEFR úrovně, pro niž je tato pravděpodobnost nejvyšší.

Poznámky k algoritmu SCATE

Na závěr je potřeba uvést ještě několik poznámek. Přestože se v algoritmu testu SCATE pracuje i s úrovní C2, úloh na této úrovni je v bance relativně málo, a test má proto svá omezení v rozlišování mezi úrovněmi C1 a C2. Navíc SCATE testuje pasivní znalosti jazyka, což pro přesné určení mezi úrovněmi C1 a C2 nemusí být již postačující. Proto i respondenti, u nichž algoritmus vyhodnotí, že patří do úrovně C2, obdrží jako výsledek úroveň C1 jakožto oficiálně možný maximální výsledek.

Za druhé, maximální počet úloh, na základě nichž může být respondent zařazen, je 40 (8 inicializačních a maximálně 32 adaptivně předkládaných). Do každého testu je ovšem zařazeno ještě 5 úloh bez statistik, které na zařazení respondenta nemají vliv. Tímto způsobem se provádí kalibrace nových úloh, a tak se postupně rozšiřuje stávající banka úloh.

V neposlední řadě, kromě celkového zařazení do jedné z CEFR úrovní se obdobně počítá i nejpravděpodobnější CEFR úroveň zvlášť pro poslech a čtení, která je také obsažena ve výsledné zprávě pro respondenta.

Výhody a rizika adaptivního testování

K rychlému rozvoji adaptivního testování přispívá mimo jiné to, že s sebou tato forma testování přináší řadu výhod, jak popisuje například van der Linden (2008) či Wainer, Dorans, Flaugher, Green & Mislevy (2000). Je na místě zmínit alespoň některé z nich. V prvé řadě lze adaptivním testem dosáhnout obdobné přesnosti při určení úrovně dovedností respondenta jako u papírového testu, avšak s menším počtem úloh, neboť respondentům jsou předkládány úlohy tak, aby přinesly co nejvíce informací právě o jejich úrovni dovedností, zatímco úlohy, které pro ně nejsou relevantní (jsou příliš lehké či příliš těžké) jim předloženy nejsou. Respondenti jsou navíc při řešení adaptivního testu většinou více motivovaní a snáze udrží pozornost, protože test je jim „šit na míru“, takže nejsou odrazeni příliš těžkými úlohami, nebo znuděni úlohami příliš lehkými. Další výhodou z čistě praktického hlediska je snížení finančních a časových nákladů spojených s administrací testu a snazší zabránění opisování a podvádění díky tomu, že každý respondent má svůj unikátní test.

Existují však samozřejmě i určité nevýhody adaptivního testování, mezi nimi například fakt, že je nutné vybudovat dostatečně velkou banku obsahově kvalitních, zkalibrovaných úloh, což může být nákladné. Navíc je potřeba příslušné procedury potřebné pro adaptivní testování naprogramovat, což bývá náročné na know-how a finanční zdroje, a je rovněž nutné opřít algoritmus o některou z vhodných teorií, například MDT či IRT, každá z nichž má svá omezení (například MDT se hodí pro rozřazování respondentů do kategorií, což však nemusí být vždy cílem, IRT naopak zase vyžaduje splnění řady předpokladů, mezi nimiž je například předpoklad unidimenzionality měřeného latentního rysu, tedy předpoklad, aby test měřil právě jednu dovednost; existují sice i multidimenzionální IRT modely, ty ale také nemusí být

vhodné vždy a nesou s sebou další potenciální komplikace[4]). Jednou ze zásadních nevýhod adaptivního testování a brzdou rozšíření jeho použití u high-stake testů je fakt, že nelze přímo porovnat dva testované, protože každý má svou vlastní sadu úloh. Jejich výsledek je sice srovnatelný, avšak je obtížné respondentům jakožto laikům srozumitelně objasnit, jak se k takovému výsledku dospělo a jak je možné, že mají například horší výsledek než jiný respondent, který přitom na větší podíl úloh odpověděl nesprávně, což je při adaptivním testování poměrně běžná situace. To může ztěžovat použití adaptivního testování při přijímacím řízení jakéhokoli druhu, kde existuje velký tlak na transparentnost a na co nejlepší srovnatelnost výsledků.

Vybraná zjištění

Na závěr uveďme ještě vybraná zjištění získaná díky dosavadnímu použití testu SCATE, a to konkrétně rozložení CEFR úrovní mezi studenty 4. ročníku středních škol. Následující statistiky se týkají celkem 5 955 respondentů, kteří skládali test SCATE během podzimního testování v letech 2012 až 2014, a to napříč všemi typy škol. Tabulka uvádí procentuální zastoupení jednotlivých CEFR úrovní v této skupině.

Tabulka 1: Rozložení CEFR úrovní mezi studenty 4. ročníku SŠ

A0	A1	A2	B1	B2	C1
3%	9%	27%	36%	19%	6%

Z tabulky je patrné, že pouze okolo 60 % respondentů splňuje požadavek na maturitní zkoušku, podle níž má být maturant nejméně na úrovni B1. Je samozřejmě otázkou, nakolik jsou výše uvedené statistiky reprezentativní, neboť by bylo možné předpokládat, že o test SCATE mají zájem především školy, které kladou na výuku angličtiny určitý důraz. Je proto možné, že napříč celou populací 4. ročníků SŠ by byly výsledky horší, což ovšem, na druhou stranu, není plně v souladu s relativně dobrými výsledky žáků u státních maturit, které lze pozorovat v posledních letech: například v roce 2015 neuspělo v didaktickém testu z anglického jazyka, který je na úrovni B1, pouhých 6,3 % studentů (Centrum pro zjišťování výsledků vzdělávání, 2015).

Závěr

Cílem článku bylo na příkladu testu SCATE přiblížit čtenáři oblast adaptivního testování, zejména ukázat, co vývoj adaptivního testu obnáší. Jelikož s sebou adaptivní testování nese mnoho výhod, lze očekávat, že jeho využití bude společně s rozšiřováním výpočetní techniky na školách stoupat. Otázkou však zůstává využitelnost adaptivních testů u přijímacího řízení. V současné době se nejeví jako příliš pravděpodobné, že by mělo v kontextu přijímacích zkoušek dojít k přechodu od papírových testů k adaptivním, a to

[4] Kromě větší složitosti oproti unidimenzionálním IRT modelům mohou multidimenzionální modely přinášet problémy čistě praktického rázu, a to ve chvíli, kdy je potřeba jedinci dát právě jeden výsledek testu, čili jedno jediné číslo, které co nejlépe shrnuje jeho výkon v daném testu.

zejména kvůli vysokému tlaku na transparentnost výsledků v této oblasti, která je u adaptivních testů ve srovnání s papírovými testy problematičtější. Na druhou stranu, do budoucna nelze posun od papírových přijímacích testů k adaptivním zcela vyloučit, tím spíš, že v zahraničí adaptivní testování u přijímacího řízení využíváno je. Jedním z nejznámějších příkladů jsou americké přijímací zkoušky na magisterské obory, GRE, které byly do roku 2011 v zásadě plně adaptivním testem (Educational Testing Service, 2011). V roce 2011 pak došlo ke změně, a v současné době je jak tzv. Verbal Reasoning, tak tzv. Quantitative Reasoning, z nichž se GRE mimo jiné skládá, adaptivní po oddílech. To znamená, že v rámci každého ze dvou oddílů obsažených ve Verbal Reasoning a Quantitative Reasoning jsou úlohy pevně dané, a respondent se tedy může při řešení daného oddílu vracet k úlohám již dříve vyřešeným. Avšak platí, že druhý oddíl jako celek je respondentovi předložen adaptivně na základě jeho odpovědí v oddílu prvním. Podá-li tedy respondent například v prvním oddíle nadprůměrný výkon, je tomu přizpůsobena obtížnost druhého oddílu jako celku. Zůstává ovšem otázkou, jestli je možné obdobný postup aplikovat i v České republice, a tak nejspíš nezbývá než počkat, zda dlouholeté zahraniční zkušenosti společně s rostoucími možnostmi výpočetní techniky přispějí k rozšíření adaptivního testování i do těch oblastí, kde dosud příliš využíváno není.

Literatura

Centrum pro zjišťování výsledků vzdělávání (2015). *Agregovaná data za školy a jejich obory vzdělávání pro rok 2015* [internetový zdroj], staženo 23. 10. 2015 z. http://vysledky.cermat.cz

Eduational Testing Service (2011). *GRE: How the test is scored* [internetový zdroj], staženo 23. 10. 2015 z https://www.ets.org/gre/revised_general/scores/how/

Figueras, N., North, B., Takala, S., Verhelst, N., & Van Avermaet, P. (2005). Relating examinations to the Common European Framework: A manual. *Language Testing*, *22*(3), 261-279.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation*, *14*(8), 2.

van der Linden, W. J. (2008). Adaptive models of psychological testing. *Zeitschrift für Psychologie/Journal of Psychology, 216*(1), 1-2.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., & Mislevy, R. J. (2000). *Computerized adaptive testing: A primer*. Routledge.

Lenka Fiřtová, Martin Holík (2015): SCATE (Scio Computer Adaptive Test): The New Role of A Test

Computerized Adaptive Testing is a way to create customized tests and to develop new, interactive types of items. The results may then be used to create a tailor-made learning program. Scio started to develop adaptive tests in 2011, the adaptive test in English (SCATE) being one of the first. The aim was to create an item pool covering all the categories defined by CEFR. In addition, we also created items for complete beginners and labeled this category "A0". The items were divided into categories with respect to their difficulty and piloted using several hundreds of students whose level of English was known from prior testing (holders of internationally recognized certificates, such as FCE, CAE, TOEFL). When taking the test, respondents are presented with items which are appropriate for their ability level. The main idea is to present respondents with items which they don't find too easy but which they are still able to solve. First, respondents are presented with a set of randomly chosen items of varying difficulty. This results in a rough estimate of their ability level. The remaining items are then selected with respect to the respondents' previous answers. The algorithm is based on the Measurement Decision Theory. Adaptive testing ensures students are presented only with such items which they are likely to find interesting. The test then becomes not only an assessment tool, but also a tool facilitating the learning process.

Keywords: *Computerized Adaptive Testing, Common European Framework of Reference for Languages, Measurement Decision Theory*