# Určovanie hraničných skóre pre kriteriálne testy[1]

Andrej Mentel
Univerzita Komenského, Fakulta sociálnych a ekonomických vied

**Abstrakt:** Stať skúma proces určovania štandardov (hraničných skóre) pre kriteriálne pedagogické testy. Hlavným cieľom štúdie je poskytnúť ucelený rámec pre určovanie hraničných skóre založený hlavne na anglo-americkej výskumnej literatúre, ako aj na *Štandardoch pre pedagogické a psychologické testovanie* (AERA, APA, & NCME, 1999). Hlavný dôraz je venovaný procesu vytvárania opisov úrovní výkonu a metódam na určovanie hraničného skóre. Opisy úrovní výkonu sú ukázané na príkladoch zameraných na testovanie čítania s porozumením a sú založené hlavne na testovaní MCAS. Spomedzi metód určovania hraničného skóre, ktoré sú zamerané na testové položky, sú opísané Angoffova a Nedelského metóda a metóda záložiek. Metódy zamerané na osobu žiaka sú zastúpené prístupom založeným na kontrastných skupinách. Reprezentantom metód využívajúcich prvky oboch prístupov je klasifikácia na základe measurement decision theory. V závere sú krátko diskutované otázky validity týchto metód.

**Kľúčové slová:** kriteriálne testy, metódy určovania štandardov (hraničných skóre), čítanie s porozumením, kultúrny kapitál

## The Standard (Cut Score) Setting for Criterion-referenced Educational Tests

**Abstract:** The study investigates the process of standard (cut score) setting for criterion-referenced educational tests. The main goal of this study is to provide the comprehensive framework on standard setting based mainly on Anglo-American research literature as well as on the *Standards for educational and psychological testing* (AERA, APA, & NCME, 1999). Main emphasis is dedicated to the process of creating description of performance levels and to the methods of cut score setting. Description of performance levels is shown on some examples concerning the reading comprehension and is based on MCAS testing. Among the test-item centered methods, Angoff, Nedelsky and bookmarking methods are desribed. The person-centered methods are represented by contrasting group approach. As a representant of methods combining both approaches, the measurement decision theory classification is described. The validity issues of these methods are briefly discussed.

**Keywords:** criterion-referenced tests, methods of standard (cut score) setting, reading comprehension, cultural capital

**DOI:** 10.14712/23363177.2015.76

V procese pedagogického hodnotenia stojíme veľmi často pred rozhodnutiami týkajúcimi sa študijného úspechu žiaka, splnenia podmienok absolvovania predmetu,

[1] Príspevok vznikol vďaka grantu VEGA 1/0445/14 Kultúrny kapitál a školská úspešnosť.

prijatia na ďalší stupeň vzdelávania, zaradenia žiaka do kurzu podľa stupňa kompetencie apod. Tieto situácie majú napriek mnohým rozdielom jeden aspekt spoločný: určité dopredu definované hraničné skóre, ktoré keď žiak dosiahne alebo prekoná, tak „uspel“, v opačnom prípade „neuspel“. V mnohých prípadoch sa táto hranica určuje arbitrárne; ak by sme sa chceli dopátrať k zdôvodneniam, pravdepodobne by sme neuspeli.

Napríklad podľa slovenského školského zákona[2] tí žiaci, ktorí v celoslovenskom testovaní žiakov deviateho ročníka dosiahli v každom testovanom predmete úspešnosť aspoň 90 %, majú byť prijatí na väčšinu stredných škôl bez prijímacej skúšky (vyňaté sú z toho napr. stredné školy umeleckého zamerania). Prečo je táto hranica práve 90 %, prečo sa pri tom nezohľadňuje obťažnosť testu, faktické rozdelenie jeho skóre – toto sú otázky, ktoré si tvorcovia zákona pravdepodobne vôbec nepoložili. Prinajmenšom táto hranica nemá žiadne obsahové zdôvodnenie; je čisto arbitrárna. Nesmieme sa nechať pomýliť – pri tejto hranici v žiadnom prípade nejde o 90 % učiva, ktoré žiaci zvládli. Je to jednoducho hranica, ktorá hovorí, že v teste, ktorý mal napríklad 30 úloh bodovaných dichotomicky (jeden bod za správnu, žiaden za nesprávnu odpoveď), žiak nedostal bod v maximálne troch úlohách. Či to boli úlohy týkajúce sa podstatných častí učiva alebo naopak okrajových detailov, tento výsledok neberie do úvahy.

Podobná situácia nastáva v hodnotení na vysokých školách, ktoré prešli na medzinárodný kreditový systém (ECTS). Ak v rámci tohto systému žiak dosiahne v celkovom hodnotení úspešnosť menej než 60 %, neuspel; v opačnom prípade dostane známku (na Slovensku na stupnici A–E) na základe podobne definovaných hraničných skóre. Opäť, neexistuje nič, čo by vysvetľovalo, prečo práve 60 % je tou hranicou (a nie napríklad 55 alebo 68 %). Mohli by sme síce povedať, že v dlhodobom meradle to približne zodpovedá akémusi implicitne akceptovanému „primeranému“ podielu neúspešných študentov, ale z obsahovej stránky sa z tejto hranice nedozvieme nič.

Aby sme mohli interpretovať výsledok založený na percentuálnej úspešnosti, musíme ho porovnať s nejakou referenčnou hodnotou. Ak táto referenčná hodnota pochádza z rozdelenia úspešností ostatných žiakov, hovoríme o tzv. štatisticko-normatívnom testovaní (resp. o testovaní vzťahujúcom sa na normu, ang. *norm-referenced testing*). Vzhľadom na svoju povahu sa tento prístup nazýva aj „testovanie relatívneho výkonu“. Mnohé štandardizované didaktické testy používané na Slovensku majú túto povahu, napr. Testovanie 9 alebo Externá časť maturitnej skúšky (obe pripravované a administrované Národným ústavom certifikovaných meraní vzdelávania, NÚCEM).

Ak nás však príliš nezaujíma, ako dopadol určitý študent v porovnaní s ostatnými, ale chceme vedieť, nakoľko zvládol presne špecifikovanú oblasť učiva, potrebujeme testy kriteriálne. Pre tento typ testov sa používajú aj iné názvy, napr. overovacie testy, testy absolútneho výkonu, resp. vzťahujúce sa na kritérium – z ang. *criterion-referenced*, skr. CR (Chráska, 2009, s. 596). Využitie takýchto testov sa len

[2] § 65 ods. 4 zákona č. 245/2008 Z. z.

zriedka zaobíde bez definovania štandardov hodnotenia v príslušnej oblasti učiva. Len málokedy totiž dokážeme testovanú oblasť učiva (doménu) popísať dostatočne jednoznačne a zároveň všeobecne na to, aby sa priamočiarym deduktívnym postupom dalo vytvoriť dostatočné množstvo vzájomne zastupiteľných úloh.

Cieľom tohto článku je sumarizovať poznatky o praktických postupoch určovania štandardov (hraničných skóre) pre kriteriálne testovanie. Keďže sa tejto téme slovenská a česká pedagogika venovala dosiaľ len pomerne málo, nie je zatiaľ možné ponúknuť výsledky empirických výskumov tvorby a použitia kriteriálnych testov a v ich rámci definovaných štandardov výkonu. Tam, kde sa odvolávam na empirické dôkazy, teda spravidla využívam anglosaskú literatúru.

Z nedostatku domácich empirických výskumov vyplýva aj čiastočne preskriptívny charakter tejto state. Píše sa v nej skôr o tom, ako sa štandardy pre kriteriálne testy majú nastavovať, a len menej o tom, ako sa naozaj nastavujú. Navyše, pri súčasnom stave výskumu na Slovensku alebo v Českej republike nie je možné odpovedať na žiadnu z veľmi zaujímavých otázok týkajúcich sa sociálnych účinkov testovania s vysokým vplyvom. Táto stránka praxe pedagogického testovania (a zvlášť uplatňovania definovaných kritérií na zaraďovanie žiakov do úrovní výkonu) je však v súčasnej sociológii vzdelávania veľmi intenzívne a často veľmi kriticky diskutovaná (napr. Allensworth, 2005; Amrein & Berliner, 2002; Jacob, 2005; Au, 2008 atď.), a to čiastočne už aj na Slovensku, hoci tiež len v podobe prehľadu (Minarechová, 2012).

## 1 Terminologické poznámky

V nasledujúcom texte budem opakovane používať viaceré výrazy, ktoré sa pokúsim objasniť na tomto mieste. V prvej časti tohto článku sa v zmysle minimálneho akceptovateľného skóre v teste, ktoré ešte stačí na dosiahnutie určitej úrovne, vyskytuje výraz „štandard“. To je síce v súlade s anglosaskou literatúrou (napr. Hambleton, 1980; Hambleton & Pitoniak, 2006), ktorí v tejto súvislosti hovoria o tzv. štandardoch výkonu (*performance standards*), ale v našich (českých a slovenských) podmienkach by to mohlo miasť. Tento výraz by sa totiž ľahko mohol mýliť s výkonovými štandardmi definovanými štátnym kurikulom, ktoré slúžia ako kritérium splnenia pedagogických cieľov v príslušnej oblasti vzdelávania (porov. Slavík, 1999; Trna, 1996). V slovenčine je preto vhodnejším výrazom „hraničné skóre“, ktoré budem používať aj ďalej v tomto článku. V angličtine sa používa ako *cut score*, resp. *cut off score*. Ak sa budem odvolávať na dokument *Štandardy pre pedagogické a psychologické testovanie* (AERA, APA, & NCME, 1999), tak budem jeho názov okrem prvého použitia skracovať na *Štandardy* a zvýrazňovať kurzívou.

V súvislosti s určovaním hraničného skóre sa hovorí o tzv. „minimálne kompetentnom kandidátovi“ (ang. *minimally competent candidate*). Ide v zásade o konceptuálny model – predstavu, akými minimálnymi vedomosťami, schopnosťami a zručnosťami musí disponovať testovaný, aby dosiahol určitú úroveň na dolnej hranici (napr. úroveň „uspel“, alebo pri podrobnejšom delení ľubovoľnú z definovaných úrovní).

Synonymicky s týmto výrazom budem používať výraz hraničný kandidát alebo hraničný žiak.

Proces určovania štandardov pre testy si vyžaduje skupinovú prácu odborníkov, ktorí sa v tejto súvislosti v anglosaskej literatúre nazývajú *panelists* alebo *judges* (Hambleton & Pitoniak, 2006, s. 451). V tomto texte ich spravidla nazývam panelisti alebo tam, kde nemôže dôjsť k nejasnosti, jednoducho „odborníci", pretože nejde o posudzovateľov alebo recenzentov. Ich úloha je veľmi špecifická. Ako prvé musia vypracovať konkrétne opisy výkonu žiakov na príslušných úrovniach, pričom tieto opisy sa na jednej strane napájajú na kurikulum, na druhej strane sa prispôsobujú obsahu konkrétneho testu. Druhá časť ich práce spočíva v aplikácii niektorej z metód na určenie odborne podložených hraničných skóre.

Testy, pre ktoré je potrebné vypracovávať kritériá úspešnosti tak starostlivo a rigorózne, ako je to popísané v tomto článku, sú spravidla určené na použitie v celom štáte alebo v podobnom širokom rozsahu. Zvyčajne pri tom ide o „testy s vysokými stávkami" (high stakes tests). Tento výraz v ďalšom texte prekladám ako testy s vysokým vplyvom. Ide o také testy, ktorých výsledok významne ovplyvňuje študijnú alebo profesionálnu budúcnosť žiaka (napr. vďaka získaniu či nezískaniu certifikátu, diplomu, oprávnenia vykonávať určitú činnosť apod.). Takéto testy vytvára a administruje väčšia organizácia, ktorú v ďalšom texte nazývam testovacia agentúra (*testing agency*). V našich podmienkach ide predovšetkým o štátne organizácie, napríklad NÚCEM, ale v zahraničí je známych viacero renomovaných súkromných testovacích agentúr (napr. americká Educational Testing Service alebo holandská CITO).

Nechcel by som sa tu púšťať do diskusie, pre koho vlastne majú *high stakes* testy najvyšší vplyv; je zrejmé, že na jednej strane pôsobia ako selekčný mechanizmus, ktorý sa podieľa na reprodukcii kultúrneho kapitálu (Bourdieu & Passeron, 1990), no na druhej strane zasahujú zároveň samotnú školu a učiteľov. Testovaním konkrétneho (a nie iného) výberu z učiva sa nepriamo ovplyvňuje vyučovanie v školách podľa toho, čo testovacia agentúra považuje za prioritu vo vzdelávaní. V odbornej literatúre (u nás najmä Kaščák & Pupala, 2012; prehľad pozri v Rosenkvist, 2010, s. 26) je často opakovane popísaný proces „vyučovania pre testovanie", ktorý môže výrazne obmedziť vzdelávacie obsahy a stáva sa tak prostriedkom nepriamej sociálnej kontroly. Navyše, ak by škola bola na základe niektorých neoliberálnych koncepcií odmeňovaná či sankcionovaná na základe podielu žiakov, ktorí uspeli v testoch, predstavovalo by to výrazný vplyv aj na školu ako celok, nielen na jednotlivých žiakov. Tieto témy by si zasluhovali rozpracovanie v samostatnej stati; tu ich zmieňujem len preto, aby som upozornil na možné problémy a na potrebu veľmi zodpovedného prístupu pri určovaní hraničných skóre.

## 2 Zásady určovania hraničných skóre pre testy

Ako bolo uvedené vyššie, ak má test skutočne slúžiť ako nástroj kriteriálneho hodnotenia, je preň zvyčajne potrebné určiť hodnoty hraničného skóre pre toľ-

ko úrovní, koľko potrebujeme u žiakov rozlíšiť. Ak chceme obmedziť arbitrárnosť rozhodnutia, musíme proces určovania hraničných skóre, ako aj výber odborníkov (panelistov), ktorí ich budú určovať, podriadiť istým zásadám. Vo všeobecnosti tieto zásady popisujú *Štandardy pre pedagogické a psychologické testovanie* (Štandardy – AERA, APA, & NCME, 1999). Určovania hraničných skóre sa týka šesť článkov (štandardy 1.7; 2.14; 4.19; 4.20; 4.21 a 14.17). Hambleton a Pitoniaková (2006) podávajú stručné zhrnutie týchto článkov: Štandard 1.7 vyžaduje detailný popis kvalifikácie a skúseností panelistov, ich špecifickej prípravy na prácu s konkrétnym testom a aj toho, ako postupovali pri práci. Štandard 2.14 je technickej (štatistickej) povahy a žiada, aby testovacia agentúra zverejnila podmienenú štandardnú chybu merania pre všetky úrovne hraničných skóre a podľa možnosti aj pre susedné hodnoty skóre (toto sa týka konzistentnosti rozhodnutia). Štandardy 4.19 až 4.20 sa týkajú samotného procesu: 4.19 vyžaduje logické zdôvodnenie a dokumentáciu postupov určovania hraničných skóre. Štandard 4.20 si žiada uvedenie dôkazov o externej validite kategorizácie vytvorenej pomocou hraničných skóre (t.j. má byť jasne zdokumentované, že napr. medzi žiakmi klasifikovanými stupňom „neuspel“ a tými, čo „uspeli“, je rozdiel prejavujúci sa aj v iných relevantných kritériách). Štandard 4.21 hovorí o tom, že ak sa hraničné skóre vymedzujúce úrovne výkonu určujú na základe priameho posudzovania výkonov žiakov pri danej úlohe, potom sa má proces rozhodovania usporiadať tak, aby panelisti mohli najlepším možným spôsobom využiť svoje vedomosti a skúsenosti. Tento štandard sa teda týka usporiadania procesu rozhodovania hlavne pri metódach založených na posudzovaní testov. Napokon štandard 14.17 hovorí o tom, že pri testoch určených ako skúšky na získanie diplomu či certifikátu sa majú hraničné skóre určovať s ohľadom na to, aké vedomosti a zručnosti sa budú vyžadovať v povolaní alebo v ďalšom vzdelávaní; tvorcovia sa majú vyhnúť prispôsobovaniu štandardov výkonu potrebám regulovať počty absolventov, ktorí uspeli v teste (AERA, APA, & NCME, 1999; Hambleton & Pitoniak, 2006, s. 434).

V mnohých obsahových doménach učiva nie je jednoduché zrozumiteľne popísať, čo znamená, že žiak ovláda učivo na príslušnom stupni. Pedagógom tu môžu pomôcť rôzne taxonómie vzdelávacích cieľov, ktoré tieto ciele vyjadrujú v konkrétnych termínoch opisujúcich žiakovo správanie. Na tomto princípe je založená napr. Bloomova taxonómia (Bloom et al., 1956), ale aj iné prístupy. Wiersma a Jurs (1990, s. 229) vo svojej učebnici pedagogického testovania uvádzajú, že mnohé školy si vytvorili vlastné kurikulá založené na tom, aby vyučovanie priamo zodpovedalo špecifickým cieľom (ang. *objectives*[3]). Aj testy, ktorými tieto školy posudzovali progres žiakov, boli prispôsobené týmto cieľom.

Mohlo by sa zdať, že akékoľvek testy, kde nejako definujeme hraničné skóre pre to, či žiak uspel alebo nie, by sme mohli považovať za kriteriálne. Wiersma a Jurs

[3] Typicky sa rozlišujú dva pojmy zodpovedajúce cieľom: Všeobecné ciele (*goals*) sa odlišujú od špecifických cieľov (*objectives*). Rozdiel medzi nimi je ten, že kým mojím všeobecným cieľom môže byť napríklad stať sa dobrým plavcom, pričom mi je takmer jedno, kedy v živote to dosiahnem a ako sa to bude merať, mojím špecifickým cieľom je do konca tohto mesiaca zaplávať 100 metrov kraul pod jednu minútu.

(1990, s. 229) však zdôrazňujú, že je to omyl. To, či je test naozaj kriteriálny, závisí od toho, ako adekvátne je definovaná oblasť žiakovho správania tvoriaceho kritérium. Ak kriteriálny test neplánujeme využiť na rozhodnutia o dosiahnutí či nedosiahnutí určitej úrovne vedomostí či zručností, potom v zásade nepotrebujeme určovať hraničné skóre. V takom prípade stačí dostatočne precízne definovať testovanú doménu, rozlíšiť v nej pojmovú štruktúru, odlíšiť centrálne poznatky od okrajových a vybrať dostatočne reprezentatívne úlohy. Naopak, aj pri normatívnom teste môžeme stanoviť hraničné skóre definované na základe relatívneho výkonu (napr. v podobe percentilov alebo násobkov smerodajnej odchýlky rozdelenia).

Vo väčšine prípadov sa však kriteriálne testy používajú práve na klasifikáciu žiakov do rôznych úrovní výkonu. V takom prípade je kľúčom nájsť vhodný opis minimálneho akceptovateľného výkonu na danej úrovni zvládnutia učiva. Tomu slúži veľké množstvo metód, ktoré môžeme pracovne rozdeliť na dve základné skupiny: metódy zamerané na žiakov a metódy zamerané na položky testu. Nejde o striktnú dichotómiu, lebo existujú aj zmiešané či kompromisné metódy. Ak však toto delenie berieme len ako orientačné, tak základný rozdiel medzi týmito dvomi skupinami metód spočíva v tom, čo je centrom analýzy. Pri metódach zameraných na položky panelisti odhadujú úspešnosť, s akou môže minimálne kompetentný žiak vyriešiť danú úlohu. Pri metódach zameraných na žiaka sa naopak panelisti rozhodujú na základe iných dát, ako rozdeliť pilotnú vzorku žiakov na skupinu tých, ktorí dosiahli želanú úroveň, a tých, ktorí ju nedosiahli. Hambleton a Pitoniaková (2006) opisujú vo svojom prehľadovom článku osemnásť základných metód (nepočítajúc ich varianty). Všetky postupy nastavovania hraničných skóre majú spoločné to, že sú založené na kvalifikovaných odhadoch skupiny odborníkov. To je však zároveň ich najslabším miestom, preto *Štandardy* (AERA, APA, & NCME, 1999), ale aj ostatná literatúra zdôrazňujú potrebu starostlivého výberu a prípravy expertov. Hambleton a Pitoniaková (2006) rozdeľujú proces nastavovania hraničného skóre do viacerých krokov.

### 2.1 Opis procesu nastavovania hraničných skóre

Prvým krokom je výber konkrétnej metódy, resp. metód určovania hraničného skóre. Tu je dôležité zvážiť druh úloh použitých v teste, s ktorým pracujeme, časovú a finančnú náročnosť jednotlivých metód, predchádzajúce skúsenosti testovacej agentúry a v neposlednom rade empirické dôkazy o validite jednotlivých metód pre danú oblasť testovania. Ideálne by azda bolo, keby sa pri každom určovaní hraničného skóre použilo viacero metód a porovnali by sa ich výsledky. Autori článku, na ktorý sa odvolávam, však realisticky priznávajú, že toto sa dá očakávať jedine v pilotných alebo výskumných štúdiách, keďže ide o časovo aj finančne veľmi náročnú úlohu (Hambleton & Pitoniak, 2006, s. 436).

Druhým krokom je výber skupiny panelistov. Tu treba pamätať na viacero vecí: Ich výber musí rešpektovať rôznorodosť cieľovej skupiny (ak napríklad pripravujeme štandardy pre maturitný test, musia byť v paneli zastúpení učitelia rôznych typov škôl, optimálne z rôznych regiónov a s rôznymi skúsenosťami týkajúcimi sa študijnej

a profesionálnej budúcnosti študentov). V paneli by mali byť zastúpení aj odborníci na kurikulum a zástupcovia ďalších zúčastnených strán. Dôležitým hľadiskom výberu panelistov je aj zvolená metóda určovania hraničného skóre. Typicky si metódy zamerané na položky vyžadujú isté skúsenosti s používaním psychometrických metód. Metódy založené na posudzovaní žiakov si vyžadujú oveľa menej skúseností s analýzou testov, ale naopak, potrebné je tu kvalitné zázemie v hodnotení študentov.

Tretím, z hľadiska úspechu procedúry určovania hraničných skóre zásadným krokom je tvorba deskriptorov pre jednotlivé úrovne. Deskriptory sú opisy znalostí, schopností a zručností žiakov dosahujúcich príslušnú úroveň. Keďže ide o veľmi dôležitú časť postupu, venujeme jej v samostatnej časti článku.

Štvrtým krokom je tréning panelistov v tom, ako sa používa zvolená metóda. Je potrebné, aby im testovacia agentúra zrozumiteľne vysvetlila proces a ciele určovania hraničných skóre, oboznámila ich s testom, ktorý sa bude posudzovať, a jeho kľúčom (resp. s kritériami skórovania odpovedí) a aby sa panelisti oboznámili s tým, čo sa od nich očakáva – ako čo robiť a ako vyplniť posudzovací formulár. Potom sa panelisti musia dôkladne oboznámiť s deskriptormi jednotlivých úrovní a s tým, ako ich prepojiť s konkrétnymi úlohami, ktoré sa môžu objaviť v testoch. Napokon by si mali procedúru vyskúšať na podobnom teste, ako je ten, s ktorým budú pracovať. Testovacia agentúra by mala mať k dispozícii štatistické analýzy cvičného testu, aby bolo možné poskytnúť panelistom spätnú väzbu. Tieto poznámky sa pochopiteľne týkajú hlavne metód založených na posudzovaní položiek testu, ale prenos na metódy zamerané na žiaka je zrejmý.

Piaty krok spočíva v samotnom posudzovaní (či už úloh alebo žiakov). Konkrétny postup závisí od zvolenej metódy. Prakticky sa zvyčajne odhady panelistov zapisujú do pripraveného formulára a potom sa centrálne spracujú. Po prvom kole, keď sa zozbierajú odhady od všetkých panelistov, má prebehnúť diskusia. Testovacia agentúra by mala mať k dispozícii výsledky štatistických analýz založených na pilotných testovaniach, ktoré môže použiť ako spätnú väzbu pre panelistov. V diskusii by sa mali vyjasniť prípadné nesprávne pochopenia alebo odlišné interpretácie (či už cieľov, deskriptorov alebo znenia úloh). Môže nasledovať druhé a po ňom prípadne aj tretie kolo; panelisti tu môžu, ak chcú, zmeniť názor.

Šiesty krok predstavuje výpočet finálnych hraničných skóre založených na poslednom kole posudzovania. Ako výsledná hodnota sa používa buď priemer alebo medián posudkov jednotlivých panelistov. Hambleton a Pitoniaková (2006, s. 439) tu uvádzajú, že hoci medián je vhodnejší v prípade menšej skupiny posudzovateľov alebo v prípade asymetrického rozdelenia hodnotení, priemer umožňuje odhad štandardnej chyby priemeru, čo je užitočné pre posúdenie stability hodnotení a čo medián neumožňuje. Moderné robustné štatistické metódy odhadu však dokážu kompenzovať aj tento problém, pokiaľ použijeme niektorú z robustných mier polohy rozdelenia (Wilcox, 2012).

Finálne kroky zahŕňajú evalváciu panelistov, sumarizáciu celého procesu a prípravu záverečnej správy, z ktorej je jasné, čo, ako a prečo sa robilo (Hambleton & Pitoniak, 2006, s. 439).

Takto opísaný postup sa môže zdať príliš zložitý a formálny. Rigoróznosť jednotlivých krokov súvisí s tým, že v prostredí USA, kde tento proces predpisujú *Štandardy* (AERA, APA, & NCME, 1999), by mohlo byť arbitrárne a nepodložené nastavenie hraničného skóre právne neobhájiteľné. Týka sa to predovšetkým testov s vysokým vplyvom na budúcnosť testovaného (*high stakes tests*), teda takých testov, od ktorých závisí napríklad prijatie na vysokú školu, vydanie certifikátu alebo diplomu.

V nasledujúcich častiach sa podrobnejšie vrátim k viacerým praktickým otázkam, ktoré môžu zaujímať tvorcov testov. Zmienim sa teda o výbere panelistov, tvorbe deskriptorov a o najčastejšie používaných metódach určovania hraničných skóre. V nasledujúcej diskusii priblížim empirické poznatky z výskumov, ktoré sa vzťahujú na viaceré obmedzenia týchto metód. Výber metód určovania hraničných skóre v tomto článku je nevyhnutne subjektívny, ale je aspoň čiastočne odôvodniteľný tým, že sa im venovala aj pomerne veľká pozornosť empirického výskumu.

### 2.2 Koho vyberať ako panelistov a koľko ich má byť

Určovanie hraničného skóre pre testy je proces, ktorého úspech je závislý na úsudku odborníkov, preto je dôležité, aby ich výber aj počet bol verejne obhájiteľný. Pri výbere odborníkov sa treba rozhodovať na základe dvoch kritérií:

- Kvalifikácia odborníkov: Je potrebné, aby väčšina panelu bola tvorená skúsenými učiteľmi vyučujúcimi daný predmet. Je pritom dobré, keď sa orientujú v problematike didaktických testov, a to aspoň na úrovni poučeného klienta (rozumejú správam zo psychometrických analýz testov). Okrem učiteľov by mali byť v paneli zahrnutí aj odborníci na kurikulum a zástupcovia ďalších zúčastnených strán (napr. riaditelia škôl, členovia školských rád apod.).
- Panel by mal dobre reprezentovať obyvateľov krajiny, čo sa týka rodu, etnicity, geografie, veku a profesionálnej skúsenosti. Táto požiadavka smeruje k tomu, aby boli nastavené štandardy čo najmenej napadnuteľné politicky (napr. z dôvodov diskriminácie).

Vyhovieť obom kritériám naraz je veľmi ťažké. Hambleton a Pitoniaková (2006) tu odporúčajú uprednostniť odbornú stránku; ako príklad uvádzajú panel, ktorý používa americké Národné centrum pedagogickej štatistiky (NCES) pri nastavovaní hraničných skóre pre testy NAEP. Asi 70 % ich panelistov tvoria učitelia, kým zvyšok pozostáva z ostatných, napr. zástupcov rodičov a významných zamestnávateľov.

Počet panelistov je takisto otázka, kde sa musíme rozhodovať na základe kompromisu medzi dvomi kritériami. Na jednej strane potrebujeme z dôvodu zachovania stability rozhodnutí čo najväčšiu skupinu, na druhej strane musíme počítať s dvomi obmedzeniami: Prvým sú náklady a druhým je čas, ktorý budú odborníci môcť venovať práci v paneli. Ak máme angažovať kvalitných odborníkov, musíme brať do úvahy, že ide spravidla o ľudí, ktorí majú veľa práce. Aby bola práca v paneli kvalitná a nie len formálna, stojíme pred obmedzením tohto druhu – že jednoducho nebudú mať čas.

Odhady založené na štatistickej teórii zovšeobecniteľnosti (Raymond & Reid, 2001) ukazujú, že pri pedagogických testovaniach je potrebných 15 až 30 panelistov, aby sa zabezpečili obe stránky výberu odborníkov: ich dostatočne pestré zastúpenie na jednej a stabilita výsledkov na druhej strane. Ak by boli odborníci do panelu vyberaní tak, že by skupina bola homogénna, na zabezpečenie stability by stačil aj polovičný počet (10–15 členov), ale v prípade, že panel obsahuje zástupcov rôznych profesií a záujmových skupín, je na zabezpečenie prijateľných štandardných chýb priemerov potrebná oveľa väčšia skupina.

### 2.3 Tvorba deskriptorov pre kategórie výkonu

Kľúčovou otázkou, na základe ktorej budú panelisti rozhodovať o nastavení vhodných hraničných skóre, je, čo konkrétne musí vedieť urobiť žiak na to, aby sme ho mohli zaradiť do príslušnej úrovne výkonu. Ide o proces, v ktorom treba prepojiť všeobecné požiadavky kurikula (ako sú formulované vo výkonovom štandarde vzdelávacích programov) s konkrétnym obsahom testu. Mills a Jaeger (1998) tento proces odporúčajú robiť v niekoľkých krokoch. Panel odborníkov, ktorí budú neskôr stanovovať pre daný test hraničné skóre, sa musia najskôr oboznámiť s cieľom svojej práce a s tým, na čo konkrétne sa využije (napr. aký je účel testovania, pre ktoré idú vyvinúť štandardy?). Potom sa musia dôkladne oboznámiť so špecifikáciou testu. Obsah testu, t.j. aké oblasti má pokrývať a do akej hĺbky, musia byť z tejto špecifikácie jasné nielen kvôli výberu úloh tvorcami, ale aj pre vypracovanie adekvátnych deskriptorov. Odporúča sa, aby si panelisti sami vyskúšali vyriešiť test, s ktorým idú pracovať. Majú si pri tom všímať spôsob skórovania úloh, teda to, za čo sa pridávajú body.

V ďalšom kroku je potrebné, aby sa pokúsili prepojiť jednotlivé položky testu s obsahovou špecifikáciou. Nakoľko úlohy pokrývajú vytýčené oblasti testovania? Pokrývajú skôr jadro, či skôr okrajové oblasti? Dôležité je, aby o týchto aspektoch panelisti diskutovali, takže im na to treba nechať dosť času.

Až po tom, ako sa panelisti dôkladne oboznámia s testom, ich testovacia agentúra oboznámi s požiadavkami na počet a charakteristiky kategórií výkonu. Len málokedy (napr. pri testoch v autoškole) sa určujú len dve kategórie (uspel vs. neuspel). Napríklad vzdelávacie štandardy zo slovenského jazyka a literatúry pre ISCED 2 (ŠPÚ, 2009) určujú tri úrovne: nedostatočnú, minimálnu („dostatočný“) a maximálnu („výborný“). Napríklad v časti *próza* sa pri téme *literárna postava* za kritérium na minimálnu úroveň výkonu požaduje, aby žiak vedel určiť hlavné a vedľajšie postavy. Kritériá na maximálnu úroveň výkonu sú nasledujúce (ŠPÚ, 2009, s. 114):

- vie vysvetliť literárnu postavu ako fikciu autora, nositeľa deja;
- vie určiť hlavné a vedľajšie postavy;
- vie výstižne charakterizovať postavy.

Systém MCAS (Massachusetts Comprehensive Assessment System) rozlišuje štyri úrovne: neuspel, potrebuje sa zlepšiť, zdatný a pokročilý (Hambleton & Pitoniak, 2006, s. 453). Úrovne sa opisujú veľmi všeobecne (MCAS 2013a):

- Pokročilí: Žiaci na tejto úrovni prejavujú obsiahle a hlboké pochopenie veľmi náročného učiva a prichádzajú s premyslenými riešeniami zložitých problémov.
- Zdatní: Žiaci na tejto úrovni prejavujú solídne porozumenie náročnému učivu a riešia široké spektrum problémov.
- Potrebujú sa zlepšiť: Žiaci na tejto úrovni prejavujú čiastkové porozumenie učivu a riešia jednoduché problémy.
- Neuspeli: Žiaci na tejto úrovni prejavujú minimálne porozumenie učivu a nedokážu riešiť ani jednoduché problémy.

V tejto podobe sa pochopiteľne tieto opisy len ťažko dajú využiť na tvorbu konkrétnych deskriptorov. Preto sa napríklad sú pre každú oblasť vypracované podrobnejšie opisy. Ako príklad uvedieme opis úrovne *zdatní* z časti *porozumenie* v oblasti *jazykové zručnosti v angličtine* (MCAS 2013b):

- chápe mnoho konkrétnych myšlienok a väčšinu abstraktných alebo implikovaných myšlienok v textoch vhodných pre daný ročník;
- prepája myšlienky v texte a dokáže to podporiť argumentom.

Na pokročilej úrovni potom ide o hlboké pochopenie konkrétnych aj abstraktných myšlienok, dokáže prepájať aj zložité idey a podáva dobre premyslené a podložené argumenty. Na základnej úrovni („potrebuje sa zlepšiť“) sa naopak vyžaduje len prepojenie niektorých myšlienok (argumentácia sa nespomína) a aj pochopenie abstraktných ideí sa obmedzuje na „niektoré“.

Takto definované úrovne sú stále dostatočne všeobecné na to, aby sa dali prispôsobiť na mnohé typy textov aj úloh. Úlohou panelu odborníkov je teda vytvoriť opisy výkonu, ktoré budú zodpovedať týmto všeobecným požiadavkám, ale budú naplnené obsahom testu.

V ďalšom kroku je dobré (hlavne pri testoch obsahujúcich otvorené úlohy, ale nielen) oboznámiť panelistov s príkladmi vyplnených testov žiakov rôznych úrovní výkonu. Účelom nie je skĺznuť k povrchnému klasifikovaniu, ale skôr prediskutovať typické chyby, ktoré žiaci robia. Nápomocné sú tu výsledky štatistickej analýzy (napr. analýzy distraktorov pri úlohách zameraných na výber odpovede).

Napokon sa panelisti dostanú k samotnému písaniu deskriptorov pre jednotlivé úrovne a oblasti testovania. Ak napríklad test z matematiky obsahuje tri oblasti (napr. geometriu, algebru a spracovanie údajov), každá má byť opísaná samostatne. Je dobré, aby pracovali najskôr individuálne alebo v menších skupinkách a potom svoje názory prediskutovali v celej skupine. Odporúča sa (Mills & Jaeger, 1998), aby pri vypracúvaní opisov jednotlivých úrovní začali na strednej úrovni výkonu a potom ju rozvinuli smerom k pokročilému stupňu a zúžili smerom k základnému. Tento postup je užitočný na to, aby sa dodržala primeraná postupnosť nárokov. Zvlášť veľkú pozornosť je potrebné venovať opisu výkonu minimálne kompetentného kandidáta.

Celý proces je veľmi náročný na čas a pracovné nasadenie, preto je efektívne niektoré fázy presunúť na individuálnu prácu. Až kvalitne pripravené deskriptory úrovní výkonu však umožňujú panelistom lepšiu orientáciu v úlohách, schopnostiach žiakov a v tom, ako nastaviť hraničné skóre.

Ako sme spomenuli vyššie, existuje niekoľko desiatok rôznych metód určovania hraničných skóre. Mnohé z nich sa zameriavajú na analýzu testových položiek, kým iné (ktorých je však menej) sa orientujú na rozbor výkonov žiakov. Ako príklad metód zameraných na testové položky si spomenieme dva varianty Angoffovej metódy (Angoff, 1971), Nedelského metódu (Nedelsky, 1954) a metódu záložiek (Mitzel, Lewis, Patz, & Green, 2001). Ako príklad metódy zameranej na žiaka si popíšeme metódu kontrastných skupín (Livingston & Zieky, 1982, 1989). Spomedzi zmiešaných metód stručne predstavíme klasifikáciu založenú na *measurement decision theory* (Rudner, 2009).

Pravdepodobne najznámejšie, no nie nevyhnutne najlepšie metódy určovania hraničného skóre navrhol Angoff (1971). Prvá, všeobecne rozšírenejšia metóda spočíva v tom, že panelisti pre každú položku odhadujú, aký podiel (resp. percento) minimálne schopných uchádzačov danej úrovne pravdepodobne vyrieši posudzovanú položku správne. Odhady panelistov sa zapisujú do formulára; navrhované hraničné skóre od jedného panelistu je dané súčtom jeho odhadov. Predpokladajme, že máme celkovo 4 položky, pričom za každú by žiak mohol dostať max. jeden bod. Ak by jeden posudzovateľ pre všetky položky odhadoval, že by ich vyriešilo 30 % spomedzi minimálne kompetentných uchádzačov, tak ním navrhované hraničné skóre by bolo $4 \times 0{,}3 = 1{,}2$ bodu. Pochopiteľne, tieto odhady nemusia byť pre každú úlohu rovnaké. Výsledné hraničné skóre sa určuje ako priemer návrhov jednotlivých panelistov. Kvôli problematickým vlastnostiam priemeru vo vzorkách vyberaných z iného než normálneho rozdelenia by na určenie skutočnej centrálnej hodnoty bolo vhodnejšie používať niektorú robustnejšiu metódu, napr. medián alebo orezaný (trimmed), príp. winsorizovaný priemer (Wilcox, 2012).

Táto verzia Angoffovej metódy, hoci je všeobecne oveľa známejšia, je v jeho vlastnej práci (Angoff, 1971) uvedená len v poznámke pod čiarou. Metóda, ktorú navrhoval ako základnú, by sa dalo označiť za dichotomizovanú: Panel expertov neodhaduje percentuálne úspešnosti minimálne schopných kandidátov, ale len to, či danú položku vyriešia, alebo nie.

S odhadmi úspešnosti čiastočne súvisí Nedelského metóda určovania hraničného skóre, ktorá však berie do úvahy taktiku riešenia. Nedelsky (1954) ju vyvinul len pre položky s výberom odpovede. Panelisti sa snažia odhadnúť, či minimálne schopný kandidát dokáže niektoré možnosti vylúčiť ako príliš nepravdepodobné. Počítal totiž, že títo žiaci typicky postupujú vylučovacou metódou – vylúčia najmenej vhodné odpovede a medzi zvyšnými sa rozhodnú náhodne (tipovaním). Každej úlohe sa potom pripíše skóre, ktoré vznikne ako pravdepodobnosť uhádnutia správnej odpovede po vylúčení niektorých distraktorov. Ak má napríklad položka 4 možnosti odpovede a minimálne schopný (hraničný) žiak z nich vie dve vylúčiť, potom ostávajú len dve zvyšné možnosti. Skóre, ktoré sa započítava za túto položku, je teda rovné 0,5.

Metóda záložiek (*bookmark method*) predpokladá, že predtým, ako sa stretne panel posudzovateľov, vypracuje testovacia agentúra štatistické analýzy položiek

testu (typicky na základe pilotného testovania). Každý z posudzovateľov dostane testové úlohy zoradené podľa vzrastajúcej obťažnosti. Typicky sa používa odhad obťažnosti na základe teórie odpovede na položku (pozri napr. Urbánek, Denglerová, & Širůček, 2011). Posudzovatelia majú umiestniť „záložku" pred tú úlohu, o ktorej si myslia, že ju už nevyrieši vopred dohodnutý podiel hraničných žiakov. Teoreticky by malo byť jedno, na akom percente úspešnosti sa panelisti dohodnú (v rámci teórie odpovede na položku sa tieto hodnoty navzájom dajú prepočítať). Výskumná literatúra (prehľad pozri v Hambleton & Pitoniak, 2006, s. 443–444) prináša pomerne rozporné výsledky týkajúce sa toho, aký podiel hraničných žiakov treba použiť, aby boli odhady najkonzistentnejšie. Citovaní autori sa prikláňajú k hodnote 67 % (resp. 2/3), ktorá je podľa nich pre panelistov psychologicky najpriateľnejšia (ibid., s. 444). Štúdia PISA používa ako referenčný podiel 62 % žiakov danej charakteristiky (porov. OECD, 2009).

Metóda záložiek má napriek nevýhode vyplývajúcej z vysokej kognitívnej náročnosti pre panelistov (odhad pravdepodobnosti riešenia; podrobnejšie pozri diskusiu v tomto článku) veľmi výraznú výhodu, ktorá ju môže predurčovať pre použitie v plošných testoch vysokého vplyvu, kde by bolo potrebné, aby sa hraničné skóre počas rokov nemenilo. Táto výhoda vyplýva z toho, že pri záložkovej metóde určujeme obťažnosť položiek na objektívnej intervalovej škále. Vyžaduje si však dostatočne bohatú banku úloh, ktoré majú známe psychometrické charakteristiky a ktorú testovacia agentúra priebežne dopĺňa úlohami, ktoré predtým okalibruje v pilotnom testovaní.

Napokon si priblížime metódu založenú na kontrastných skupinách (Livingston & Zieky, 1982, 1989). V tejto metóde panelisti neposudzujú úlohy testu, pre ktorý sa má nájsť hraničné skóre, ale žiakov. Na základe deskriptorov testovanej oblasti rozdelia vzorku žiakov na dve skupiny, a to na skupinu spĺňajúcu minimálne požiadavky na začlenenie do príslušnej úrovne a na tú, ktorá tieto požiadavky nespĺňa. Dôležité tu je, aby boli tieto dve skupiny rozlíšené nezávislými metódami na zabezpečenie súbežnej validity. Potom sa všetkým žiakom zadá test. Hraničné skóre sa potom dá určiť viacerými spôsobmi. Najjednoduchší spôsob je založený na porovnaní histogramov: Tam, kde sa pretínajú rozdelenia, sa umiestni hraničné skóre. Presvedčivejší spôsob, hlavne v prípade, že sa rozdelenia výraznejšie prekrývajú, je však na základe logistickej regresie (Livingston & Zieky, 1989).

Podobne ako metóda kontrastných skupín pracuje aj klasifikácia na základe *measurement decision theory* (MDT; pozri Rudner, 2009). Táto metóda má výhodu v tom, že predpokladá len to, že položky posudzovaného testu sú navzájom štatisticky nezávislé (teda že správnosť odpovede v jednej z nich neovplyvňuje správnosť riešenia inej). Aby sme mohli túto metódu použiť na rozhodnutie, do ktorej úrovne patrí testovaný žiak, potrebujeme dopredu aspoň približne poznať podiel žiakov danej úrovne v celkovej populácii a percentuálne úspešnosti riešenia daných úloh pre jednotlivé úrovne. Rozhodnutie, do ktorej skupiny žiak patrí, urobíme pomocou Bayesovho vzorca, pomocou ktorého odhadneme, aká je pravdepodobnosť, že žiak patrí do danej skupiny, pokiaľ odpovedal určitým spôsobom. Presnosť (spoľahlivosť)

rozhodnutia sa určuje na základe pomeru vierohodností (likelihood ratio). Metóda založená na MDT však nepracuje s hraničným skóre v podobe „minimálneho počtu bodov“, ale so vzorom odpovedí (s tým, na ktoré položky žiak odpovie správne a na ktoré nie). To je na jednej strane veľmi výhodné z hľadiska samotného testovania (umožňuje to totiž klasifikáciu pomocou tzv. adaptívneho testovania, keď sa každému žiakovi vyberajú úlohy primerané jeho miere schopnosti), no na druhej strane to môže byť politicky ťažko obhájiteľné. Ako totiž ukázal Georg Rasch (1980), na to, aby bolo hrubé skóre (t.j. počet dosiahnutých bodov) jednoznačným ukazovateľom schopnosti, musí byť test jednorozmerný (t.j. úlohy musia byť indikátormi práve jednej latentnej vlastnosti), položky musia byť navzájom štatisticky nezávislé a musia mať aspoň približne rovnakú rozlišovaciu schopnosť. Oproti týmto pomerne striktným požiadavkám predpokladá MDT len štatistickú nezávislosť položiek. Túto metódu pri rozhodovaní o zaradení žiaka do príslušnej úrovne používa v Českej republike spoločnosť SCIO pri počítačovom adaptívnom testovaní angličtiny SCATE (SCIO, n.d.). Jej výhodou je, že kombinuje expertné poznanie so štatistickými postupmi takým spôsobom, ktorý nemusí nevyhnutne obsahovať odhady pravdepodobností (tie sa dajú získať aj na základe predbežného pilotného testovania).

## 3 Diskusia a záver

V tejto časti state zhrnieme a budeme krátko komentovať výsledky empirického výskumu použitia opísaných metód určovania hraničných skóre pre testy. Zameriame sa najskôr na metódy samotné a potom stručne prediskutujeme ich vzájomné vzťahy.

Oba opísané varianty Angoffovej (1971) metódy spočívajú v odhade pravdepodobnosti, čo je však úloha, ktorá je pre ľudí veľmi náročná. Je dobre známe, že ľudia pri odhadoch pravdepodobnosti majú sklon robiť systematické skreslenia (ľudia typicky pravdepodobnosti vyššie než 40 % podhodnocujú, t.j. tam, kde sa objektívne dá očakávať pravdepodobnosť napr. 60 %, odhadujú v priemere okolo 40 % apod.). Naopak, nižšie pravdepodobnosti typicky mierne nadhodnocujú (napr. namiesto 20 % odhadujú okolo 30 %). Shepardová (1995) aj iní autori, napr. Reckase a Bayová (1999), tento efekt konštatujú v aplikácii na odhady obťažnosti položiek: Panelisti systematicky nadhodnocujú výkon v ťažkých položkách (t.j. v takých, ktoré majú nízke pravdepodobnosti správneho riešenia), pričom zároveň podhodnocujú výkon v ľahkých úlohách. K zaujímavým výsledkom došli Impara a Plakeová (1998). Ich vzorku tvorilo 26 učiteľov, ktorí mali odhadovať výkony svojich žiakov v teste z prírodných vied. Mali však urobiť dva odhady: Jednak pre žiakov ako celok, jednak pre žiakov považovaných za „tesne dostatočných“ (prejavujúcich minimálnu ešte akceptovateľnú úroveň vedomostí). Ukázalo sa, že učitelia presnejšie odhadujú výkon žiakov ako celku (ale aj v tomto prípade nadhodnocovali výkony žiakov v priemere o 9 %); výkony minimálne kompetentných žiakov podhodnocovali oveľa výraznejšie, v priemere až o 16,7 % (Impara & Plake, 1998). Toto zistenie je zaujímavé preto, že ukazuje na presne opačný jav ako predchádzajúce citované

práce. Je preto otázkou, ktorá si bude vyžadovať ďalší výskum, od čoho závisia a nakoľko spoľahlivé sú v tejto metóde odhady úspešnosti pre žiakov rôznych úrovní výkonu.

Nedelského (1954) metóda na jednej strane obchádza problém s odhadovaním pravdepodobnosti riešenia, na druhej strane zas počíta len s jednou z možných taktík riešenia úlohy hraničnými žiakmi. Vôbec však nie je isté, aké taktiky riešenia testových úloh typicky používajú žiaci v našich podmienkach; autorovi tohto prehľadu nie sú známe žiadne výskumné štúdie ani diplomové práce na túto tému.

Metóda záložiek (Mitzel et al., 2001) je podobne ako Angoffova metóda založená na odhadoch pravdepodobnosti úspešného riešenia, takže napriek obrátenému formátu úlohy je možné, že sa pri nej budú prejavovať tie isté nedostatky. Obrátený formát znamená, že namiesto priameho odhadu, koľkí žiaci daných charakteristík úspešne vyriešia určitú úlohu, panelisti riešia, ktorú úlohu v rade postupne stále ťažších ešte vyriešia napríklad dve tretiny minimálne kompetentného žiaka. Výskumná literatúra sa zameriavala na overenie vplyvu tohto podielu úspešných hraničných žiakov na konzistentnosť odhadov. V rámci teórie odpovede na položku je to z matematického hľadiska jedno: Ak sa panelisti dohodnú, že budú brať do úvahy polovičnú úspešnosť, tak umiestnia záložky ďalej od začiatku (smerom k ťažším úlohám). Ak sa dohodnú na 67 %, alebo dokonca na 80 % žiakov, tak záložka bude bližšie k začiatku (medzi ľahšími úlohami). Teoreticky by sa mali tieto hodnoty medzi sebou dať ľahko prepočítať (ak vieme, akú obťažnosť majú mať hraničné úlohy, ktoré úspešne vyriešia 2/3 minimálne kompetentných žiakov, tak z toho ľahko odvodíme, akú majú mať tie, ktoré vyrieši len polovica takýchto žiakov). To je možné práve vďaka modelom používaných v teórii odpovede na položku (u nás viď Urbánek et al., 2011). Problém však je opäť v subjektívnych odhadoch pravdepodobnosti. Pri vyššom percente úspešných žiakov (napr. pri hodnote 67 %) posúvali záložku výrazne ďalej od začiatku, než by sa očakávalo podľa toho, kam umiestňovali záložku, keď podiel úspešných žiakov mal byť len 50 %. To znamená, že nadhodnocovali výkon týchto žiakov pri ťažších úlohách (Williams & Schulz, 2005); toto zistenie je konzistentné so zisteniami, ktoré sme uviedli v diskusii Angoffovej metódy.

Zaujímavou otázkou je, aká je stabilita hraničných skóre určených pre ten istý test pomocou rôznych metód. Obzvlášť zaujímavé je porovnanie niektorej metódy zameranej na položky testu (napr. Angoffova alebo metóda záložiek) s metódami založenými na hodnotení žiakov (napr. porovnanie kontrastných skupín). Tu sa však otvára široké pole pre výskumy. Výsledky doterajších výskumov poskytujú obraz pomerne vysokej zhody: V štúdii porovnávajúcej Angoffovu metódu so záložkami (Näsström & Nyström, 2008) vychádzajú obe metódy porovnateľné a poskytujúce pomerne spoľahlivé výsledky. Podobne vychádza aj porovnanie Angoffovej metódy s metódou záložiek vo dvoch kolách (Olsen & Smith, 2008), pričom tu sa ukázalo, že metóda záložiek dáva stabilnejšie výsledky, pokiaľ sa porovnávajú dve kolá odhadov. Aj Buckendahl et al. (2002) došli k podobným výsledkom: umiestnenie hraničného skóre medzi týmito metódami kolísalo v rozmedzí 1 %, čo je veľmi vysoká miera

zhody. Je však pravdepodobné, že táto vysoká zhoda bola spôsobená aj veľkosťou panelu (23 odborníkov) a predmetom – šlo o test z matematiky, kde je možné veľmi precízne sformulovať deskriptory dosiahnutia príslušnej úrovne.

K oveľa menej presvedčivým výsledkom prišli Çetin a Gelbal (2013). Ich panel bol zložený zo 17 odborníkov, ktorí analyzovali 60-položkový test jazyka a slovnej zásoby. Veľký rozptyl v odhadoch a relatívne nízka korelácia medzi odhadmi úspešnosti žiakov pri riešení položiek a ich skutočnými obťažnosťami mohli mať viacero príčin. Autori uvažujú o obmedzení vyplývajúcom z usporiadania testovania (resp. zo zloženia vzorky testovaných). Nediskutujú možnosť, že by mohlo ísť o efekt testovaného predmetu (oveľa menej precízne stanovené deskriptory pre test z jazyka). Táto problematika je však tiež otvorená ďalšiemu výskumu, takisto ako aj vzťahy medzi odhadmi založenými na testových položkách oproti odhadom založených na skupinách žiakov.

Ako bolo spomenuté vyššie, úsilie zabezpečiť čo najvyššiu mieru validity v pedagogických testoch nie je len čisto odborná záležitosť, obzvlášť ak ide o testy s vysokým vplyvom (*high stakes tests*). Na základe týchto výsledkov sa aspoň v princípe prijímajú rozhodnutia, ktoré sa podieľajú na generovaní a reprodukcii kultúrneho kapitálu v spoločnosti (Bourdieu, 1986). Samotný proces certifikácie (testu, rozhodnutia a v prípade pozitívneho výsledku získania certifikátu) predstavuje zo strany inštitúcie mocenský zásah, ktorý je vždy arbitrárny a vždy sa riadi záujmami tých aktérov v príslušnom sociálnom poli (napr. v tomto prípade v poli školského vzdelávania), ktorí disponujú kapitálom umožňujúcim im určovať pravidlá (Bourdieu & Passeron, 1990). Ak však jednou zo základných princípov, na základe ktorých má fungovať súčasné školstvo, je akontabilita (ktorej nástrojom majú byť práve testy, pozri West, 2010), potom je potrebné, aby ich tvorcovia vedeli podložiť, na akom konkrétnom poznaní sa zakladajú rozhodnutia o študijnom úspechu či neúspechu žiakov. Tomuto cieľu slúžia aj napriek ich problematickým vlastnostiam metódy opísané v tomto článku.

## Literatúra


AERA, APA, & NCME (1999). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association (AERA).

Allensworth, E. M. (2005). Dropout rates after high-stakes testing in elementary school: A study of the contradictory effects of Chicago's efforts to end social promotion. *Educational Evaluation and Policy Analysis, 27*(4), 341–364.

Amrein, A. L., & Berliner, D. C. (2002). *An analysis of some unintended and negative consequences of high-stakes testing*. Dostupné z http://greatlakescenter.org/docs/early_research/pdf/H-S%20Analysis%20final.pdf.

Angoff, W. H. (1971). Scales, norms and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., s. 508–597). Washington, DC: American Council on Education.

Au, W. W. (2008). Devising inequality: a Bernsteinian analysis of high-stakes testing and social reproduction in education. *British Journal of Sociology of Education, 29*(6), 639–661.

Bloom, B. S. et al. (1956). *Taxonomy of educational objectives: Handbook 1, the cognitive domain*. New York: McKay.

Bourdieu, P. (1986). The forms of capital. In J. E. Richardson (Ed.), *Handbook of theory of research for the sociology of education* (s. 241–258). Westport, CT: Greenwood Press.

Bourdieu, P., & Passeron, J.-C. (1990). *Reproduction in education, society and culture* (2nd Ed.). (R. Nice, prekl.). London, UK, Thousand Oaks, CA, New Delhi, I: SAGE. [Pôv. dielo vyd. 1970.]

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement, 39*, 253–263.

Çetin, S., & Gelbal, S. (2013). A comparison of bookmark and Angoff standard setting methods. *Educational Sciences: Theory & Practice, 13*(4), 2169–2175.

Hambleton, R. K. (1980). Test score validity and standard-setting methods. In R. A. Berk (Ed.), *Criterion-referenced measurement: The state of the art* (s. 80–123). Baltimore and London: The Johns Hopkins University Press.

Hambleton, R. K., & Pitoniak, M. (2006). Setting performance standards. In R. L. Brennan (Ed.), *Educational measurement* (s. 433–470). Westport, CT: Praeger Publishers.

Chráska, M. (2009). Testování výkonů ve vzdělávání. In J. Průcha (Ed.), *Pedagogická encyklopedie* (s. 594–598). Praha: Portál.

Impara, J. C., & Plake, B. S. (1998). Teachers' ability to estimate item difficulty: A test of the assumptions in the Angoff standard setting method. *Journal of Educational Measurement, 35*(1), 69–81.

Jacob, B. A. (2005), Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools. *Journal of Public Economics, 89*(5–6), 761–796.

Kaščák, O., & Pupala, B. (2012). *Škola zlatých golierov. Vzdelávanie v ére neoliberalizmu*. Praha: Sociologické nakladatelství (SLON).

Livingston, S. A., & Zieky, M. J. (1982). *Passing scores: A manual for setting standards of performance on educational and occupational tests*. Princeton, NJ: Educational Testing Service.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*, 121–141.

MCAS (2013a). *Massachusetts Comprehensive Assessment System. MCAS achievement level definitions*. Dostupné z http://www.doe.mass.edu/mcas/tdd/pld/.

MCAS (2013b). *English Language Arts. General performance level definitions*. Dostupné z http://www.doe.mass.edu/mcas/tdd/pld/ela410.pdf.

Mills, C. N., & Jaeger, R. M. (1998). Creating descriptions of desired student achievement when setting performance standards. In L. Hansche (Ed.), *Handbook for the development of performance standards: Meeting the requirements of Title I* (s. 73–85). Washington, DC: Council of Chief State School Officers.

Minarechová, M. (2012). Negative impacts of high-stakes testing. *Journal of Pedagogy / Pedagogický časopis*, *3*(1), 82–100.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure. Psychological perspectives. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (s. 249–281). Mahwah, NJ: Erlbaum.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational and Psychological Measurement, 14*, 3–19.

Näsström, G., & Nyström, P. (2008). A comparison of two different methods for setting performance standards for a test with constructed-response items. *Practical Assessment, Research & Evaluation, 13*(9). Dostupné z http://pareonline.net/pdf/v13n9.pdf.

OECD (2009). *PISA 2006 technical report*. Paris: OECD.

Olsen, J. B., & Smith, R. (2008). *Cross validating modified Angoff and bookmark standard setting for a home inspection certification*. Príspevok prednesený na stretnutí National Council on Measurement in Education, New York. Dostupné z http://siterepository.s3.amazonaws.com/00373201006251026068636.pdf.

Rasch, G. (1980). *Probabilistic Models for Some Intelligence and Attainment Tests* (Exp. ed.). Chicago and London: The University of Chicago Press.

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard setting. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (s. 119–157). Mahwah, NJ: Erlbaum.

Reckase, M. D., & Bay, L. (1999). *Comparing two methods for collecting test-based judgements*. Príspevok prednesený na stretnutí National Council on Measurement in Education, Montréal, QC. Dostupné z https://www.measuredprogress.org/documents/10157/19213/ComparingTwoMethods.pdf.

Rosenkvist, M. A. (2010). Using student test results for accountability and improvement: A literature review. *OECD Education Working Paper No. 54, EDU/WKP(2010)17,* 22. Nov. 2010. Paris: OECD, Directorate for Education.

Rudner, L. M. (2009). Scoring and classifying examinees using measurement decision theory. *Practical Assessment, Research & Evaluation, 14*(8). Dostupné z http://pareonline.net/getvn.asp?v=14&n=8.

SCIO (n.d.). *Scate. Souhrnná zpráva z testování 2013/2014*. Dostupné z https://www.scio.cz/download/skoly/SCATE/SZ_AJ_final.pdf.

Shepard, L. A. (1995). Implications for standard setting of the National Academy of Education evaluation of the National Assessment of Educational Progress achievement levels. In *Proceedings of the joint conference on standard setting for large scale assessments of the National Assessment Governing Board (NAGB) and the National Centre for Educational Statistics (NCES), Vol. 2* (s. 143–160). Washington, DC: Government Printing Office.

Slavík, J. (1999). *Hodnocení v současné škole. Východiska a nové metody pro praxi*. Praha: Portál.

ŠPÚ (2009). Vzdelávacie štandardy zo slovenského jazyka a literatúry pre 2. stupeň základných škôl a 1.–4. ročník gymnázií s osemročným štúdiom. *Štátny vzdelávací program Slovenský jazyk a literatúra (vzdelávacia oblasť Jazyk a komunikácia). Príloha ISCED 2*. Bratislava: Štátny pedagogický ústav. Dostupné z http://www.statpedu.sk/files/documents/svp/2stzs/isced2/vzdelavacie_oblasti/slovensky_jazyk_a_literatura_isced2.pdf.

Trna, J. (1996). Vzdělávací standardy pro základní a střední školy. *Pedagogika, 46*, 349–353.

Urbánek, T., Denglerová, D., & Širůček, J. (2011). *Psychometrika: Měření v psychologii*. Praha: Portál.

West, A. (2010). High stakes testing, accountability, incentives and consequences in English schools. *Policy & Politics, 38*(1), 23–39.

Wiersma, W., & Jurs, S. G. (1990). *Educational measurement and testing* (2nd ed.). Boston: Allyn and Bacon.

Wilcox, R. R. (2012). *Introduction to robust estimation and hypothesis testing* (3rd ed.). Waltham, MA: Academic Press.

Williams, N. J., & Schulz, E. M. (2005). *An investigation of response probability (RP) values used in standard setting*. Príspevok prednesený na stretnutí National Council on Measurement in Education, Montréal, QC.

Zákon č. 245/2008 Z. z. o výchove a vzdelávaní (školský zákon) a o zmene a doplnení niektorých zákonov. Dostupné z http://www.uips.sk/sub/uips.sk/images/PKvs/z245_2008.pdf.

Mgr. Andrej Mentel, PhD., Ústav sociálnej antropológie
Fakulta sociálnych a ekonomických vied, Univerzita Komenského,
Mlynské luhy 4, 821 05 Bratislava, Slovensko
andrej.mentel@gmail.com