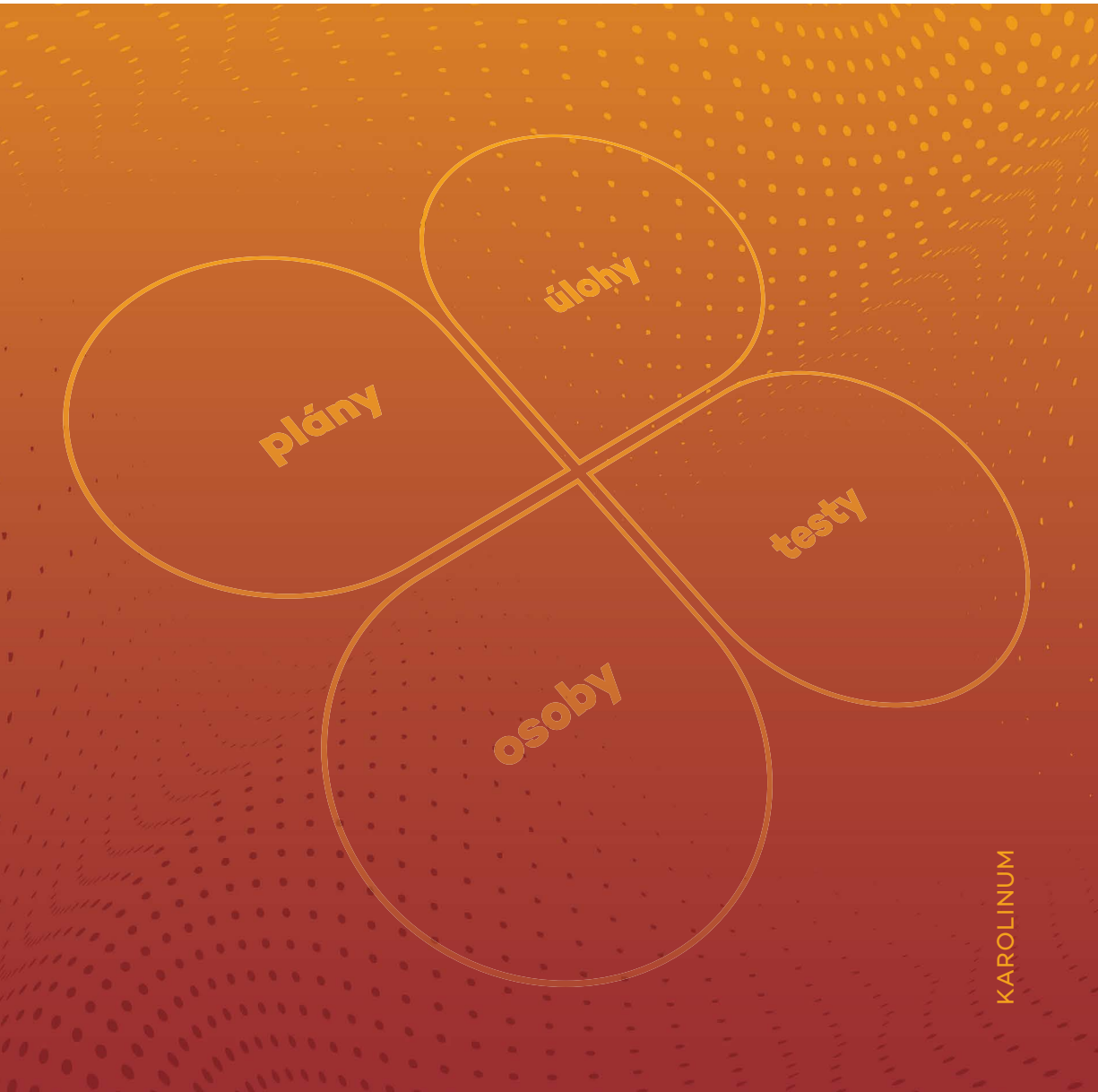
Čestmír Štuka Martin Vejražka

Testování a hodnocení studentů na VŠ



KAROLINUM

Testování a hodnocení studentů na VŠ

Čestmír Štuka, Martin Vejražka

———

Recenzovali:
PhDr. Eva Řídká, CSc.
PhDr. Iva Štětovská, Ph.D.
Mgr. Linda Nepivodová, Ph.D.

Publikace byla vydána za podpory Ministerstva školství, mládeže a tělovýchovy v rámci centralizovaného rozvojového projektu MŠMTC-13/2021 Posilování akademické integrity studujících vysokých škol se zaměřením na rizika a příležitosti distančních metod vzdělávání a hodnocení.

Vydala Univerzita Karlova, Nakladatelství Karolinum
Praha 2021
Redakce Dita Křišťanová
Grafická úprava Lukáš Kejha
Sazba DTP Nakladatelství Karolinum
Vydání první

ISBN 978-80-246-5107-1
ISBN (pdf) 978-80-246-5108-8

https://doi.org/10.14712/9788024651088

Univerzita Karlova
Nakladatelství Karolinum

www.karolinum.cz
ebooks@karolinum.cz

OBSAH

1 ÚVOD

1.1 Předmluva

Moderní metody hodnocení znalostí studentů patří v akademickém světě k velmi diskutovaným tématům. Hodnocení je formou komunikace mezi učitelem a studentem. Je to příležitost sdělit, co učitel považuje ve výuce za podstatné.

Spravedlivé, prokazatelné a objektivní hodnocení znalostí a dovedností studentů je základem jak pro posouzení výkonů studentů a kvality výuky, tak pro motivaci studentů k dalšímu studiu.

Vzhledem k tomu, že testy, jako součást hodnocení, mají významný vliv na učení studentů, je důležité testy sladit se vzdělávacími a učebními cíli.

U testů velkého významu, kdy se např. rozhoduje o postupu do dalšího studia, je důležité zajistit přesnost a spolehlivost testů. Kniha přináší souhrn postupů, doporučení a metod pro tvorbu kvalitních didaktických testů s následnou analýzou jejich výsledků. Příručka je primárně zaměřena na členy akademické komunity odpovědné za hodnocení studentů a má být praktickým nástrojem, který učiteli pomůže v průběhu celého procesu plánování, vývoje, nasazení a vyhodnocení testů.

Autoři se snaží nabídnout čtenářům prakticky použitelnou metodiku a nástroje pro sestavování a vyhodnocování jejich vlastních testů. Práci s testovými položkami, jejich organizaci v položkových bankách a tématu zabezpečení položek a testů.

Metodiky a nástroje pro přípravu a hodnocení testů určených k objektivnímu posouzení výsledků učení jsou společné napříč obory. Součástí textu jsou vybrané pasáže z publikací uvedených v seznamu použité literatury. Hlavním výchozím zdrojem byla příručka *Testování při výuce medicíny*, kterou vydalo nakladatelství Karolinum (Štuka et al. 2013).

Text je členěn tak, aby provázel postupem přípravy testů a umožnil se v problematice zorientovat. Zájemce o podrobnější studium odkazujeme na články a knihy uvedené v literatuře.

1.2 Role testování na vysokých školách

S kvalitou vysoké školy úzce souvisí kvalita studentů, proto se vysoké školy snaží pro studium vybírat nejlepší uchazeče. Ty pak, jako své studenty, připravují na úkoly, s nimiž se budou potkávat v praxi. Musí přitom ověřovat efektivitu vzdělávacího procesu a zkoumají, jak jsou studenti na svoji budoucí roli připraveni. Je zřejmé, že nejlepším měřítkem efektivnosti vzdělávání je úspěch absolventů v praxi. Taková metrika je sice nejobjektivnější, ale prodleva mezi výukou a jejím hodnocením touto formou by znemožňovala udržování účinné zpětné vazby. Abychom mohli výsledky výuky měřit v kratším časovém období, musíme volit jiné cesty. Jednou z nich je testování výsledků výuky, které pokud možno objektivně, reprodukovatelně a spravedlivě hodnotí úroveň dosažených vědomostí a schopností.

Potřebuji testovat. Jak na to?

Potřebujete-li začít testovat:

1. přečtěte si celou tuto knihu,
2. prostudujte základní prameny, z nichž tato kniha čerpá,
3. naučte se používat statistický jazyk R

... a už se k samotnému testování ani nedostanete.

Celý testový cyklus je komplexní proces, v němž můžete jednotlivé kroky neustále zdokonalovat. Nakonec se stanete uznávaným odborníkem v psychometrii. A proto: potřebujete-li začít testovat, **začněte hned**. A můžete začít právě psaním testových položek. Nalistujte v této knize kapitolu, která vám pomůže. Vlastní praxe vás posune vpřed více než sebelepší teorie.

1.3 Typologie testů

Aby test fungoval, jak si představujeme, musíme si nejprve ujasnit, co od něj očekáváme. V různých situacích se používají různé typy testů. Všechny jsou si v něčem podobné, pokaždé je ale kladen důraz na jiné vlastnosti a na některé vlastnosti se naopak rezignuje.

Podle toho, jaké **fáze výuky** je test součástí a čemu má ve výuce napomoci, můžeme testy dělit na **formativní** a **sumativní**. Účelem **formativního hodnocení** je především poskytnutí zpětné vazby studentům i učiteli o průběhu výuky. Test se stává součástí *výukového dialogu*, podporuje aktivní zapojení studentů do výuky, přispívá k jejich motivaci. Student se dozví, nakolik znalosti a dovednosti, které získal, odpovídají nárokům kurzu, a vyzkouší si jejich využití. Test mu pomůže identifikovat silné stránky i oblasti, na kterých musí ještě pracovat. Učiteli výsledky formativního testování pomáhají zefektivnit výuku, neboť mu ukáží, kterým oblastem je třeba se věnovat více a kde by to naopak bylo již zbytečné. Pro obě strany by mělo být formativní testování především orientační. Na formativní test se nekladou tak vysoké procedurální nároky. V některých případech může dokonce nedokonalost formativního testu výuce pomoci, neboť stimuluje diskusi a zapojení všech zúčastněných.

Cílem **sumativního hodnocení** je naproti tomu poskytnout celkový obraz o výsledku učení. Výsledky sumativních testů jsou často podkladem pro další postup ve studiu nebo kariéře. K tomuto typu hodnocení se nejčastěji přistupuje po nějaké celistvé části výuky nebo na závěr kurzu, anebo naopak může jít o testy, kterými se ověřuje způsobilost uchazeče nastoupit do kurzu či zahájit určitou práci.

V praxi jsou čistě formativní a čistě sumativní testy jen extrémy na spojité škále. Často se setkáme s tím, že i test, který je především formativní, se nějakým způsobem započítává do celkového hodnocení studenta a další postup se podmiňuje dosažením určitých výsledků. A obráceně, i sumativní testy a zkoušky by ve většině případů měly zahrnovat poskytnutí zpětné vazby studentovi i učiteli a pomáhat tak ke zkvalitňování kurzu a rozvoji studijních dovedností.

Při přípravě testu je také potřeba zvážit, v jakém **rozsahu** a do jaké hloubky se mají hodnotit dosažené znalosti a dovednosti.

Nejnáročnější jsou v tomto směru **testy a zkoušky (odborné) způsobilosti** (*proficiency tests*), hodnotící celkovou schopnost vykonávat nějakou činnost, například komunikovat cizím jazykem, provádět určité výkony apod. Zkoušky odborné způsobilosti většinou vyžadují nástroje praktického zkoušení (*workplace based assessment*) a samostatné písemné testování lze použít jen ve specifických případech nebo jako dílčí součást zkoušky způsobilosti.

Didaktické testy (*achievement tests*) hodnotí, do jaké míry student zvládl část kurzu nebo určitý úsek studia.

Cílem **diagnostických testů** je podrobněji popsat, jaká jsou silná a slabá místa testovaného.

Konečně **prognostické testy** a **testy studijních předpokladů** (*aptitude tests*) mají odhadnout, do jaké míry je testovaná osoba schopna úspěšně absolvovat určitý kurz a získat v něm cílové kompetence. Například Modern Language Aptitude Test (MLAT) měří studentův potenciál pro úspěšné zvládnutí cizích jazyků, Scholastic Aptitude Test (SAT) hodnotí akademické schopnosti a potenciál vystudovat vyšší nebo vysokou školu.

Standardizované provedení a hodnocení má zajistit prokazatelnost, reprodukovatelnost a dlouhodobou stabilitu výsledků nejdůležitějších zkoušek. U standardizovaného testu musí být zaručeno, že výsledek závisí především na schopnostech zkoušeného, nikoli na konkrétní variantě testu, prostředí, ve kterém se test píše, dohledu či hodnotitelích.

Pojem „standardizace“ má v psychometrii celou řadu významů a podrobněji se jím budeme zabývat později v samostatné kapitole. Standardizace testu je požadována zejména tam, kde výstupem zkoušky bývá nějaký uznávaný certifikát, či je výsledek důležitý pro další kariéru testovaného. Součástí standardizovaného testování je sběr dat o testování a jejich statistické zpracování, mimo jiné s cílem odhalit „nestandardní jevy“ (opisování, únik položek, napovídání...). Mezi základní nástroje používané při přípravě standardizovaných testů patří použití kalibrovaných testových položek, jejichž psychometrické charakteristiky se kombinují tak, aby test jako celek měl požadované vlastnosti. Pro kontrolu porovnatelnosti mezi jednotlivými běhy testu se používá **kotvících položek**, díky kterým je možno porovnat obtížnost testů v jednotlivých termínech. Ke standardizaci patří i objektivní nastavení meze pro průchod testem a důsledné zajištění srovnatelných

podmínek pro všechny testované. Standardizované testy musí být připraveny a provedeny tak, aby objektivitu jejich výsledků bylo možné prokázat, např. i před soudem. Požadavkem na standardizaci prudce rostou náklady. Vždy je proto třeba zvážit, kde jsou tyto náklady odůvodněné a kde by stačilo použít běžné, nestandardizované zkoušky a testy.

U **nestandardizované** zkoušky více záleží na konkrétním zkoušejícím nebo hodnotiteli. Ti se často soustředí více na individualitu zkoušeného a mohou lépe posoudit jeho osobní předpoklady a dosažené kompetence. Pro porovnávání zkoušených mezi sebou se však nehodí.

V situacích, kdy by byla standardizace neúčelná, nebo dokonce neproveditelná (například pro příliš malý počet zkoušených), se často provádějí kroky, které vedou k **objektivizaci** nestandardizovaného hodnocení a tím ke snížení nežádoucích vlivů na hodnocení, především subjektivity zkoušejícího.

Převažující formou testování byly vždy testy **prezenční**, při nichž jsou studenti v přímém kontaktu s vyučujícím. V posledních letech nabyly kvůli pandemii na významu testy a zkoušky vedené **distančně** a došlo k rozvoji forem testování, které nevyžadují bezprostřední kontakt učitele a studenta. Zásadního pokroku se dosáhlo v metodách **proktorovaného testování**, rozvíjí se také **testování s otevřenou knihou**.

1.4 Meze testování

Přestože má vzdělávání na univerzitách a obecně vysokých školách za sebou již několikasetletou historii, stále nepanuje jasná shoda, co je vlastně jeho cílem (Council on Higher Education 2013). Pravděpodobně je takových cílů více a záleží i na zaměření dané vysoké školy. V obecné rovině asi můžeme říci, že absolvent vysoké školy by měl odcházet jako profesionál připravený pro výkon určité profese, povolání či role. Tradiční představa je, že předpokladem k tomu je osvojení si řady znalostí a dovedností. To však samo o sobě nestačí. Absolvent vysoké školy by měl být schopen víceméně samostatně pracovat, tj. vykonávat určité činnosti. Ostatně, získání vysokoškolského titulu je mnohdy s oprávněními pro výkon povolání přímo svázáno.

K výkonu konkrétní činnosti nestačí jen soubor faktografických znalostí. Je třeba určité oblasti rozumět, mít určité dovednosti, ale také zaujímat profesionální postoje. Máme-li garantovat, že vysokoškolský absolvent je schopen odborně práci v daném oboru vykonávat, měli bychom ověřit, že je dostatečně kompetentní nejen co do znalostí, ale i v odpovídajících „vyšších“ úrovních.

1.4.1 Dimenze znalostí, dovedností a postojů

Když se popisují cíle výuky, například při tvorbě anotace nějakého předmětu, nebo při jeho členění do sylabu, vychází se obvykle z jeho věcného obsahu – seznamu témat, která chceme vyučovat. Takové členění ale samo o sobě nestačí. Pro kvalitně vedenou výuku a poté i pro hodnocení jejích výsledků je užitečné k seznamu tematických okruhů přidat ještě druhý rozměr – rozčlenit každé téma do několika úrovní podle komplexity vzdělávacích cílů.

Nejpoužívanějším modelem, který komplexitu cílů vzdělávání, výchovy a odborné přípravy popisuje, je **Bloomova taxonomie**. Jde vlastně ne o jeden, ale o tři hierarchické modely (označované také jako *domény* nebo *okruhy*):

- **kognitivní** doména čili okruh **vzdělávacích cílů**,
- **afektivní** doména čili **emoční okruh** a
- **psychomotorická** doména čili **senzomotorický okruh**.

Ve vzdělávání se nejčastěji pracuje s prvním okruhem, tj. okruhem vzdělávacích cílů, který odpovídá **znalostem**, schopnosti je zapojovat a využívat. Mnozí autoři ale opakovaně upozorňují, že neméně významnou, byť hůře uchopitelnou součástí vzdělávání jsou i zbylé dva okruhy. Vzdělávání či výchova v emočním okruhu vede k rozvoji profesionálních **postojů**. Psychomotorická doména pak zahrnuje osvojení praktických **dovedností**.

Bloomova taxonomie se v současnosti používá v několika verzích. V 90. letech 20. století byla rozsáhle revidována a vznikla dvourozměrná mapa. Namísto tří okruhů tato revidovaná verze pracuje se čtyřmi *znalostními dimenzemi* – faktuální, konceptuální, procedurální a metakognitivní znalostí. Každá z těchto znalostních dimenzí pak obsahuje šest *dimenzí poznávání*: zapamatovat, rozumět, aplikovat, analyzovat, hodnotit a tvořit.

Vzhledem k tomu, že se termínem Bloomova taxonomie označuje několik různých pojetí a verzí, dovolíme si podrobněji pracovat jen s jednou z nich, tradičně označovanou jako **Bloomova taxonomie vzdělávacích cílů**. Víceméně odpovídá kognitivní doméně původní Bloomovy taxonomie z přelomu 50. a 60. let 20. století a dimenzi faktuálních znalostí z výše zmiňované revize.

1.4.2 Bloomova taxonomie vzdělávacích cílů

Taxonomie v okruhu vzdělávacích cílů popisuje úrovně schopností a dovedností, které se týkají faktických znalostí, jejich pochopení a porozumění kontextu. Má šest úrovní: nejnižší je znalost, pak pochopení, aplikace, analýza, hodnocení a nejvýše tvorba. Tradiční vzdělávání má sklon pracovat právě s tímto souborem cílů, především pak s jeho nižšími úrovněmi.

Znalost (též zapamatování)

Nejnižší úrovní vzdělávacích cílů je **znalost**, tj. **zapamatování** a **schopnost vybavit** si fakta a nejzákladnější koncepty. Student se učí základní pojmy, definice. Převážně po něm chceme, aby něco vyjmenoval, zopakoval, vysvětlil pojem, něco zařadil do určitého klasifikačního schématu apod. To může učinit i bez plného porozumění pojmům, se kterými se pracuje – student například může správně zařadit rostlinný druh do čeledi čistě díky tomu, že má naučené, jaké druhy do příslušné čeledi patří. Nemusí přitom vůbec tušit, jak rostlina, o které mluví, vypadá, jaké jsou charakteristiky příslušné čeledi a proč vlastně daná rostlina do té čeledi patří. Znalosti v této úrovni mají často povahu izolovaných údajů z encyklopedického slovníku.

Pochopení (též porozumění)

Naplnění tohoto vzdělávacího cíle lze typicky demonstrovat jako schopnost vysvětlit či interpretovat určitou látku. Student, který dosáhl porozumění, dokáže **vysvětlit hlavní myšlenku**.

S porozuměním také dokáže něco popsat, porovnat, seřadit, přeložit do jiného jazyka, který ovládá, apod. Přestože dané věci rozumí, nemusí ji ještě umět využít – typicky se to odráží v tom, že ji nedokáže kombinovat s jinými znalostmi, které rovněž má a kterým rozumí. Student tak například chápe pohyb Země kolem Slunce a jeho souvislost se střídáním ročních období, stejně tak je mu dobře známý tvar Země, nedokáže si ale poradit s otázkou, jaká je poloha slunce nad obzorem v různých zeměpisných oblastech v době letního slunovratu.

Aplikace

Pro dosažení této úrovně je typická schopnost **využít nabyté znalosti v nových situacích**. Student dokáže řešit úkoly a problémy, se kterými se dosud nesetkal. Musí k tomu rozpoznat souvislosti a vztahy a najít cestu, jak jich k řešení využít. Mnohdy musí využít pravidel či postupů, které již zná, ale jinak, než byl dosud zvyklý. Dokáže však pracovat jen s relativně nekomplikovaným zadáním. Na řešení komplikovaně zadané úlohy není tato úroveň postačující, neboť např. ještě nerozliší, které údaje jsou pro řešení podstatné, které bezvýznamné a jaké údaje případně chybí a musí je ještě zjistit. Komplikovanější problém je proto pro něj neřešitelný, přestože by měl potřebné dílčí znalosti, neboť se v něm ztratí.

Analýza

Analýzou se v Bloomově taxonomii myslí rozbor složitého celku nebo problému na **menší části**, což umožní jeho **lepší pochopení**. Analytické dovednosti jsou potřebné například pro rozlišení příčin a následků nebo pro hledání důkazů, které podporují nějaké zobecňující tvrzení. Do této úrovně patří také schopnost rozpoznat strukturu nějaké informace, rozčlenit ji na jednotlivé součásti, posoudit vztahy mezi nimi a díky tomu odhadnout důvěryhodnost informačního zdroje. Dosažení tohoto vzdělávacího cíle lze také demonstrovat třeba myšlenkovým experimentem, ve kterém student dokáže odhadnout, jak by určitý zásah změnil nějaký děj. Musí k tomu děj analyzovat, rozpoznat jeho součásti a vazby mezi nimi, určit, co se přesně změní zvažovaným zásahem a k jakým to povede důsledkům.

Hodnocení

Hodnocení patří mezi nejvyšší cíle v Bloomově taxonomii. Myslí se jím schopnost posuzovat informace a na základě toho **přijímat informovaná rozhodnutí**, zaujímat stanoviska nebo hájit názory. Hodnocení vyžaduje schopnost informaci nejprve analyzovat, je tedy provázané s předchozím cílem. Podstatou hodnocení je zhodnotit jednotlivé části informace a posoudit jejich význam a validitu. Výsledný verdikt by přitom měl být podstatný pro řešení určitého problému nebo pro tvorbu něčeho nového. Typickým úkolem, který demonstruje dosažení tohoto cíle, může být např. zpracování odborné recenze vědeckého článku, včetně doporučení k jeho publikování, odmítnutí či úpravám. Je zřejmé, že ke splnění takového úkolu je zapotřebí nejen dosažení všech předchozích cílů, ale i určité kreativity. Není tedy divu, že v různých verzích Bloomovy taxonomie se pořadí dvou nejvyšších cílů někdy liší.

Tvorba

Tvorba je vrcholným cílem, jehož může student učením dosáhnout. Je to cíl vysloveně kreativní, produktivní. Při dosažení tohoto cíle dokáže student **navrhnout nové originální řešení**, projektovat, vynalézat apod.

Bloomova taxonomie nás nutí zamyslet se nad způsobem, jakým se člověk učí, a díky tomu je cenná i k úvahám, jak výsledky učení hodnotit. Přestože jde o nesporně nepoužívanější „rozškatulkování" vzdělávacího procesu, má i své nevýhody a kritiky. Již z předchozího popisu je zřejmé, že zatímco charakteristika nižších kategorií Bloomovy taxonomie je vcelku jednoznačná a snadno v praxi použitelná, vyšší cíle jsou definovány stále abstraktněji, méně jednoznačně, a dokonce jsou pochybnosti, jak je přesně uspořádat. Omezení Bloomovy taxonomie je více (Soozandehfar a Adeli 2016, Case 2013, Tutkun et al. 2012), např.:

- Bloomova taxonomie předpokládá, že se člověk učí lineárně, sekvenčně – že postupuje od nejnižších cílů k vyšším. Ve skutečnosti tomu tak není, mezi jednotlivými „patry" se přeskakuje a opakovaně se vrací zpět k nižším úrovním. Naučí-li se člověk něco zhodnotit nebo vytvořit, výsledek své práce pak znovu analyzuje, doplňuje si znalosti a učí se jim porozumět apod.
- Učit se člověk může i od špičky pyramidy směrem k její základně, tím, že něco vytvoří. Začnou se pak uplatňovat jiné procesy, které Bloomova taxonomie příliš nepopisuje: průzkum, zkoušení nějakého řešení, vytváření prototypu, revize či kritické posuzování. To teprve nutí k vyhledávání zdrojů a získávání nových faktických znalostí (Berger 2018).
- Bloomova taxonomie předpokládá, že se člověk učí izolovaně od ostatních, je individualistická. Pomíjí sociální a konektivistické aspekty učení, které jsou ve vysokoškolském vzdělávání velmi důležité (Teachers Commons 2008).

1.4.3 Jiné taxonomie

Bloomova taxonomie je relativně složitá – připomeňme, že jsme v textu výše pracovali jen s jedním z okruhů znalostních dimenzí, takže jsme zcela pominuli vše, co je spojené s procedurálními dovednostmi či postoji. A to společně s její neúplností vedlo ke vzniku jiných, různě pojímaných modelů učení a úrovní dosahovaných kompetencí (O´Neil 2010; Malamed 2020).

Specificky pro hodnocení znalostí a dovedností se v devadesátých letech 20. století začala v medicíně používat tzv. **Millerova pyramida** (Miller 1990), která se postupně rozšířila i do jiných oborů (Peñalver 2015, Krevič 2019). Původní čtyři úrovně hodnocených kompetencí byly později doplněny o úroveň pátou (Cruess 2016).

- **Znalost** (student **zná**; v anglicky psané literatuře *knowledge*, úroveň *knows*).
- **Porozumění** (student **ví jak**; *competence*, *knows how*). Zkoušený dokáže znalosti z předchozí úrovně **zapojit** do kontextu.
- **Dovednost** (dokáže **ukázat jak**; *performance*, *shows how*). Dovednost je již komplexní, zkoušený se „sám vyzná" a kombinuje široké spektrum znalostí a schopností, kterých často nabyl v různých předmětech a částech studia.
- **Činnost** (v praxi **provádí** správně veškeré potřebné úkony; *action*, *does*). Této úrovně by měl dosáhnout např. kandidát u státní závěrečné zkoušky.
- **Identita** (skutečně **je** profesionálem). U člověka, který dosáhl této úrovně, lze konzistentně pozorovat postoje, hodnoty a chování, které se očekávají od zástupce určité profesní skupiny. Lze říci, že daný člověk *myslí, chová se a cítí se jako* lékař, učitel, právník, projektant apod.

Přistoupíme-li na to, že absolvent vysoké školy má být profesionál připravený k výkonu určité činnosti, měli bychom v průběhu či nejpozději v závěru studia ověřit, že skutečně nabyl příslušných kompetencí. Zatímco znalosti a porozumění obvykle můžeme velmi dobře hodnotit pomocí standardizovaných testů i nestandardizovaných metod (třeba ústním zkoušením), hodnocení nejvyšších úrovní kompetencí je složitější. Jistě bychom považovali za absurdní, kdyby např. orchestr přijímal houslistu na základě písemného testu nebo ústní zkoušky z houslové hry. Stejně absurdní by bylo na základě pouhého písemného testu či ústní zkoušky o někom prohlásit, že je např. dobře připraveným učitelem, právníkem, historikem nebo lékařem. Ověřili bychom sice s větší či menší věrohodností, že nabyl znalostí a porozumění, jež jsou nezbytným předpokladem pro výkon povolání, ale nijak bychom nezjišťovali, zda jich také dokáže odpovídajícím způsobem využít, zda nabyl potřebných dovedností a zda skutečně dokáže vykonávat činnosti, které jsou součástí konkrétní práce.

Dobře připraveným a standardizovaným písemným testováním lze kvalitně hodnotit znalosti a porozumění. Hodnocení dovedností písemným testem je už možné jen v konkrétních případech. Můžeme jej použít, pokud je dovedností například schopnost vyřešit matematickou úlohu nebo popsat nějakou reakci chemickými rovnicemi. Většinu dovedností ale písemným testem ani ústní zkouškou hodnotit nelze – těžko bychom takto zkoušeli např. laboratorní úkony, praktickou práci se zeměměřičskými přístroji nebo odběr krve.

Pro ověřování dovedností a činností lze využít techniky, která se obecně označuje jako **praktické zkoušení** (*practical examination*) nebo **hodnocení na pracovišti** (*workplace (based) examination*) (The Royal College of Pathologists 2019; Prakash et al. 2020). Z principu jde o metodu, která může být standardizována jen těžko. Při prakticky zkoušených úkonech totiž nelze jednoduše zajistit stejné (standardizované) podmínky většímu počtu kandidátů. Kromě toho se mnohdy hodnotí i dovednosti, které ve své úplnosti standardizovaně klasifikovat nelze, například je součástí zkoušky i komunikace s klientem nebo týmová spolupráce.

I tyto praktické zkoušky však mohou být **objektivizované**, tj. uspořádané tak, aby byl potlačen vliv nežádoucích faktorů – například subjektivity zkoušejícího nebo variability podmínek, za nichž zkouška probíhá. Dalším krokem je pak validace praktických zkoušek, tj. ověření, že výsledek zkoušky skutečně odpovídá získaným dovednostem potřebným pro reálnou praxi.

Objektivizované metody praktického zkoušení mívají několik typických charakteristik:

- Sleduje se dlouhodobý nebo opakovaný výkon, nikoli jednorázový výkon v rámci jednoho testového sezení. Pokud praktická zkouška probíhá v krátkém časovém úseku (např. během jediného dne), je rozdělena do několika samostatných částí (označovaných často jako *stanice*). Každou z nich posuzují jiní hodnotitelé a každá je zaměřená na jiných okruh dovedností a činností.
- Zkoušeného hodnotí nezávisle na sobě větší počet hodnotitelů. Mezi hodnotiteli jsou odborníci v dané oblasti, ale často i další osoby – například spolužáci, figuranti, kteří v průběhu zkoušky vedou se zkoušeným modelovou komunikaci, někdy i technický personál.
- Zkouška a její hodnocení jsou strukturované, tj. hodnotitelé se vyjadřují podle předem daných kritérií ke sledovaným aspektům výkonu (tzv. *rubriky* zkoušky).
- Zkouška je validovaná jako celek, nebo jsou validované její jednotlivé části.

Velký posun ve formátech praktického zkoušení přineslo zavedení tzv. **objektivních strukturovaných klinických zkoušek** (*objective structured clinical examination*, **OSCE**) v medicíně v polovině 70. let 20. století. O dvacet let později se tento přístup začal používat i v jiných oborech, někdy se proto označuje i jako *objektivní strukturované praktické zkoušení*, **OSPE**. V průběhu OSCE studenti procházejí řadou stanic, v nichž jsou konfrontováni s běžnými situacemi každodenní praxe a mají provést určitou proceduru. Jsou hodnoceni pomocí strukturovaného dotazníku, který vyplňují přítomní hodnotitelé. Větší váhu mají kroky a postupy obecnější povahy (v medicíně např. prevence šíření infekce, komunikace s pacientem v průběhu výkonu, poučení pacienta a vysvětlení výkonu apod.), menší váhu pak mají úkony, které jsou úzce specifické.

Jiným způsobem, jak hodnotit dosažení vyšších vzdělávacích cílů, je sestavování **portfolií**. Portfolio je systematicky tvořený soubor ukázek práce studenta, který demonstruje jeho úsilí, pokrok a dosažení cílů vzdělávání v průběhu celého kurzu nebo i kurikula (Klenowski et al. 2006, Seifert 2011, Herman a Zuniga nedatováno, Hill nedatováno). Jde o vysloveně konstruktivisticky pojatý nástroj hodnocení. Jeho výhodou je, že věrně odráží dosažení nejvyšších, kreativních vzdělávacích cílů a také získání profesionálních návyků a postojů, hodnotových žebříčků apod. Na druhou stranu jde o vysoce individualizovaný hodnotící nástroj, který neumožňuje standardizaci, a obtížná je i jeho objektivizace. Ve srovnání s jinými nástroji jsou portfolia poměrně náročná na čas studenta i učitele. Zavedení hodnocení pomocí portfolia vyžaduje pečlivou přípravu a dokonalou integraci s kurikulem (Driessen et al. 2007).

1.4.4 Kvantitativní a kvalitativní formy hodnocení

V předchozím textu jsme vycházeli především z pohledu na hodnocení výsledku vzdělávání jako na **měření**, do jaké míry student dosáhl očekávaných znalostí a dovedností. Výsledkem hodnocení je v tomto pojetí určitá kvantita, známka nebo číselná hodnota. Půvab tohoto pojetí spočívá v jeho pochopitelnosti a v tom, že lze oprávněnost závěrů, které učiníme, snadno zkoumat vědeckými metodami. Můžeme mluvit o přesnosti a spolehlivosti takových závěrů, kde míru přesnosti můžeme kvantifikovat statistickými metodami, můžeme dokonce měřit míru nejistoty, s jakou výsledek sdělujeme. Můžeme také sledovat dopady každé změny, kterou v rámci výuky a zkoušení uděláme. Vše dohromady nám umožňuje zkoušky standardizovat – zajistit, že výsledky hodnocení jsou podložené, reprodukovatelné, objektivní a validní.

Už hierarchizace výukových cílů podle nejběžnějšího pojetí Bloomovy taxonomie ale ukazuje, že standardizovaným testováním a zkoušením nelze hodnotit celou šíři vysokoškolského vzdělávání. Ukázali jsme, že dokáže postihnout jen nižší úrovně poznání. Vyšší vzdělávací cíle lze hodnotit sice nestandardizovanými, ale stále ještě objektivizovanými metodami a výsledkem pořád může být nějaká hodnota. Pro hodnocení nejkomplexnějších cílů, dovedností a postojů nám však pouhé jednorozměrné vyjádření nestačí.

Komplexní kompetence, postoje, chování či dosažení profesionality v určité oblasti nemůžeme měřit číslem. Přestože nejsou měřitelné (nebo je aspoň nelze vyjádřit jednorozměrnou veličinou), jsou popsatelné. Lze je popsat slovním hodnocením. To ale má vždy subjektivní složku, jde typicky o nestandardizované hodnocení.

Přístupy k hodnocení výsledků vzdělávání se neustále vyvíjejí. Ve světě se v minulých desetiletích intenzivně rozvíjelo standardizované hodnocení a ve vyspělých zemích se stalo nedílnou součástí vzdělávací činnosti. Díky přezkoumatelnosti a obhajitelnosti výsledků v mnoha oblastech postupně zcela vytlačilo nestandardizované metody. V posledních letech se ale upozorňuje na limity takového přístupu (Vleuten 2019a). Standardizované metody zůstávají nejlepším známým nástrojem v určitých fázích výuky, někteří autoři ale upozorňují, že by neměly být nástrojem jediným (Fairtest 2007, Riffert 2005). Nestandardizované metody tedy nejsou méněcenné – nejsou ale ani nadřazené standardizovaným. Jde o různé nástroje, z nichž každý se hodí k něčemu jinému.

2 PLÁNOVÁNÍ TESTU

Test má kvalitně hodnotit výsledky výuky a často je také sám součástí výuky jako takové. Pro přípravu výuky a jejího následného hodnocení je proto nutné, abychom dokázali co nejpřesněji definovat, co má absolvent umět, tj. jaké jsou **cíle výuky** (*learning objectives*). V praxi se však skutečná výuka těmto cílům jen blíží. V některých oblastech studenti dokonce cílů výuky zcela nedosáhnou, současně ale mnohdy získají jiné, neplánované, znalosti a dovednosti. Tomu, co absolvent skutečně umí, říkáme **výstupy výuky** (*learning outcomes*).

Test, kterým ověřujeme, zda absolvent dosáhl požadovaných znalostí a dovedností, by měl svým obsahem i rozsahem co nejlépe odpovídat cílům výuky a současně by testu měly co nejvíce odpovídat i výstupy výuky. V praxi nikdy nebude shoda úplná. Aby byla co nejlepší, je třeba řádně plánovat jak výuku, tak i test. Pokud studenti znají cíle, kterých mají dosáhnout, a vědí, jak bude jejich výkon hodnocen, jsou motivovanější ke studiu (Herman 1992). Pokud se naopak obsah testu nebo zkoušky liší od skutečné náplně (tj. výstupů) výuky, cítí se studenti podvedeni a označují zkoušení za nespravedlivé a motivaci k učení často nahrazují získáním „v testu požadovaných odpovědí“.

Dalším důvodem, proč pečlivě plánovat testy, je skutečnost, že mnohdy nemůžeme vyzkoušet všechny studenty najednou a potřebujeme vytvořit více paralelních forem (variant) testu. Pokud při tvorbě všech variant postupujeme podle stejného, dostatečně podrobného plánu, lze se jejich rovnocennosti přiblížit v celku snadno.

2.1 Specifikační tabulka (blueprint)

Jedna z nejosvědčenějších metod, jak připravit plán testu, je konstrukce tzv. **specifikační tabulky** (v anglické literatuře *blueprint*) (Vleuten et al. 1991). Prvním krokem je vytvoření **řádků tabulky**, které odpovídají **obsahovým cílům**. Tento krok je poměrně jednoduchý – vychází se většinou ze sylabu předmětu, pořadí kapitol v základní učebnici apod. Každý řádek odpovídá jednomu tématu. Je vhodné, aby členění témat bylo dostatečně podrobné – jedné přednášce, lekci či kapitole v učebnici tak většinou odpovídá několik řádků s dílčími podtématy.

Sloupce specifikační tabulky odpovídají **aspektům** (přesněji *výukovým doménám*), z nichž se na témata můžeme dívat (Abdellatif a Al-Shahrani 2019). Nalezení takových pohledů, z nichž

lze popisovat většinu obsahových cílů (tj. řádků tabulky) najednou, je klíčový, a přitom většinou nejobtížnější krok při tvorbě specifikační tabulky. Nejobecnější radou je vycházet z cílů Bloomovy taxonomie, např.:

- znalost – např. znalost terminologie, definic, pojmenování určitého fenoménu;
- porozumění – např. porovnávání, interpretace grafů a dat;
- aplikace – např. řešení nového problému pomocí analogie;
- analýza – např. identifikace příčin a důsledků, schopnost vysvětlit určitý jev;
- syntéza – například předpověď výsledku nějakého děje, odhad důsledků.

Jelikož domény vytvořené podle Bloomovy taxonomie jsou poměrně podrobné, a zejména pro tvorbu kratších testů je jich příliš mnoho, používají se někdy jednodušší schémata, např. *vybavení si znalosti – aplikace – řešení problému*. V některých oborech společné aspekty přirozeně vyplývají z vyučované látky a nacházejí se relativně snadno, např. v klinických oborech medicíny lze často sloupce nadepsat *etiologie – příznaky – diagnostika – léčba – prognóza* atd.

V každém případě je třeba tabulku sestavovat tak, aby dávalo smysl co nejvíce kombinací řádků a sloupců. Do jednotlivých **políček** se pak zaznamenává plánovaný **počet úloh**, případně **typ úloh** či způsob zkoušení. Není nutné vyplnit všechna políčka tabulky, nicméně plán testu by měl být vyvážený – neměl by zůstat žádný prázdný, nebo téměř prázdný, řádek ani sloupec, a neměly by zůstávat ani rozsáhlejší prázdné oblasti.

Při vyplňování počtů úloh již bereme v potaz celkový rozsah testu. Může se stát, že některá políčka tabulky, odpovídající méně podstatným znalostem a dovednostem, se využijí jen v některých verzích testu a jiné verze budou namísto nich obsahovat jiné úlohy. V každém případě takto sestavený plán testu pomůže dosáhnout toho, že počet úloh věnovaných určitému tématu odpovídá jeho významu a že se adekvátně zkoušejí všechny aspekty určitého problému (Patil et al. 2015).

Dvourozměrná specifikační tabulka je mnohdy velmi podrobná a rozsáhlá, připomíná proto svými rozměry technický výkres – i proto se pro tento způsob plánování testu vžil anglický termín *blueprinting*. Podrobná specifikační tabulka je zpravidla neveřejná, slouží pouze tvůrcům testu. Její zveřejnění by totiž mohlo vést ke spekulativnímu jednání testovaných, kteří by přípravu více soustředili na test samotný než na získání znalostí a dovedností potřebných pro praxi (Chvál et al. 2015). Na druhou stranu by vždy měl být zveřejněn obsah testu (tj. řádky specifikační tabulky), popřípadě i podíl každé oblasti na celkovém rozsahu testu.

3 TESTOVÉ ÚLOHY

Základním stavebním prvkem každého testu jsou *úlohy* nebo též *položky*. Záměrně se budeme vyhýbat často používanému termínu *otázka*, který jak uvidíme dále, má trochu jiný význam.

Obecně můžeme dovednosti zkoušet *přímo* nebo *nepřímo*. Při **přímém zkoušení** dostane student za úkol přímo vykonat určitou činnost. Přímé zkoušení často využívá technik praktického zkoušení (*workplace based assessment*), některé přímé úlohy ale mohou být i součástí písemných testů.

Příklady:
Napište svůj strukturovaný životopis v anglickém jazyce.
V programovacím jazyce R vytvořte funkci, která...

Pomocí přímých úloh lze dobře hodnotit, do jaké míry kandidát dosáhl cílových kompetencí a nakolik je např. připraven pro výkon určité pracovní činnosti. Jejich nevýhodou je obtížná klasifikace. Výkon kandidáta musí hodnotit několik hodnotitelů, hodnotí se zpravidla několik oblastí či aspektů výkonu. Hodnotitelé musí být předem vyškoleni, pro hodnocení používají strukturovaný hodnotící formulář.

Častější jsou **nepřímé metody** zkoušení. Schopnosti studenta se nezkoušejí přímo, ale posuzují se na základě znalostí a dovedností, které jsou pro konkrétní schopnost nezbytným předpokladem.

Typický písemný test se snaží dosažení cílových kompetencí hodnotit nepřímo. Autoři testu vytvářejí určitý konstrukt, jaké znalosti a dovednosti jsou pro dosažení kompetence podstatné. Jejich zkoušením se snaží odhadnout, zda testovaný mohl cílové kompetence dosáhnout. Z tohoto pohledu tedy nezkoumáme, zda zkoušený skutečně něco dokáže, ale jestli má předpoklady to dokázat.

Pomocí samotných nepřímých metod zkoušení není možné spolehlivě rozhodnout, zda je kandidát schopen samostatně vykonávat určitou činnost, a není jimi možné přímé metody zcela nahradit. Tyto metody jsou však mnohem rychlejší, lépe se vyhodnocují, jsou levnější, snáze se dosáhne jejich reprodukovatelnosti.

Nepřímé testové úlohy mohou být v zásadě dvojího druhu (Chvál et al. 2015):

1. Otevřené úlohy

Zkoušený musí odpověď vytvořit – napsat text, provést výpočet, nakreslit obrázek apod. Odpověď může mít nejrůznější formát i délku – na jedné straně může jít o doplnění chybějícího písmene, na druhé o napsání několikastránkového eseje. Zkoušený svou odpověď tvoří, proto lze tyto úlohy také označit jako **produktivní**.

2. Uzavřené úlohy

Zkoušený vybírá řešení z uzavřeného okruhu možností, které mu byly nabídnuty. Řešení sám nevytváří, má za úkol jej jenom **vybrat** a vyznačit. Nejčastěji volí mezi několika odpověďmi na určitou otázku. Patří sem ale i další typy úloh, v nichž zkoušený vybírá z několika předem definovaných alternativ, např. k sobě má navzájem přiřadit související pojmy, uspořádat položky v určitém pořadí, doplnit v textu na vynechaná místa *i* nebo *y*, nebo rozhodnout, zda se v určité situaci nějaká veličina sníží, zvýší, či zůstane beze změny.

Hranice mezi otevřenými a uzavřenými úlohami není úplně ostrá. V některých případech zkoušený volí odpověď z prakticky neomezeného okruhu možností, nejde však o vysloveně produktivní úlohu. Příkladem mohou být třeba úlohy, ve kterých má student vyznačit (tj. vybrat) určité místo na fotografii mikroskopického preparátu.

Otevřené i uzavřené úlohy mají v testování a ve vysokoškolské výuce své nezastupitelné místo, každá se však hodí k něčemu jinému:

Otevřené úlohy umožňují dobře hodnotit komplexnější dovednosti, především dovednosti produktivní, kreativní povahy (Schindler 2006). Pro formulaci odpovědi jsou studenti nuceni aktivně používat odbornou terminologii. Často je možné sledovat myšlenkový postup, který testovaného vedl k řešení. Při vyhodnocování lze rozpoznat, nakolik testovaní porozuměli zadání a jestli úloha není špatně formulovaná. Otevřené úlohy jsou neocenitelným nástrojem pro průběžné formativní testování během výuky, které má především poskytnout zpětnou vazbu studentům i vyučujícím. Zpětnou vazbu poskytují účinněji než výběrové úlohy a lze je velmi dobře využít jako východisko pro diskusi o probíraných tématech. Otevřené úlohy mohou být také součástí závěrečných sumativních testů, kde se jimi ověřuje dosažení nejen dílčích znalostí a porozumění naučeným faktům, ale i jejich využití a zapojení ve složitějších úkonech. Příprava otevřených úloh pro sumativní testy je ale náročná, stejně jako hodnocení otevřených úloh.

Otevřené úlohy nelze klasifikovat automaticky, odpovědi musí posoudit kvalifikovaní *hodnotitelé*. Může se pak stát, že hodnocení je zatíženo subjektivní chybou. Každou otevřenou úlohu hodnotí postupně několik navzájem nezávislých hodnotitelů, kteří je pokud možno dostávají anonymizované. Pro hodnotitele se předem připravují velmi detailní pokyny. Přesto mohou zkoušení napadat objektivitu testu a může být obtížnější rozhodnutí hodnotitelů nezpochybnitelně odůvodnit. Pokud se tedy otevřené úlohy mají použít v testu, který má zásadní význam, je potřeba tyto úlohy a pravidla, podle kterých se bodují, velmi pečlivě připravit. Otevřené úlohy se často hodnotí na širší škále než jen „správně/nesprávně“ a všichni hodnotitelé musí stejně přidělovat i dílčí body za částečně správné řešení. Obecně platí, že čím je úloha otevřenější, tím je obtížnější její objektivní hodnocení zajistit. Také je nutné mít připravené postupy pro případ, že se názory hodnotitelů na

konkrétní řešení liší. Příprava otevřených úloh proto bývá zdlouhavá a její náročnost roste s významem testu.

Otevřené úlohy mohou znevýhodňovat komunikačně slabší studenty, neboť formulace odpovědi často ovlivňuje hodnocení.

Uzavřené úlohy zahrnují výběrové, přiřazovací a uspořádací úlohy (Schindler 2006). Bývají jednodušší pro zpracování, testy je často možné ohodnotit automaticky počítačem, popřípadě je může opravovat i méně kvalifikovaný pracovník. Ve většině situací jde o nejrychlejší a nejefektivnější nástroj pro zjištění, nakolik si student osvojil znalosti a porozuměl probírané látce. V omezené míře můžeme těmito položkami hodnotit i zvládnutí některých jednoduchých dovedností.

Velkou výhodou uzavřených úloh je, že lze snadno rozhodnout, zda testovaný odpověděl správně. Hodnocení testu je díky tomu reprodukovatelné. Obodování testu je také velmi rychlé. Odpovědi nejsou závislé na formulačních dovednostech testovaných, jejich grafomotorické zdatnosti, rychlosti psaní na klávesnici počítače apod. Naproti tomu uzavřené úlohy neumožňují zkoušet mnoho typů dovedností. Uzavřené úlohy znevýhodňují studenty, kteří jsou méně pozorní, popřípadě méně přesně pracují při stresu.

3.1 Úlohy s výběrem z možností

Výběrové úlohy dnes v písemných testech převažují. Jejich hlavní výhodou je, že se snadno hodnotí. Jak uvidíme dále, mohou mít řadu forem. Všechny spojuje, že zkoušený vybírá jednu nebo více odpovědí z nabídnutých možností. Není rozhodující, jakým způsobem se možnosti nabízejí – může jít např. o zaškrtávání, „radiobutton“, „checkbox“, výběr z roletky.

Z hlediska vlastností i využití v testech je důležité výběrové úlohy rozdělit – bez ohledu na jejich formální vzhled – do dvou skupin:

Dichotomické úlohy (úlohy typu ANO/NE)

Jedna nabízená možnost (či více z nich) je *zcela* správná, ostatní jsou *zcela* chybné.

> **Příklad:**
> Označte, zda tvrzení platí:
> Plejtvák je savec, který žije v moři ANO – NE

Bodování dichotomických úloh je jednoduché. Nejčastěji se za správnou odpověď udělí jeden bod, za chybnou nic. Méně častá jsou skórovací schémata, v nichž se přidělují jiné počty bodů, např. bodový zisk se určí podle náročnosti úlohy, nebo se za chybnou odpověď body strhávají.

Nevýhodou jednotlivých dichotomických úloh je, že prostým tipováním lze získat průměrně 50 % maximálního možného skóre. To na první pohled nemusí vadit, pokud je správně nastavená hranice, které musí student dosáhnout, aby v testu uspěl. Snižuje se tím ale rozlišovací

schopnost testu. Někteří autoři proto doporučují různé modifikace dichotomických úloh, například se vyžaduje, aby spolu s každou odpovědí „NE“ student napsal, jak by se musela otázka změnit, aby na ni byla odpověď „ANO“ (Kubiszyn a Borich 2000). Vzniká tím vlastně kombinace výběrové úlohy s úlohou otevřenou.

Svazky dichotomických úloh

(Též *multiple true/false*, **MTF**; *multiple response question*, MRQ.)

Někdy se několik dichotomických úloh kombinuje do *svazku* se společným kmenem.

Příklad:
Pes je rozšířené domácí zvíře. Chová se mnoho plemen, která se liší velikostí, barvou a povahou. Které tvrzení o psech je pravdivé?
a) Některá plemena psů nemají vůbec žádné chlupy. ANO – NE
b) Bez ohledu na velikost a barvu, všechna plemena psů patří do jediného biologického druhu. ANO – NE

Důležitou vlastností svazků dichotomických úloh (MTF) je, že zkoušený má o každém tvrzení rozhodnout samostatně, nezávisle na ostatních tvrzeních ve svazku. Jinými slovy, výše uvedený svazek dichotomických úloh můžeme rozepsat do dvou samostatných dichotomických úloh:

Úloha 1:
Pes je rozšířené domácí zvíře. Chová se mnoho plemen, která se liší velikostí, barvou a povahou.
Označte, zda platí tvrzení:
Některá plemena psů nemají vůbec žádné chlupy. ANO – NE

Úloha 2:
Pes je rozšířené domácí zvíře. Chová se mnoho plemen, která se liší velikostí, barvou a povahou.
Označte, zda platí tvrzení:
Bez ohledu na velikost a barvu, všechna plemena psů patří do jediného biologického druhu. ANO – NE

Formální vzhled dichotomických úloh a jejich svazků může být různý. Nejčastěji se ke každému tvrzení vybírá odpověď ANO/NE, nebo PRAVDA/NEPRAVDA. Méně vhodné je vyzvat testovaného, aby označil tvrzení, která jsou pravdivá, a nepravdivá tvrzení ponechal neoznačená. V tomto případě se totiž svazek dichotomických úloh (MTF) podobá úlohám s jedinou správnou odpovědí (SBA), které mají ale jiné vlastnosti a neodpovídá se v nich na každou z nabízených možností zvlášť.
Jsou ovšem i další možnosti, kterými se vzájemně se vylučující alternativy označí.

Příklad:
Máme k dispozici pět zkumavek. V každé z nich je 1 ml roztoku jednoho z níže uvedených sacharidů o koncentraci 1 g/l.

Do každé ze zkumavek přidáme 1 ml roztoku hydroxidu draselného (2 g/l) a směs krátce povaříme. Poté do všech zkumavek přidáme roztok s komplexně vázanou dvojmocnou mědí. Výsledná barva směsi v některých zkumavkách je modrá, v jiných červená.
Pro každý sacharid zakroužkujte, jakou barvu směsi po skončení popsaného pokusu očekáváte:

a) amylóza	MODRÁ – ČERVENÁ
b) fruktóza	MODRÁ – ČERVENÁ
c) glukóza	MODRÁ – ČERVENÁ

Od jednotlivých dichotomických úloh se jejich svazek často liší bodováním. Používají se různá skórovací schémata:

1. Vše, nebo nic – pokud jsou všechny odpovědi ve svazku správné, hodnotí se 100 % (nejčastěji 1 bodem), ve všech ostatních případech 0 body.
2. Dílčí skóre – každá dílčí dichotomická otázka se boduje nezávisle např. 0,25 body.
3. Dílčí vážená skóre – každá dílčí dichotomická otázka se boduje nezávisle, každá má jinou bodovou hodnotu (např. dle významnosti nebo obtížnosti).
4. Bodování s penalizací – za některé nesprávné odpovědi se dává záporný počet bodů.
5. Parciální skórování – např. PS_{50}: Pokud testovaný odpoví celý svazek správně, obdrží 100 %. Pokud odpoví správně na více než polovinu dílčích dichotomických otázek, obdrží 50 %. V ostatních případech nedostane nic.
6. Korekce na tipování – odhaduje se, jakého skóre za svazek mohl zkoušený dosáhnout náhodným tipováním, a výsledek se koriguje.
7. Další složitější metody.
   Nejvíce se používají metody vše, nebo nic a PS_{50}, ostatní se opouštějí.

Úlohy s jedinou správnou odpovědí

(v anglické literatuře *single best answer*, **SBA**).

Nejpoužívanějším a současně nejefektivnějším typem úloh s výběrem z možností jsou úlohy s jedinou správnou odpovědí. Vzhledem mohou připomínat svazky dichotomických úloh, avšak jejich konstrukce je odlišná. I v tomto případě má úloha kmen následovaný nabídkou několika možností. Úkolem je vybrat odpověď, která je výrazně lepší než všechny ostatní. Zkoušený tedy nevyhodnocuje každou možnost zvlášť a nesnaží se určit, zda platí, či nikoli, jako ve svazku dichotomických úloh, ale porovnává nabízené možnosti mezi sebou. Žádná z nabízených možností přitom nemusí být správná zcela bezvýhradně a za všech okolností a žádná nemusí být zcela chybná. Na druhou stranu nabízené možnosti musí být možné seřadit od nejlepší po nejhorší.

> Obě úlohy v příkladu se ptají na totéž a nabízejí stejná řešení. V obou případech autor považuje za správnou odpověď možnost 2. Všimněte si ale, že položka typu ANO/NE není zcela jednoznačná: lze namítat, že Země nemá *přesně* tvar elipsoidu, a na druhou stranu lze pro některé účely její tvar dostatečně přesně aproximovat koulí.
> V případě položky typu SBA je situace jiná: zkoušený má vybrat *nejpřesnější* (nikoli nutně zcela přesnou) odpověď. Řešení je jednoznačné.

Srovnání výběrové položky typu ANO/NE a typu SBA

Položka typu ANO/NE		**Položka s jedinou nejlepší odpovědí**
Země se tvarem blíží		Země se svým tvarem blíží rotačnímu tělesu. Kterému z uvedených se podobá nejvíce?
1. kouli 2. elipsoidu 3. ovoidu 4. válci	ANO – NE ANO – NE ANO – NE ANO – NE	1. Kouli. 2. Elipsoidu. 3. Ovoidu. 4. Válci.

V obecné rovině se dá říci, že **pro vysokoškolské vzdělávání jsou položky typu SBA vhodnější** než svazky dichotomických úloh. To, že o jednotlivé možnosti v SBA nelze s absolutní platností říci, zda je zcela správná, nebo naopak zcela chybná, odpovídá reálnému životu. Zkoušení pomocí SBA lépe připravuje studenty pro praxi. Naproti tomu bývá obtížné vytvořit větší počet MTF k určitému tématu tak, aby úlohy byly opravdu jednoznačné. Snaha o jednoznačnost vede mnohdy k zpřesňování zadání, které je pak stále delší a detailnější, ale často také návodnější, takže výsledkem může být sice jednoznačná, ale současně velmi snadná úloha MTF. Lze tedy říci, že k určitému tématu lze vytvořit více kvalitních úloh SBA než MTF. Častá obava, že úloha s jedinou správnou odpovědí bude snazší a uhádnutelnější než svazek dichotomických úloh, který může mít správných odpovědí více, není odůvodněná. V praxi se naopak ukazuje, že správně zkonstruované úlohy typu SBA mají tendenci být obtížnější a obvykle lépe rozlišují než úlohy MTF.

Podrobnější informace o tvorbě úloh typu SBA nalezne čtenář v kapitole Doporučení pro tvorbu testových úloh.

Poznámka: Multiple-choice questions (MCQ)
Často se setkáváme s termínem *úloha s mnohočetným výběrem*, anglicky *multiple-choice question*, **MCQ**. Jde o obecnější pojem, který zahrnuje svazky dichotomických úloh (MTF), úlohy s jedinou správnou odpovědí (SBA) a další typy úloh. V této publikaci se pojmu MCQ záměrně vyhýbáme, neboť jeho význam není jednoznačný. V běžné komunikaci se totiž obsah termínu MCQ často zužuje jen na nejběžnější typ úloh a podle zvyklostí v konkrétní zeměpisné oblasti se jím pak myslí pokaždé něco jiného:

- V anglicky psané literatuře jsou MCQ nejčastěji synonymem pro položky *s jedinou správnou odpovědí*, tj. **SBA**.
- V naší jazykové oblasti se MCQ používá nejčastěji jako označení pro *svazky dichotomických úloh*, tj. **MTF**.

Vzhledem k zásadním odlišnostem v konstrukci i vlastnostech SBA a MTF může termín MCQ způsobovat nepříjemná nedorozumění.

Přiřazovací úlohy

Přiřazovací úloha je tvořena souborem premis a odpovědí. Úkolem zkoušeného je ke každé premise přiřadit nejlepší odpověď. Přiřazovací úlohy mohou mít různé poměry mezi počtem premis a odpovědí a podle toho se někdy rozlišují nejrůznější podtypy. V nejjednodušším případě je premis a odpovědí stejný počet a je dáno, že každá odpověď náleží k právě jedné premise. Jinou možností je, že je premis více (a některé odpovědi se použijí vícekrát).

Příklad:
Zařaďte každé zvíře do skupiny podle typu jeho potravy

1. Prase domácí _____	A. Masožravci
2. Lev pustinný _____	B. Všežravci
3. Zebra stepní _____	C. Býložravci
4. Kůň Převalského _____	
5. Krokodýl nilský _____	

Může také být naopak více nabízených odpovědí než premis. Extrémním případem jsou tzv. **rozšířené přiřazovací úlohy** (*extended matching questions*, **EMQ**). V mnohém připomínají několik úloh SBA za sebou, ale nabídka možností je podstatně rozsáhlejší (typicky více než deset) a stejný soubor odpovědí se používá pro více premis. Úlohy typu EMQ se rozšířily v medicínské oblasti, kde byly využívány především pro zkoušení klinických oborů.

Příklad:
Ke každé kazuistice bolesti v zádech zvolte nejpravděpodobnější diagnózu z této nabídky:

A. Ankylozující spondylitida
B. Disekce aorty
C. Vyhřeznutí meziobratlové ploténky
D. Lumbální spondylóza
E. Zlomenina obratle
F. Infekce meziobratlového disku
G. Defekt pars interarticularis
H. Metastáza do obratlového těla
I. Renální kolika
J. Herpes zoster

Úloha 1:
23letý muž má půlroční anamnézu bolesti v dolní části zad. Bolest zasahuje převážně thorakolumbální spojení a pravou hýždi. Bolest bývá nejhorší ráno, dělá mu obtíže vstát z postele. V průběhu dne dochází k částečnému zlepšení. Při vyšetření nacházíme omezenou pohyblivost lumbální páteře, především laterální flexe.

Úloha 2
32letá žena přichází pro náhle vzniklou bolest v dolní části zad. Bolest je setrvalá, nezávisí na poloze. Všechny spinální pohyby jsou omezené a bolestivé. Před třemi týdny prodělala infekci močových cest, která byla přeléčena amoxicilinem.

Přiřazovací úlohy mohou mít nejrůznější grafickou podobu. Testovaný může například ke každé premise vypisovat písmenné nebo číselné označení zvolené odpovědi, nebo může premisy a odpovědi spojovat čarou. Při testování na počítači se často odpovědi volí z rozbalovacího seznamu, nebo se odpovědi přetahují myší k premisám. V širším pohledu mezi přiřazovací úlohy patří třeba také umisťování popisků do obrázku.

Přiřazovací úlohy se hojně využívají například při výuce jazyků. Jejich vlastnosti jsou do značné míry podobné úlohám s jedinou správnou odpovědí, v podstatě jde o jakýsi svazek úloh SBA. V mnoha oborech se od přiřazovacích úloh postupně upouští, jsou nahrazovány právě úlohami typu SBA. Menší počet typů úloh, které jsou použité v určitém testu, bývá výhodou, neboť testovaný nemusí tolik přemýšlet, jaká forma odpovědi se od něj očekává, a může se lépe soustředit na samotné odpovídání na otázky. Tím se také test stává „přátelštějším“, snižuje se testová úzkost.

Pro bodování přiřazovacích úloh se používají podobné postupy jako pro bodování svazků dichotomických úloh (MTF), nejčastěji metody Vše, nebo nic, Dílčí skóre nebo PS_{50}.

Uspořádací úlohy

Zkoušený má za úkol seřadit nabídnuté položky (např. pojmy, děje) podle určitého pravidla. Může jít např. o seřazení kroků určitého postupu nebo uspořádání nějakých objektů podle nějaké veličiny či vlastnosti.

Příklad:
Seřaďte kapaliny od nejvyššího po nejnižší bod tuhnutí.
Voda
Nafta
Líh
Glycerín

Z formálního hlediska mohou uspořádací úlohy připomínat úlohy přiřazovací, neboť zkoušený ke každé položce přiřazuje její pořadí. V některých případech mohou uspořádací úlohy mít více správných řešení, např. roční období po sobě následují v pořadí jaro – léto – podzim – zima, ale také podzim – zima – jaro – léto apod.

Slabinou uspořádacích úloh je obtížné hodnocení. Někdy se používá metoda Vše, nebo nic, toto hodnocení ale mívá malou citlivost. Nejčastěji se proto hodnotí postupně po dvojicích a zkoumá se, zda položky ve dvojici jsou uspořádané správně, či nikoli:

Příklad:
Na stole leží čtyři stejně velké krychle, každá odlitá z jednoho kovu – železa, hliníku, mědi a zlata. Seřaďte krychle od nejlehčí po nejtěžší.

Správné pořadí: hliník – železo – měď – zlato
Odpověď zkoušeného: hliník – měď – železo – zlato

hliník – měď: správné pořadí
měď – železo: chybné pořadí
železo – zlato: správné pořadí

Zkoušený obdrží 2/3 bodů za úlohu.

3.2 Otevřené úlohy

Úloha s krátkou tvořenou odpovědí

též *úlohy se stručnou odpovědí*, *short-answer question, SAQ*. Na úlohu s krátkou tvořenou odpovědí má zkoušený nejčastěji odpovědět jedním slovem nebo slovním spojením. Často je odpovědí také výsledek nějakého výpočtu, načrtnutý graf nebo obrázek, chemický vzorec, matematická rovnice apod. Podle konstrukce se někdy tyto úlohy dělí na úlohy produkční a doplňovací:

- Produkční úloha:
  Jak se jmenuje hlavní město Velké Británie?
- Doplňovací úloha:
  Hlavní město Velké Británie se jmenuje _____ .

Úlohy s krátkou tvořenou odpovědí jsou vynikající součástí formativních testů. Jejich přínos spočívá v tom, že učiteli poskytují informaci nejen o tom, které části učiva studenti zvládli a do jaké míry, ale současně se pomocí nich dozví o případných omylech, neporozuměních a chybných konceptech, z nichž studenti vycházejí. Díky tomu je možné cíleněji reagovat při přípravě další výuky a studenty lépe vést.

Úlohy s krátkou tvořenou odpovědí jsou užitečné i pro sumativní testy. V některých oblastech jsou zcela obvyklé – např. při výuce jazyků, matematiky a geometrie apod. Jejich příprava pro sumativní test však musí být velmi pečlivá, aby se předešlo nejednoznačnostem.

Příklad:
Správně konstruovaná SAQ:
Jak se jmenuje hlavní město Velké Británie?
Správná odpověď: Londýn, London.
I v tomto jednoduchém případě se musí hodnotitelé domluvit, jak budou klasifikovat např. odpověď Londres.[1]

[1] Londres je jméno Londýna ve francouzštině, španělštině, portugalštině, katalánštině, baskičtině, galicijštině a dalších jazycích. V celosvětovém měřítku tento název používá víc než miliarda lidí. Je to název používaný větším počtem mluvčích než „London“ a „Londýn“ dohromady. Pokud bude ve výuce student-cizinec, takové odpovědi se objeví. Jde-li o test ze zeměpisu, je třeba takovou odpověď považovat za správnou.

Špatně zkonstruovaná SAQ:
Hlavní město Velké Británie je _____ .
V tomto případě není jasné, na co se autor úlohy ptá: hlavní město Velké Británie je Londýn, velké, v Anglii, historické, na Temži...

Obecně se dá říci, že na většinu úloh SAQ je více správných odpovědí a většinou je musí vyhodnocovat kvalifikovaný hodnotitel – expert v příslušném oboru. Hodnotitelé se buď musí shodnout, za jaké odpovědi přidělí plný počet bodů, za jaké částečné skóre a jaké odpovědi budou považovat za chybné, nebo musí dostat podrobný návod.

Co je jednotkou hmotnosti?
Mezi odpověďmi může být např. kilogram, gram, tuna, metrický cent, libra, unce, karát, kvertlík...

Jaké zvíře je ve znaku Moravy?
Červenostříbrně šachovaná orlice se zlatou korunkou
Jak hodnotit vynechání barvy, popřípadě popis barevné kombinace jako červenobílé?
Jak hodnotit odpověď „jednohlavý orel"? Je chybou vynechání informace o korunce?

Kvůli množství možných odpovědí se dokonce může stát, že vůbec nelze spolehlivě posoudit, zda je odpověď správná.

Napište jméno alespoň jednoho malíře.
Kolik bodů přidělíte za jméno, které vůbec neznáte? Lze spolehlivě ověřit, zda je odpověď správná? Kolik času a úsilí takové ověřování bude stát?

Mimo sumativní testování lze této vlastnosti SAQ využít. Rozličné odpovědi na stejnou otázku mohou povzbudit diskusi během výuky a aktivizovat studenty. Mohou nám také pomoci při tvorbě uzavřených úloh typu SBA: potřebujete-li pro takovou úlohu najít distraktory, zeptejte se na totéž pomocí SAQ. Získané odpovědi mohou být cennou inspirací.

Doplňovací úloha

V **doplňovací úloze** (cloze) dostává zkoušený text s vynechanými místy, která má doplnit. Nejčastěji má na několika různých místech doplnit vynechaná slova. V některých případech má doplnit některou část slova.

Příklad:
Doplňte vynechaný text
Rovinné těleso, které má tři strany a tři vrcholy, se označuje jako t_____helník. Zvláštní kategorii tvoří prav_____lé t_____níky. Jejich nejdelší strana, tzv. p_____a, je proti pravému _____, který je sevřen od_____ami.

Doplňovací úlohy se hodně používají při výuce jazyků, např. pro testování slovní zásoby nebo pro zkoušení schopnosti porozumět mluvenému slovu.

Příklad:
Na základě poslechu doplňte chybějící slova a údaje
Jediný přirozený satelit Země se jmenuje _____. Jeho průměrná vzdálenost od středu Země je _____ km. Zhruba jednou za _____ oběhne kolem Země. Člověk na něj poprvé vstoupil v roce _____.

V této podobě jsou doplňovací úlohy vlastně svazkem úloh s krátkou tvořenou odpovědí. Někdy se jako doplňovací úlohy (cloze) označují i podobně konstruované položky, v nichž zkoušený své odpovědi vybírá z předem daného seznamu slov, nebo je počet možných odpovědí omezen (např. úlohy typu *Doplňte -mě-/-mně-*). V tom případě jde vlastně jen o jinak označené přiřazovací, tedy uzavřené úlohy.

Při bodování doplňovacích úloh se nejčastěji přiděluje dílčí skóre za každé správné doplnění, nebo se používá parciální skóre PS_{50}.

Modifikovaný esej

Modifikovaný esej (*modified-essay question, MEQ*) je jiný typ svazku úloh s krátkou tvořenou odpovědí. Po úvodním medailónku následuje první otázka, pak se střídají doplňující informace a další otázky.

Příklad:
78letý muž, vdovec, který žije sám, přišel na ambulanci pro únavu a pokles tělesné hmotnosti. Byl přijat na všeobecné interní oddělení, na němž pracujete, k dalšímu vyšetření.

Otázka 1: Jaké jsou tři nejpravděpodobnější diagnózy?

Otázka 2: Napište pět otázek, které pacientovi položíte a které vám nejlépe pomohou mezi těmito třemi diagnózami rozlišit.
Laboratorní vyšetření ukázala mírnou chudokrevnost s koncentrací hemoglobinu 104 g/l. Objem červených krvinek je menší, než by odpovídalo referenčnímu rozmezí. Uzavíráte proto, že pacient trpí mikrocytární anémií.

Otázka 3: Napište dva typické klinické příznaky, po kterých budete při vyšetření pacienta pátrat.

Otázka 4: Stručně napište, jak uvedený výsledek změní vaši prvotní diagnózu.

Modifikovaný esej je na pomezí úloh s krátkou tvořenou odpovědí a široce otevřených úloh. Mají velkou hodnotu zejména ve formativních testech, v nichž do určité míry mohou simulovat dialog nad řešením problému mezi studentem a učitelem. Příprava tohoto formátu pro sumativní testy je však náročná. Ke všem požadavkům a omezením zmiňovaným u krátkých tvořených úloh přibývá ještě skutečnost, že chyba na začátku řešení modifikovaného eseje může mít vliv i na následující dílčí otázky. K tomu by v ústně vedeném dialogu nedošlo, neboť učitel by prvotní chybu vhodným způsobem korigoval. Samozřejmě má také vliv, jestli se student může při vyplňování testu vracet zpět, nebo nikoli.

Všimněte si, že v modifikovaném eseji bývají jednotlivé otázky pojaty mnohem šířeji než v typických úlohách s krátkou tvořenou odpovědí. Tentokrát již neočekáváme, že testovaný odpoví slovem nebo slovním spojením. Jeho odpověď bude spíše zahrnovat více samostatných výroků. Student musí nejen ovládat testovanou látku, ale také musí být schopen v krátkém čase stručně a přesně svou odpověď formulovat.

Esej

Esej je úloha s **dlouhou tvořenou odpovědí**. Zkoušený píše rozsáhlejší text (od jednoho odstavce do několikastránkové práce).

Esej obvykle tvoří samostatnou část zkoušky, jeho hodnocení se nekombinuje s jinými úlohami. Hodnocení může být zatíženo subjektivitou hodnotitele. Pokud se esej používá k sumativnímu hodnocení, hodnotí jej proto většinou více hodnotitelů a v ideálním případě dostávají eseje anonymizované. Aby bylo hodnocení objektivnější, hodnotí se většinou podle předem dané osnovy, tj. boduje se, nakolik zkoušený v eseji naplnil určité hodnocené aspekty („rubrics“). Kromě znalostí, tj. obsahové stránky, mají typicky velký vliv na hodnocení i schopnost podat dobře uspořádaný výklad, analyzovat a popsat souvislosti, přehledně, strukturovaně a srozumitelně se vyjadřovat a správně k tomu používat odbornou terminologii, dodržovat konvence obvyklé v oboru atd.

Využívání eseje jako nástroje pro hodnocení je různé v různých částech světa. Do určité míry se dá říci, že alternativou k eseji je **ústní zkouška**, při níž vede zkoušející se zkoušeným rozhovor. Výhodou ústní zkoušky může být dialog, díky němuž lze přesněji identifikovat slabé i silné stránky zkoušeného. Nevýhodou je ovšem nereprodukovatelnost hodnocení. Zatímco esej lze kdykoli dát znovu oznámkovat jinému hodnotiteli, ústní zkoušku nelze zopakovat. Dokonce i v případě, že by z ní byl pořízen audiovizuální záznam, může být opakované hodnocení svízelné, zejména pokud zkoušející v nějakém okamžiku vedl rozhovor nevhodným nebo chybným způsobem. Ke zvýšení objektivity může přispět zkoušení před komisí, avšak v praxi větší počet zkoušejících automaticky neznamená, že by je bylo možné považovat na vzájemně nezávislé hodnotitele, kteří se při známkování neovlivňují. Přes všechny tyto výhrady ale ústní zkouška má svou pedagogickou hodnotu, Právě díky možnosti založit zkušební akt na odborném dialogu a interakci. Podmínkou je ovšem vysoká erudice a profesionalita zkoušejícího.

3.3 Další typy testových úloh

Test shody se scénářem

Úloha, též označovaná jako *script concordance test* (SCT), začíná medailónkem podobným jako v SBA. Následuje otázka, která jednak nabízí možné řešení (hypotézu), jednak přináší novou informaci a ptá se, do jaké míry nová informace nabídnutou hypotézu podporuje. Testovaný vybírá odpověď obvykle z pěti možností (od hypotéza je velmi nepravděpodobná po hypotéza je velmi pravděpodobná).

Příklad:
Vyšetřujete 14měsíční holštýnskou jalovici, která je nadmutá a má anorexii. Její tělesná teplota je 39,5 °C, tepová frekvence 115/min., dechová frekvence 64/min. Není

dehydratovaná, kontrakce bachoru jsou neslyšné, trusu je velmi málo.
Pokud jako možnou příčinu zvažujete dislokaci slezu doleva a v laboratorním nálezu je fibrinogen 10 g/l, stává se tato příčina

−2 velmi nepravděpodobnou
−1 méně pravděpodobnou
0 ani méně ani více pravděpodobnou
1 pravděpodobnější
2 velmi pravděpodobnou

Za vinětou často následuje několik otázek:

Vyšetřujete 48letého muže s Fourniérovou gangrénou, který opakovaně podstoupil chirurgické odstranění nekrotické tkáně. Nemocný je léčený širokospektrými antibiotiky.

Pokud jste měl v plánu...	**... a zjistil jste, že**	**je plánované řešení**
1. transplantaci kůže na skrotální defekt	v části defektu je granulační tkáň, ale onemocnění dále postupuje do třísel, kde jsou nové nekrotické oblasti,	-2 : absolutně kontraindikované -1 : relativně kontraindikované 0 : stejně indikované nebo kontraindikované 1 : indikované 2 : velmi indikované

Pokud jste měl v plánu...	**... a zjistil jste, že**	**je plánované řešení**
2. další debridement	pacient je septický, zaintubovaný, kardiopulmonálně nestabilní,	-2 : absolutně kontraindikované -1 : relativně kontraindikované 0 : stejně indikované nebo kontraindikované 1 : indikované 2 : velmi indikované

Pokud jste měl v plánu...	**... a zjistil jste, že**	**je plánované řešení**
3. hyperbarickou oxygenoterapii	pacient je septický, zaintubovaný, kardiopulmonálně nestabilní,	-2 : absolutně kontraindikované -1 : relativně kontraindikované 0 : stejně indikované nebo kontraindikované 1 : indikované 2 : velmi indikované

Bodování odpovědí se stanovuje zvlášť pro každou otázku na základě názoru skupiny expertů. Každý expert označí jednu možnost, kterou považuje za správnou. Možnost, kterou označil největší počet expertů (tzv. *modální možnost*), se boduje plným počtem bodů. Skóre za ostatní možnosti se přiděluje podle vztahu

[Skóre za možnost] = [počet expertů, kteří označili tuto možnost] / [počet expertů, kteří označili modální možnost]

Například:

Úlohu posuzovalo 10 expertů. Správné možnosti označili takto:

Možnost	-2	-1	0	1	2
Počet expertů, kteří možnost označili jako správnou	0	0	2	5	3

Modální možností je odpověď „1“, kterou označil největší počet expertů – za tuto možnost tedy bude náležet plný počet bodů. Celé bodování bude vypadat takto:

Možnost	-2	-1	0	1	2
Skóre za možnost	0/5 = 0	0/5 = 0	2/5 = 0,4	5/5 = 1	3/5 = 0,6

3.4 Doporučení pro tvorbu testových úloh

Psaní položek je úkol, který vyžaduje představivost a kreativitu, ale zároveň vyžaduje značnou disciplínu při práci a znalost výukových cílů. Tvorba položek musí vycházet z jasné představy o cílech učení. Test by měl měřit jednu kognitivní oblast.

Než začnete úlohy tvořit

Dříve, než dojde na vlastní tvorbu úloh, je třeba se vrátit k otázce, co vlastně chceme zkoušet a proč. Nestačí jen vzít učebnici, prolistovat kapitoly, které obsahově pokrývají připravovaný test, a vytvořit úlohy k textu, na nějž padne zrak.

V ideálním případě máme k dispozici předem zpracovaný **plán testu** (specifikační tabulku). Pokud tomu tak není, měl by plán testu vzniknout, než začneme psát úlohy. Není-li ani to z jakéhokoli důvodu možné, měli bychom přinejmenším mít co nejpodrobněji sepsané **cíle výuky** (tedy nikoli jen tematické okruhy, kterých se výuka týká). K cílům výuky je vhodné doplnit představu, jak velká část testu by se jimi měla zabývat.

Plán testu nebo podrobný seznam cílů výuky nám dává jasné vodítko, jaké úlohy pro test potřebujeme a v jakém množství. Současně bychom již v této fázi měli mít základní představu o typech úloh, které v testu použijeme.

Volba typů úloh

Při sestavování sumativního testu neuděláme chybu, pokud většina úloh budou úlohy s jedinou nejlepší odpovědí (*single-best answer*) a případně je doplníme otevřenými otázkami s krátkou tvořenou odpovědí. Další typy úloh bychom měli používat uvážlivě, případně podle zvyklostí konkrétního oboru. Úlohy s jedinou nejlepší odpovědí poskytují nejlepší poměr cena/výkon. Čas potřebný na jejich zodpovězení je relativně krátký, takže takových úloh může být do testu zařazen dostatečný počet, a současně umožňují dostatečně citlivě rozpoznat schopnosti studentů. Úlohy s krátkou tvořenou odpovědí se hůře hodnotí, na druhou stranu jsou ale vhodným doplněním, neboť poskytují učiteli lepší zpětnou vazbu.

Ve formativních testech, zvláště, pokud je skupina testovaných malá, může být poměr obrácený – většinu testu mohou tvořit úlohy s krátkou tvořenou odpovědí, které jsou doplněné výběrovými úlohami s jedinou nejlepší odpovědí. Ve formativních testech není třeba se bát ani dalších formátů úloh, je-li jejich použití účelné. Test by ale nikdy neměl kombinovat příliš mnoho různých formátů (ne více než tři nebo čtyři), jinak bude nepřehledný a studenti stráví hodně času zkoumáním, co se po nich vlastně chce a jak mají na kterou úlohu odpovídat.

Pokud se v testu kombinuje více typů úloh, je třeba studentům jednoznačně říci, co se od nich čeká. Pokyn musí být zcela konkrétní.

Příklad:
Pokyn k úlohám s jedinou nejlepší odpovědí
Vhodný: Zakroužkujte nejlepší odpověď.
Nejasný: Vyberte nejlepší odpověď.

Pokyn k úlohám s krátkou tvořenou odpovědí
Vhodný: Odpovězte jedním slovem nebo slovním spojením.

3.4.1 Doporučení pro tvorbu výběrových úloh

Doporučení pro tvorbu uzavřených (výběrových) úloh uvedeme na tomto místě pro úlohy s jedinou nejlepší odpovědí (*sigle best answer*, SBA), tedy typ úloh, který by měl být základem většiny testů. Z velké části lze stejná doporučení použít i pro ostatní formáty výběrových úloh.

Častým problémem při použití výběrových úloh je nejednoznačnost. Většinou ji způsobí to, že autor úlohy při jejím sestavování má na mysli nějakou konkrétní situaci, ale pak se snaží zadání napsat co nejstručněji. Tím se z jejího textu ztratí podrobnosti a předpoklady, které jsou ale pro zodpovězení důležité. Student, který řeší test, pak nejprve musí odhadnout, co vlastně měl autor testu na mysli, a pak teprve může odpovídat. Úloha pak zákonitě měří spíše schopnost studenta odhadnout, na co se učitel chtěl zeptat, než samotné znalosti a dovednosti ve zkoušené oblasti.

Základem kvalitní výběrové úlohy je proto dobře napsaný kmen. Úlohy s jedinou správnou odpovědí mívají kmen poměrně dlouhý, na několik řádek (někdy se označuje jako *medailonek*). Kmen by měl **vyprávět** příběh – popsat jednoduchou, ale reálnou situaci, nebo třeba pokus. Jasným popisem situace, kterou měl autor na mysli, se předejde většině nejednoznačností. Pro studenty jsou navíc takto sestavené úlohy motivující – příběhy, které připomínají reálnou praxi, jim připomínají, že se učí něco praktického, co budou potřebovat ve svém budoucím zaměstnání.

Pro začínající autory testových úloh je někdy psaní medailonků obtížné. Základní doporučení říká, že mají úlohy psát stejně, jako by navrhovali výzkum k zodpovězení konkrétních (ale samozřejmě jednoduchých) otázek ze svého oboru.

Za medailonkem následuje vlastní **otázka**. Ta má být krátká, jednoznačná, musí se ptát jen na jednu věc. V úlohách s jedinou nejlepší odpovědí musí být z otázky jasné, že se po studentovi chce, aby vyznačil opravdu jen jednu možnost, která je lepší odpovědí než všechny ostatní. To je důležité hlavně v případě, kdy by i další z nabízených možností dávaly aspoň

za určitých podmínek smysl. Správně konstruované otázky mívají podobu např. **Jaká je nejpravděpodobnější příčina? Jaký je nejvhodnější další postup? K čemu nejspíše povede popsaný děj?**

Z výše uvedeného vyplývá další jednoduchá zásada: Kvalitní položky mají obvykle **relativně dlouhý kmen (medailonek), za kterým následuje stručná, ale jasná otázka. Následuje nabídka možností, které by měly být rovněž krátké.**

Co se týče nabízených možností, bývá jednoduché vytvořit správnou odpověď (tzv. *klíč*). Platí tady zásada, že odborník by měl správnou odpověď říci po přečtení kmene a otázky, i bez nabídky možností. Výběr z několika možností vlastně jen usnadňuje vyhodnocení testu a poskytuje určitou pomoc studentovi, který v dané oblasti ještě není tak zběhlý jako odborník.

Obtížněji se učitelům navrhují nesprávné odpovědi – **distraktory**. Autoři úloh si nemusí již zcela vybavovat, v čem spočíval problém při osvojování zkoušeného tématu, a navrhují pak distraktory, které jsou pro studenty irelevantní. Může proto být výhodné úlohu nejprve dát studentům jako otevřenou otázku, nejlépe ve formativním testu. Distraktory se pak vytvoří z chybných odpovědí. Učitel současně získá představu, jak bude vytvářená úloha obtížná. Pokud takovýto postup zkombinuje s diskusí nad odpověďmi, popřípadě vyzve studenty, aby při řešení „přemýšleli nahlas“, poskytne mu to další cenné informace jak pro výuku, tak i pro další úpravy testových položek.

Zběhlejším autorům testových úloh často pomůže pětice doporučení, jak by měla položka s jedinou nejlepší odpovědí vypadat. Těchto pět doporučení pomůže také recenzentům při oponentuře nových úloh.

1. Zaměřte se na významný problém.

Vysokoškolské vzdělávání má studenty připravovat pro reálnou praxi, a proto i testové úlohy se mají ptát na problémy relevantní pro praxi. Neztrácejte čas triviálními ani příliš složitými otázkami. Nemá smysl testovat málo podstatné, marginální znalosti – takové úlohy obvykle nic nevypovídají o připravenosti pro praxi, spíš zkouší, jak který student „umí psát testy“. Nepoužívejte „chytáky“, vyhýbejte se negativně formulovaným úlohám. Zkoušejte znalosti a porozumění, nikoliv pozornost.

2. Zkoušejte využití znalostí, nikoliv vybavení pojmu nebo izolovaného faktu.

Delší medailonek úlohy vyžaduje, aby student nějakým způsobem vyhodnotil popsanou situaci a interpretoval ji. V sumativních testech se vyhýbejte zkoušení definic a klasifikací – zkoušejte, jestli student rozumí obsahu pojmů a dokáže s nimi zacházet, popřípadě jestli má zařazení pojmu do určité kategorie spojeno také s porozuměním, jaké vlastnosti daná kategorie má. Zda se toto doporučení podařilo naplnit, lze mnohdy jednoduše ověřit: zkopírujte celý kmen úlohy do internetového vyhledávače – neměla by se před vámi objevit správná odpověď.

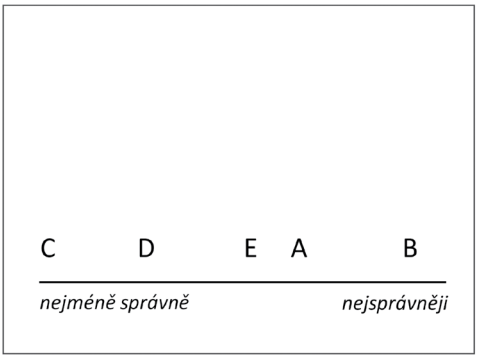
3. Na úlohu musí být možné odpovědět i se zakrytými možnostmi.

Nechejte otázku zkontrolovat svými kolegy. Nejprve jim ji dejte bez nabízených možností – měli by na ni dokázat správně odpovědět. Pokud ne, úlohu přepište.

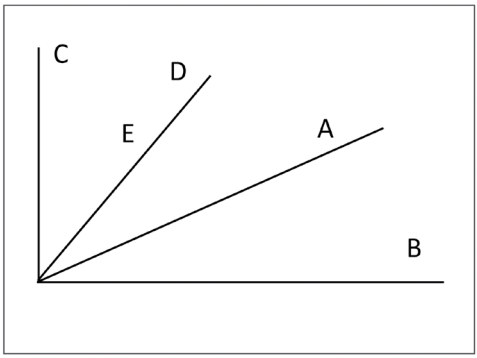
4. Nepoužívejte relativizující ani absolutní termíny.
Text kmene ani nabízené možnosti by neměly obsahovat relativizující termíny jako *často, zřídka, výjimečně, většinou* apod. Použití takových slov způsobí, že úloha nebude jednoznačná, v relativizujícím termínu se skrývá určitý pohled autora, který student nemusí správně odhadnout. Určitá situace například může být vzácná z pohledu běžné populace, současně je ale relativně častá z pohledu odborníka, který se řešením takové situace profesionálně zabývá.

Mezi „zakázaná slova“, především v nabízených možnostech, patří také absolutní termíny *vždy, nikdy* apod. Jen málokdy něco platí opravdu na 100 %, takže použití takovéhoto slova v některé z nabízených možností obvykle znamená, že jde o distraktor – a studenti to snadno rozpoznají.

5. Všechny nabídnuté odpovědi musejí být homogenní.
Všechny nabídnuté možnosti musejí vypadat podobně – měly by být psané podobným stylem a měly by mít podobnou délku. Musejí spadat do jedné kategorie – například všechny nabízené možnosti jsou možné příčiny nějakého děje, slovní druhy, biologické taxony, pracovní úkony, chemické látky apod. Odpovědi musí být možné seřadit od nejlepší po nejhorší.



Správně: Nabídnuté možnosti jsou homogenní



Špatně: Nabídnuté možnosti jsou neporovnatelné

Obr.3.4.1 Na příkladu vlevo lze možnosti nabídnuté k otázce seřadit od nejhorší (možnost C) po nejlepší (možnost D). Svědčí to o homogenitě nabídnutých možností. Pokud stejnou otázku položíte několika expertům, všichni nabídnuté odpovědi seřadí ve stejném pořadí. Naproti tomu v případě vpravo není možné nabídnuté možnosti seřadit – každá možnost spadá do jiné kategorie, odpovídá na položenou otázku z jiného aspektu, takže bychom při pokusu o seřazení „srovnávali jablka s hruškami“.

Příklad otázky s nehomogenními možnostmi
Vyberte nejlepší tvrzení o Marfanově syndromu:

A. Postihuje častěji muže.
B. Jde o poruchu kolagenního vaziva.
C. Léčí se kyselinou hyaluronovou.
D. Bývá spojen s oligofrenií.
E. Projevuje se nápadně krátkými končetinami.

Nabízené možnosti mají být seřazeny náhodně nebo podle abecedy. Pokud možnosti obsahují číselný údaj, měly by být seřazené podle něj.

3.4.2 Doporučení pro tvorbu otevřených úloh

Při vytváření otázek s krátkou odpovědí pro písemné testy dbejte těchto doporučení (Jolly 2010):

- **Otázky formulujte jasně a jednoduše, vystříhejte se jazykových záludností a chytáků.**

Dobrá otázka s krátkou odpovědí testuje znalost konkrétních fakt nebo schopnost analyzovat a interpretovat nějaký scénář. Není vhodné ve stejné úloze současně testovat schopnost porozumět složitě konstruované otázce – výsledky hodnocení by pak byly prakticky neinterpretovatelné.

- **Pokuste se na otázku odpovědět z různých úhlů pohledu.**

Otázka, která se ptá na jednu konkrétní skutečnost, by měla mít jedinou správnou odpověď. Naopak otázka, která se ptá na možné varianty (např. na diferenciální diagnózu), bude mít více správných řešení. Počítejte s tím, že i otázku, která vám připadá jednoznačná, mohou různí čtenáři pochopit různě. Vždy je vhodné, aby otázky zkontroloval recenzent.

- **Napište, jak dlouhou odpověď očekáváte, např. *Odpovězte jedním slovem nebo slovním spojením* nebo *Načrtněte graf funkce....* Napište také, jak bude úloha hodnocena.**
- **Negativně formulované otázky používejte opatrně.**

Kladně formulované otázky („Jaký je nejlepší postup...“, „Jaká je nejpravděpodobnější příčina...“) mají větší didaktickou hodnotu, než negativní otázky („Jaký je nesprávný postup“). Pokud už používáte negativně formulovanou otázku, zdůrazněte zápor např. použitím verzálek („Které antibiotikum v této situaci NENÍ vhodné?“).

- **Dbejte, aby odpovědi nenapověděla např. velikost místa pro její vepsání.**
- **Pečlivě připravte pokyny pro hodnotitele a nechejte je zkontrolovat svým kolegům společně s úlohou.**

3.5 Techniky tipování odpovědí

Jako **testová moudrost** (*test-wise*, *testwiseness*) se označuje schopnost, která umožňuje studentovi správně vybrat odpověď na otázku, aniž by znal oblast, na kterou se úloha ptá (American Psychological Association 2020, The University of Kansas nedatováno). Využívá toho, že autoři výběrových úloh mají obvykle nejprve dobře rozmyšlenou správnou odpověď a k ní až dodatečně vymýšlejí distraktory. Často tak dělají stereotypně, takže lze jejich úvahy odhadnout, nebo se dopouštějí některých typických nedopatření (Berk 2002).

Nejdelší odpověď je správná

Testová moudrost velí studentům preferovat nejdelší odpověď. Někdy je také správná možnost na první pohled kompletnější než ostatní možnosti, nebo je specifičtější či detailnější (Al-Faris et al. 2010). Je to tím, že autoři úloh ve snaze o jednoznačnost napíší správnou možnost velmi detailně, distraktorům pak věnují menší péči. Může se také stát, že pro tvorbu

distraktorů použijí neúplnou část textu správné odpovědi. Typickým příkladem mohly být dřívější položky např. v autoškole, nebo v testech bezpečnosti práce, kde nejdelší z nabídnutých odpovědí bývala zpravidla ta správná.

Odpověď „uprostřed“

Pokud lze odpovědi logicky seřadit (např. jde o čísla) a student neví, která je správná, tipne si některou z prostředních. Když autoři úloh vymýšlejí distraktory, často zvolí nějaké menší a nějaké větší hodnoty. Studenti, kteří si jsou vědomí tohoto pravidla, proto tipují některou z odpovědí uvnitř nabízených možností. Tím, že vyloučí nejnižší a nejvyšší hodnotu, se pravděpodobnost správného tipu podstatně zvýší. Nabízené odpovědi také někdy „krouží kolem“ správného řešení, takže stačí vysledovat, v čem spočívá vzájemná podobnost, a správnou odpověď lze odhadnout.

Příklad:
Jaká je plocha papíru formátu A4 ve srovnání s formátem A6:
a) poloviční,
b) trojnásobná,
c) čtyřnásobná,
d) osminásobná.

Student, který nezná odpověď, vyloučí extrémy, tj. možnosti a) a d). Volí tedy mezi b) a c). Odpověď c) je podobná možnostem a) i d) (jde o mocniny čísla 2), zatímco možnost b) se liší. Zkusí tedy vybrat c) – což je správná odpověď.

Gramatická návodnost

Při tvorbě kmene položky a distraktorů je třeba si dát pozor, aby gramatická forma nenapovídala správnou odpověď. Při tvorbě kmene úlohy autor obvykle již má vymyšlenou správnou odpověď a formulace kmene jí bude odpovídat. To nemusí platit o distraktorech, které mnohdy vymýšlí dodatečně, nebo je na poslední chvíli mění.

Příklad:
Zdroj, který umožňuje nejrychleji zkontrolovat, zda je v oboru něco nového, je:
a) knihy,
b) odborné časopisy,
c) internet,
d) vědecké konference.

Absolutní a relativizující výrazy

Při konstrukci odpovědí se autor pokouší vyloučit nedorozumění tak, že použije upřesňující výrazy, které ale poskytují vodítko pro tipování. Pokud některá možnost obsahuje některý zpřesňující („extrémní“) termín, např. vždy, nikdy, pouze, nezbytně, musí, všechny, žádný, nemožné, stále, jde o nesprávnou odpověď (distraktor). Naopak relativizující výrazy často, zřídka, možná, někdy, obvykle, většinou, může apod. bývají ve správných odpovědích.

Příklad:
Které tvrzení o savcích je správné?
a) Jde výhradně o suchozemské živočichy.

b) Žádný savec neumí létat.
c) Mohou mít ploutve.
d) Všichni mají vyvinutý zrak.

Logický klíč – protiklady nebo vyčerpávající výčet možností

Pokud jsou mezi nabízenými odpověďmi dvě nebo tři, které pokryjí všechny možnosti připadající v úvahu, jistě bude jedna z nich správná.

Příklad:
Uvažujme matematické kyvadlo s hmotností závaží m a délkou závěsu l. Pokud zvětšíme hmotnost m,
a) doba kyvu se zkrátí,
b) doba kyvu se nezmění,
c) doba kyvu se prodlouží,
d) maximální výchylka se zmenší,
e) maximální úhlová rychlost se sníží.

Možnosti a), b) a c) společně pokrývají všechny možné případy (doba kyvu se zkrátí, nezmění nebo prodlouží). Student, který využívá testové moudrosti, bude uvažovat jenom o těchto možnostech a vůbec nebude ztrácet čas možnostmi d) a e).

Příliš jednoduchá odpověď

Studenti předpokládají, že položky budou obsahovat chytáky a složitosti. Mají proto tendenci nezvolit odpověď, která je jednoduchá a samozřejmá. Občas je vhodné takovou správnou odpověď zařadit, aby se toto schéma porušilo.

Vše výše uvedené

Pokud je mezi nabízenými odpověďmi „všechny uvedené odpovědi“, nebo „nic z uvedeného“ (a jim podobné), pak tuto odpověď studenti upřednostní. Navíc se ukazuje, že formulace tohoto typu v odpovědích nediskriminují dobře mezi lepšími a horšími studenty. Neměly by se proto v testech používat.

Opakující se je správně

Učitel často připravuje distraktory tak, aby se mu zdály co možná podobné správné odpovědi. Může se tak stát, že lze odpověď uhádnout porovnáváním nabízených možností a hledáním, v čem se shodují.

Příklad:
Vyjádřete jednotku ohm v základních jednotkách SI:
a) $m \cdot kg \cdot s^{-3} \cdot A^{-2}$,
b) $m^2 \cdot kg \cdot s^{-2} \cdot A^{-2}$,
c) $m^3 \cdot kg \cdot s^{-3} \cdot A^{-2}$,
d) $m^2 \cdot kg \cdot s^{-3} \cdot A^{-2}$.

Vyjádření jednotky ohm v základních jednotkách je nesporně náročná úloha. Vidíme však, že se odpovědi liší prakticky jen v exponentech. A exponenty se v odpovědích opakují.

Vybereme tedy tu variantu odpovědi, v níž se objevují všechny opakující se znaky. V našem případě je to odpověď d).

Nápověda mezi položkami

V testech se mohou vyskytnout položky, které jsou nápovědou pro jinou položku. Zvláště u delších testů je velmi obtížné udržet tento aspekt pod kontrolou. Pro prevenci této situace se používají dvě strategie. Položky v položkové bance mají nastaveny vzájemné relace a systém nedovolí vybrat do stejného testu „příbuzné" položky. Druhou strategií je nabízet studentům otázky postupně a neumožňovat návrat k otázkám již dříve zodpovězeným.

Verbální podobnost

Pokud jsou mezi kmenem položky a jednou z nabízených odpovědí verbální podobnosti, bývá tato odpověď správná.

3.6 Automatizace tvorby testových úloh

S tím, jak přibývá počítačem podporovaného testování, a zvláště pak s rozvojem adaptivního testování, přitahují pozornost metody, kterými by se tvorba testových úloh mohla zjednodušit. V tradičním přístupu ke konstrukci testů vytvářejí jednu každou položku specialisté na konkrétní oblast. Nejprve úlohu napíše autor, potom ji další odborníci oponují, následně ji pedagog prověří v pilotním testu a podle výsledku ji revidují a upravují. Teprve poté se položka konečně použije pro testování. Celý proces je dlouhý a nákladný. V důsledku toho je stále obtížnější pokrýt rostoucí poptávku po zkušebních položkách (Drasgow et al. 2006). **Automatické generování položek** (*automatic item generation*, **AIG**) by mohlo představovat velkou úsporu času i prostředků, a je proto předmětem intenzivního výzkumu. Některé koncepty řešení tohoto úkolu se dostaly již do stádia praktického testování.

V prvním z konceptů můžeme proces automatického klonování položek rozdělit do dvou kroků. Autoři testových úloh nejprve vytvářejí modely položek, které slouží jako jakési šablony. Snaží se z úloh vydestilovat jejich podstatu, která je zásadní pro prokázání znalostí. Na vhodná místa v těchto šablonách jsou pak navrhovány různé alternující termíny (často strojově, např. pomocí slovníků synonym). S využitím sady zástupných termínů pak algoritmus změní tuto šablonu na skupinu souvisejících položek vytvořením všech možných permutací. Tím dochází ke generování „nových", nikoliv však nezávislých položek. V konkrétním běhu testu nemůže být použita víc než jedna položka z každé skupiny klonů. Navíc některé permutace povedou ke vzniku nesmyslných, nebo nepravděpodobných kombinací, takže musí být vyloučeny (Gierl a Haladyna 2012; Gierl a Lai 2012; Gierl et al. 2012).

Je pak předmětem diskuse, zda v položkových bankách mají být až výsledné klony, nebo zdrojové šablony a variabilní součásti položek. Účelem je získat položky, jejichž psychometrické charakteristiky by byly odhadnutelné ze známých výsledků jiné položky ze stejné série klonů. Díky náročnému procesu tvorby, který vede k nutnosti ujasnit si podstatu každé položky, jsou takto vzniklé úlohy často překvapivě kvalitní. Poznamenejme, že použitelnost strojově generovaných variant je závislá i na konkrétním jazyce. Například v češtině s její komplikovanou gramatikou by to bylo mimořádně obtížné.

Podobný postup byl testován i při snahách vytvářet položkově srovnatelné testy pro ověřování reliability metodou test-retest. Ukázalo se, že pokus modifikovat původně funkční položky změnou alternujících termínů vedl k vytváření položek s vyšší obtížností (Fiřtová 2021). To poněkud nabourává původní představu, že klonované položky budou mít stejné psychometrické parametry jako originál. Je tedy otázkou, zda celý proces dává smysl, když sice vzniknou nové položky, ale tak jako tak je třeba je kalibrovat.

Druhý koncept automatizované tvorby testových úloh otevírají první práce zabývající se využitím umělé inteligence. Modelový postup byl předveden na workshopu při jednání Evropské rady lékařských hodnotitelů v Braze v Portugalsku. Skupina autorů dostala za úkol vytvořit k danému tématu (bolest břicha) kognitivní mapu. Kognitivní mapa pomáhá popsat problém po prvcích (např. věk, pohlaví, kontext, vitální funkce, příčina, diagnóza). Každý z těchto prvků může mít sadu různých hodnot. Zkušení vývojáři testů potřebují na vytvoření kognitivní mapy několik hodin. Poté počítač vygeneroval sadu položek, které představují různé kombinace prvků kognitivní mapy. Při workshopu bylo toto namíchání prvků provedeno s pomocí aplikace Excel. Autorům testů by nasazení podobného systému mohlo v budoucnosti usnadnit život (Vleuten 2019b). Problém tohoto přístupu spočívá v časové náročnosti a nákladnosti vytváření kognitivní mapy. V publikaci věnované automatizované tvorbě položek z matematiky pro první stupeň se ukazuje, že automatizovaná tvorba je nákladově výhodná (oproti tradiční tvorbě), pokud lze z jednoho kognitivního modelu vygenerovat sadu více než 200 položek (Kosh et al. 2018).

V témže roce byl prezentován i další systém, který pomocí umělé inteligence dokáže dolovat data z oborové bibliografické databáze a využít je pro tvorbu kmenů položek i návrhy distraktorů. Tyto návrhy položek mohou sloužit lidským autorům jako polotovar pro snazší tvorbu nových úloh (Davier 2019).

3.7 Recenze testových úloh

V procesu přípravy testů, zvláště u zkoušek s velkým významem, má nezastupitelnou roli kontrola položek pomocí recenze expertů před jejich použitím v testu (tzv. *panel review*). Zatímco u kvízu, kterým zjišťuje učitel znalosti žáků z přírodovědy ve čtvrté třídě, nemusí být nutné, aby obsah testu posoudili další učitelé, u testů, které jsou součástí přijímací zkoušky nebo odborné certifikační zkoušky, to již potřeba je. Položky projdou několika úrovněmi nezávislé kontroly, než je uvidí první účastník testu.

Oponentura, nebo též **recenze položek** je rozdělena do několika fází, které se vždy zaměřují na specifickou oblast. Jejím cílem je odhalení nedostatků, které zpravidla položky a testy ve své počáteční podobě obsahují. Motivací je zajištění správnosti, optimalizace testu a odstranění subjektivních vlivů. I když recenze bývá zpočátku časově a organizačně poněkud náročnější, její přínos je nepopiratelný a roste s významem testu. Po úspěšném zvládnutí všech níže uvedených revizí (obsahová revize, revize férovosti, redakční revize) by měl finální podobu jednotlivých úloh znovu projít autorský tým a všechny provedené změny odsouhlasit.

Proč je kontrola položek a celého testu potřeba?
Testové položky jsou součástí nástroje, jímž měříme nějakou schopnost testovaných. Kontrola správnosti, formulační přesnosti a nerozpornosti položek dělá test lepším měřícím nástrojem a snižuje pravděpodobnost, že test bude neférový a že si někdo z účastníků na něj nebo na jeho jednotlivou položku bude stěžovat.

Kdo má položky kontrolovat?
To se může značně lišit podle významu testu. U testování menšího významu bohatě postačí jeden další recenzent. Požádáte kolegu, aby vám test prošel a zkontroloval. U zkoušek velkého významu, jako jsou přijímací zkoušky, maturitní testy a podobně, musí být položka zkontrolována několika recenzenty s jasně přidělenými rolemi. Recenzující experti musí být současně experty na danou oblast a současně by měli znát testovanou populaci.

Co kontrolující kontrolují?
Záleží na typu recenzenta a jeho roli v procesu recenze. Testující instituce často vytvářejí kontrolní seznamy, podle kterých recenzenti postupují. Recenzent může kontrolovat, zda je kmen položky dobře formulovaný. Zda není gramaticky návodný a neusnadňuje tak výběr správné odpovědi. Zda je klíč správný a distraktory nesprávné a zda jsou všechny možnosti srovnatelně dlouhé. Korektor může zkontrolovat správnost interpunkce, správné použití horních a dolních indexů, dodržení zvyklostí při zápisu proměnných a jednotek.

Jak se recenzní práce organizuje?
I když může být recenzní formulář (kontrolní seznam) i v papírové podobě, je běžnější, že má podobu elektronickou. Často je přímo integrován v položkové bance, takže položky ani při recenzi neopouštějí bezpečné prostředí banky. Administrátor testu může v položkové bance kontrolovat stav recenzí a motivovat recenzenty k vyšším výkonům.

Pro oponenturu je, podobně jako přípravu kompletní testové agendy, základem týmová spolupráce. Několik zainteresovaných odborníků nezávisle na sobě posuzuje vhodnost jednotlivých úloh a společnými silami se snaží o odstranění všech nedostatků, které by mohly při praktické realizaci vadit. Týmová spolupráce hraje při oponování testů a testových položek zcela klíčovou roli.

Proces oponentury položek a testu lze rozdělit do tří fází, kterými oponenta provede **formulář pro recenzenty úloh** (podrobněji rozebrán dále v textu).

3.7.1 Obsahová revize

Jsou odpovědi správně a přesně formulované? Nejsou distraktory diskutabilní?
V rámci revize obsahu je velmi vhodné, aby zadání otázek a nabízené odpovědi zkontrolovali jak spoluautoři celého testu, tak i nezávislí odborníci, kteří nebyli zapojeni do jejich vytváření. Subjektivní postoj autora může být příčinou nejednoznačné, tedy nesprávně utvořené testové položky, jejíž použití by snížilo hodnotu testu.

Zvláště obtížnou činností při vytváření položek pro většinu pedagogů bývá formulace alternativních odpovědí (distraktorů). Obecně by distraktory neměly být nesmyslnými tvrzeními nebo absurdními možnostmi, které testovaný automaticky vyloučí, ale naopak by jej měly

donutit k zamyšlení a následné eliminaci po logickém zdůvodnění. Mimořádně náchylné k nejednoznačným formulacím distraktorů jsou dosud velmi rozšířené položky s mnohočetným výběrem odpovědí typu MTF.

U jiných typů otázek mohou vyvstat jiné typy obsahových nedostatků. Otázky s jedinou nejlepší odpovědí (SBA) musí být revidovány tak, aby existovala shoda expertů o jednoznačně nejlepší odpovědi.

Stojí-li pedagog před úkolem vytvořit více testových úloh a krom správných odpovědí i navrhnout řadu vhodných distraktorů, může si pomoci tím, že v rámci formativního testování zadá studentům své nové položky jako úlohy s krátkou tvořenou odpovědí. Studenti při tvorbě odpovědí často navrhnou skvěle fungující a atraktivní distraktory.

V obecné rovině se po obsahové stránce doporučuje kontrolovat zejména:

- přesnost formulace zadání/kmene položky,
- zda jsou nabízené možnosti v každé úloze formulovány tak, aby za žádných okolností, v žádné interpretaci ani v žádném uvažovaném případě nemohl být distraktor správnou odpovědí a obráceně (platí zejména pro MTF),
- zda položky v testu odpovídají plánu testu (*blueprint*).

3.7.2 Redakční revize

Jsou otázky dostatečně srozumitelné, typograficky jednotné a bez typografických chyb?

Redakční revize se může na první pohled jevit jako nepříliš časově náročná, nicméně v praxi to může být složitější. Je nutné projít všechny testové položky a ověřit, zda jsou dostatečně čitelné, srozumitelné a formálně i typograficky jednotné. Složitá větná souvětí, dvojité zápory a krkolomná zadání úloh je vhodné přepracovat do jednodušší formy tak, aby student nemohl ve formulaci zabloudit. Zadání úlohy i nabízené možnosti by měly být konstruovány co možná nejsrozumitelněji. Jednotnost a styl vytváření testových položek se liší podle autorů. V této fázi oponentury se provádí sjednocení jak po stránce terminologické, tak po stránce typografické. Nedílnou součástí kontroly jakýchkoli textů je gramatická správnost. To platí i pro vytváření testovacích položek. Eliminace veškerých gramaticky nesprávných či sporných výrazů dle pravidel pravopisu by měla být závěrečnou fází redakční revize.

V praxi se ukazuje, že jedna recenze je zcela nedostatečná. Ideálního stavu, kdy je recenzí 5–7, je s omezenými finančními prostředky těžké dosáhnout, ale jako použitelné minimum se jeví 3 recenze. Často přitom na problém upozorní jen jeden z recenzentů. Proto musí být zpracovatel recenzí velmi pozorný k návrhům recenzentů, aby nepřehlédl možný problém.

Příklad:
Při redakční revizi můžeme odhalit i gramaticky nebo graficky návodné formulace otázek (tzv. sugestivní zadání):
Místem narození Jana Amose Komenského byl:
Uherský Brod

Nivnice
Komňa
Brno

3.7.3 Formulář pro recenzenty úloh

Z praktického hlediska je výhodné vybavit recenzenty formulářem, který je bude oponenturou testových položek „provádět“. Tím, že recenzent odpovídá na jednotlivé otázky ve formuláři, musí se testovou položkou zabývat ze všech úhlů pohledu, které formulář postihuje. Není nezbytně nutné, aby každá testová položka zcela vyhověla ve všech sledovaných parametrech; oponent by však případné odchylky měl zaregistrovat a komentovat. Příklad takového formuláře pro recenzenty úloh najdete v tab. 3.7.1.

Tab. 3.7.1 Recenze otázky s jedinou nejlepší odpovědí

Zadání otázky		
Recenzent		
	Ano ✓ nebo Ne ×	**Poznámky**
Zkouší podstatnou znalost.		
Odpovídá tématu dle plánu testu.		
Zkouší aplikaci znalostí, nikoli jen vybavení izolovaných údajů.		
Odpovídá požadované úrovni znalostí.		
Zadání je jasně formulované.		
Zadání je bez chytáků (např. dvojí zápor).		
Správná odpověď odborníka napadne, i když nezná nabízené možnosti.		
Distraktory jsou homogenní.		
Formulace možností nenapovídá správnou odpověď.		
Žádná možnost není nepřiměřeně obtížná.		
Nemá podobu „které tvrzení je správné“ nebo „všechna tvrzení jsou správná kromě“.		
Neobsahuje slova „vždy“, „obvykle“, „zřídka“, „nikdy“ apod.		
Právě jedna z nabídnutých možností je nejlepší.		
Nabídnuté možnosti jsou seřazené abecedně či v jiném logickém pořadí.		
Možnosti mají podobnou délku a obsah.		
Možnosti jsou kompatibilní s otázkou.		

3.7.4 Revize férovosti

Měří úlohy pouze požadovanou konkrétní znalost či dovednost a nic jiného?
Každá položka, každý test by měly testovat právě požadovanou vědomost, znalost či schopnost a nic jiného. Podle definice je férovost testu míra, do jaké jsou závěry učiněné na základě výsledků testů validní pro různé skupiny účastníků testů.

Pokud jsou k zodpovězení otázky nutné znalosti a dovednosti, které z jakéhokoli důvodu nebyly srovnatelně dostupné všem testovaným osobám, tedy pokud neměli všichni testovaní stejnou možnost požadované znalosti či dovednosti získat, není položka férová. Taková otázka je snazší pro skupinu studentů, kteří byli nějakým způsobem zvýhodněni, a naopak obtížnější pro druhou skupinu, která byla bez vlastního zavinění znevýhodněna. Příkladem může být nadbytečné používání odborných výrazů nebo složitých větných konstrukcí, které nemusí být pro všechny srozumitelné. Ačkoli chtěl autor otázky ověřit určitou znalost, současně v tomto případě nechtěně testuje jazykovou vybavenost a zběhlost v odborné terminologii. V této souvislosti může být další komplikací také testování pozornosti studentů prostřednictvím „chytáků v zadání“, případně používání dvojitých záporů a podobně.

Položka by neměla zvýhodňovat žádnou skupinu podle věku, pohlaví, původu, společenského a ekonomického postavení, víry, rasy, mateřského jazyka atd. Vzhledem k tomu, že členění na skupiny není nijak omezené, není reálné zkoumat férovost pro všechny možné skupiny v populaci účastníků testování. Doporučuje se proto zkoumat spravedlivost vůči těm skupinám, u nichž zkušenosti nebo výzkumy ukázaly, že by mohly být nepříznivě ovlivněny. Často se jedná o skupiny, které byly diskriminovány na základě takových faktorů, jako je etnický původ, zdravotní postižení, pohlaví nebo rodný jazyk. Studenti z různých skupin *se shodnou úrovní znalosti* by měli na danou otázku odpovídat správně se stejnou pravděpodobností.

Základní doporučení a pravidla tvorby testových položek a testů týkající se férovosti položek jsou obsaženy například ve standardech ETS pro férovost a kvalitu testů (ETS Standards for Quality and Fairness) (Educational Testing Service 2014). Tyto standardy doporučují ověřit, že testové položky:

- nejsou urážlivé ani kontroverzní,
- neposilují stereotypní pohledy na žádné skupiny,
- jsou bez rasových, etnických, genderových, socioekonomických a jiných forem zaujatosti,
- nemají obsah, který by byl považován za nevhodný nebo hanlivý vůči jakékoli skupině.

Neférovost položek lze často odhalit důkladnou revizí férovosti samotného zadání. Někdy ji však neodhalí ani zkušený oponent. Proto při analýze výsledků testu zkoumáme i diferenciální chování položek, jak ukážeme v kapitole věnované položkové analýze.

4 PROVEDENÍ TESTU

4.1 Pilotování testů

Důvěryhodné testování výsledků výuky, zvláště pokud ovlivňuje další postup studentů, předpokládá, že vlastnosti používaného testu budeme znát ještě před jeho ostrým použitím. K odhadu vlastností testu slouží pilotní testování a pretestování.[2]

Pretestování používá pro vyhodnocení testu srovnatelné postupy, které se provádějí při vyvozování závěrů z „ostrého" testování. Zatímco pro samotný pilotní běh testu stačí menší skupina studentů, například 20 (Alderson et al. 1995), s odpovídající úrovní znalostí a motivací, jako má cílová skupina, pro pretest, sloužící k výpočtu statistických parametrů položek, je třeba skupina větší, nejméně 100 respondentů.

Vzhledem k nárokům na sestavení relevantní skupiny a mnohdy i časové náročnosti se jako pretest často používá první „ostrý" běh samotného testování. Podněty získané z vyhodnocení předběžných testů je zapotřebí zapracovat v návrhu ostré verze testu. Zpravidla je třeba upravit některé položky. Pokud pretest prokáže významné nedostatky, může však jít i o přepracování celé koncepce testu (Komenda a Pokorná 2011).

4.1.1 Subjektivní zpětná vazba

Subjektivní zpětná vazba poskytuje velmi důležitou informaci od vybraného vzorku z cílové skupiny respondentů – typicky od vybraných studentů. Ti nám mohou svými subjektivními názory pomoci identifikovat nejasnosti či chyby v zadání. Názory každého člena zvolené skupiny je nutné brát v úvahu a zvážit jeho poznámky a podněty. Skladba pilotní skupiny by měla být vyvážená, nemělo by se tedy například jednat o žáky s nadprůměrnými výsledky, nebo naopak o vyloženě slabé žáky. Prostředků pro samotnou realizaci je více. Vzhledem

[2] Terminologická poznámka: Oba pojmy se částečně překrývají. Termín **pilotní testování** se v této práci většinou používá jako širší označení obou kroků. Pokud je třeba oba kroky rozlišit, rozumí se pojmem *pilotní testování* obecnější „*proof of concept*" – jakási studie proveditelnosti, která na malé skupině studentů odhaluje případné chyby v konceptu a designu testu a může přinést i užitečnou subjektivní zpětnou vazbu. Termínem **pretest** se pak myslí formálnější a podrobnější předběžné prověření testu, které umožňuje odhadnout psychometrické vlastnosti otázek, jejich obtížnost, schopnost testu rozlišit mezi dobrými a slabšími účastníky testu a které umožňuje získat *subjektivní* i *objektivní zpětnou vazbu* od testované skupiny.

k efektivitě dalšího zpracování je nejrozšířenější dotazníková forma v elektronické podobě, kde je možné odpovědi snadno zpracovat a předat v přehledném formátu pracovní skupině. Níže je uveden výčet vhodných možností, jak lze subjektivní zpětnou vazbu provádět:

- dotazník,
- diskusní fórum,
- diskuse ve frontální výuce (v případě menšího množství studentů, při větším počtu se tato varianta stává neefektivní),
- poznámky v testu nebo tzv. přemýšlení nahlas, „*think aloud*“ (Tavakol a Dennick 2011a), kdy jsou studenti žádáni, aby při řešení testu komentovali nebo zaznamenávali své myšlenkové pochody.

4.1.2 Objektivní zpětná vazba

Objektivní zpětná vazba je důležitá pro svou nepopiratelnost, která se opírá o matematické zpracování výsledků pilotního testu. Závěry objektivní zpětné vazby dávají indicie k případné modifikaci nevyhovujících testových položek. Mezi nejznámější a nejhojněji užívané výstupy hodnocení testů patří:

- zhodnocení **obtížnosti** testových úloh (identifikace snadných a obtížných úloh, nevyhovujících položek, možnost uspořádání úloh podle obtížnosti),
- určení **citlivosti** jednotlivých úloh (analýza a korekce nebo vyřazení úloh s nevyhovující citlivostí),
- vyhodnocení kvality testu jako celku, především jeho **reliability** a **validity**.

Při vyhodnocování výsledků testu pilotní skupiny musíme mít na paměti možné odlišnosti pilotní skupiny od cílové, způsobené např. odlišnou motivací obou skupin. Tyto odlišnosti je dobré předem minimalizovat, např. vhodnou „legendou“ provázející pilotní test.

4.2 Cvičné testy

Pokud mají studenti příležitost absolvovat cvičné testy z látky, kterou se učí, má to často příznivý vliv na výsledky výuky – jde o tzv. **efekt testování**. U důležitých testů se proto často organizuje „test na nečisto“ (anglicky *mock test*). Studenti díky tomu získávají možnost ujistit se, že nebude problém s technickou stránku věci (u počítačových či distančních testů), vyzkoušet si testový formát (položky s jednou nebo více správnými odpověďmi apod.), ověřit si předem znalosti a časovou náročnost. Tyto zkoušky na nečisto významně snižují *testovou úzkost*, zvyšují motivaci a „efektem testování“ i připravenost studentů. Z pohledu organizace tyto testy umožňují kalibrovat nové položky, vyzkoušet si organizaci zkoušek před ostrým během a zejména podpořit přípravu studentů poskytnutím zpětné vazby o jejich aktuálním výkonu.

Zkouška na nečisto obvykle obsahuje tytéž součásti jako ostrý test. Všechny části testu se vyhodnocují a účastníkům se nabízí zpětná vazba poukazující na jejich silné a slabé stránky. Studentům to umožňuje poučit se ze svých chyb a získat praxi a sebejistotu před finálním testem.

Pro studenta mají cvičné testy řadu přínosů:

- Nastavení správné strategie. Pokud je test časově stresující, může si to student při vypracovávání cvičného testu uvědomit a přizpůsobit tomu svou strategii práce s časem (např. časově náročné úlohy odloží na konec).
- Získání praxe. Zkoušky velkého významu mohou účastníky stresovat a snižovat tak jejich přirozený výkon. To je možné omezit přípravou v podmínkách podobných skutečné zkoušce.
- Analýza vlastního výkonu. Po každém testu by studenti měli věnovat čas analýze svých chyb. Měli by pečlivě projít každou část testu, aby zjistili, kde dělají nejčastěji chyby, a mohli svou přípravu na tato místa zaměřit. Pomocí tohoto druhu přípravy mohou studenti lépe porozumět otázkám, které by mohly být použity v závěrečném testu.

Metaanalýza provedená v roce 2017, ale i další práce ukázaly, že cvičné testy a jejich zpětná vazba mají velký vliv na výsledky učení a lze je použít jako účinný nástroj pro podporu učení. Ukazuje se, že studenti, kteří se účastnili cvičných testů, často dosahují lepších výsledků než studenti, kteří se připravovali jiným způsobem, např. opakováním učiva, procvičováním apod. Podle některých studií jsou cvičné testy pro učení prospěšnější než opakování učiva a všechny ostatní metody, které se srovnávaly. Cvičné testy lze proto doporučit k efektivní podpoře učení a jako součást zpětné vazby pro studenty i učitele (Adesope et al. 2017; Yang et al. 2019).

4.3 Administrace testu

Administrace testu – prezenční, nebo distanční?
Díky rozvoji počítačové techniky je možné volit, jakým způsobem bude test administrován. Může to být písemně (*paper based test*, PBT), nebo online (*computer based test*, CBT). Každý z těchto přístupů má své výhody a svá omezení.

4.4 Papírové testování

Papírové testy tvořené položkami s výběrem odpovědí se dočkaly masivního rozšíření již za 1. světové války v reakci na personální potřeby americké armády. Bylo tehdy nutné rychle a efektivně klasifikovat velký počet rekrutů, a to se nedalo řešit do té doby obvyklou individuální prací psychologů (Dubois 1970).

Díky své efektivitě se testování pomocí předtištěných testů rychle rozšířilo do dalších oborů, které dříve spoléhaly na individuálně administrované testy – vzdělávání, testování inteligence a dalších.

V anglicky psané odborné literatuře se rozlišují dva pojmy: čistě *počítačové testování* (*computer based testing*) a *počítačem podporované testování*. Ve druhém případě může sběr odpovědí probíhat i pomocí papírových dotazníků (jde tedy o *paper based testing*), pomocí výpočetní techniky se pak ale testy vyhodnocují a dále analyzují.

Počítačové testování je jistě meta, k níž celý obor směřuje. Přesto má *papírové testování* svůj význam nejen při nedostatku počítačového vybavení, ale i jako snadný vstup do světa testování a používání souvisejících metodik. Vhodně zvolené programy a technologie nám mohou výrazně ulehčit práci.

V nejjednodušší podobě papírového testu stačí volně vytištěné úlohy s nabídnutými odpověďmi. Tradiční vyhodnocování odpovědních formulářů pomocí průsvitek se šablonou správných odpovědí vykazovalo díky lidskému faktoru velkou chybovost, často srovnatelnou s počtem chyb, které v odpovědích udělal respondent. S nástupem optických skenerů a technologie optického rozpoznávání značek (*optical mark recognition*, OMR) již čtení odpovědních formulářů není problém. Poměrně snadno lze analyzovat i opravy a změny odpovědí, které zkoušený provedl. Pro automatizované vyhodnocení je ovšem třeba navrhnout formuláře tak, aby byly snadno strojově čitelné, tedy vyhovovaly požadavkům na optické rozpoznávání značek. Příklady strojově zpracovatelných dotazníkových listů lze vyhledat na internetu pod termíny „bubble answer sheet“, „OMR answer sheet“, případně „scantron test sheets“.

Testy lze generovat a tisknout přímo i z programů podporujících testování, jako je specializovaný testový program Rogō. Ten podporuje tisk strojově čitelných formulářů, včetně vytváření několika verzí testu s různě seřazenými úlohami. Tisk testových formulářů umožňuje i LMS Moodle, který má pro tvorbu strojově čitelných formulářů rozšíření *Quiz OMR*.

Zatímco tisk testových formulářů je v testovacích programech často zahrnut, čtení a rozpoznávání vyplněných formulářů není ve jmenovaných testovacích systémech řešeno. Je nutné využít externí řešení, např. osvědčený komerční software Remark Office.

Výhody

- Papírové testování většinou využívá předtištěné formuláře, v nichž testovaný vyznačí své odpovědi. Výhodou je, že lze souběžně administrovat velký počet testů.
- Odpovídání na papíru je pro některé testované intuitivnější a komfortnější, nebudí v nich obavy, jestli zvládnou práci s technikou.

Nevýhody

- Nevýhodou je nepružnost celého procesu daná použitými technologiemi.
- Nelze získat některé informace, např. o rychlosti, kterou testovaný odpovídal.

4.5 Počítačové testování

Elektronické hodnocení se z velké části vyvinulo z konvenčních forem hodnocení. Původní papírové testy a odpovědní formuláře byly převedeny do digitální podoby a doručovány testovanému buď aplikací běžící na lokálním počítači, ale s rozvojem techniky daleko častěji on-line, prostřednictvím internetu. Masivní nárůst elektronického testování pozorujeme zvláště v posledních deseti letech (Egarter et al. 2020). K tomu se nově otevírá testování pomocí mobilních platforem (Denison et al. 2016). Pro realizaci on-line testování je k dispozici řada softwarových nástrojů. Na jednu stranu to jsou specializované programy, které řeší jen testování (např. Rogō) nebo jsou testovací moduly součástí různých komplexních nástrojů (např. LMS Moodle).

Výhody

Počítačového testování je oproti papírovému nesrovnatelně flexibilnější. Umožňuje v zadání úloh využívat multimédia, také obrázky při něm neztrácejí kvalitu. Počítačové testování také přináší řadu výhod pro administraci testu, řízení jeho průběhu (je např. možné nastavit jednosměrný průchod testem, hotové úlohy se mohou uzamknout apod.). Obrovskou výhodou pro bezpečnost testů je možnost sledovat, jak testovaný odpovídal v čase a jak dlouho mu která úloha trvala. Mimo to počítačové testování otevírá zcela nové možnosti adaptivního testování. Přímé vyplňování a zpracování testu na počítači podstatně urychluje jeho vyhodnocení, což studenti očekávající zpětnou vazbu velmi kvitují. Elektronické testování je obecně méně chybové a vede ke zvýšení kvality hodnocení (Denison et al. 2016). Díky počítačovému testování pronikají do hodnocení nové formáty testových úloh, využívající například možnost vyznačit odpověď v obraze. A v neposlední řadě je počítačové testování nákladově efektivnější a ekologicky šetrnější než jeho papírová obdoba (Dennick et al. 2009).

Nevýhody

Počítačové testování je ve srovnání s „papírovým" limitované výpočetní technikou, kterou má provozovatel k dispozici. Počítačový systém může být napaden virem, nebo hackery, případně selhat v důsledku ztráty napájení nebo konektivity. Před elektronickým testováním je třeba zaškolit testované i personál do používání elektronického testovacího systému. Případné řešení sporů s nespokojenými účastníky testu může být komplikovanější, protože není k dispozici „papírový důkaz", jaké bylo zadání a jak student odpověděl. Při testování mnoha klientů najednou mohou vznikat zvýšené požadavky na přenosovou kapacitu, zvláště v případě mobilních zařízení s bezdrátovým připojením. Od nasazení elektronických forem hodnocení mohou odrazovat vysoké vstupní náklady (HW + SW), ale tuto nevýhodu vyvažují pozdější nízké náklady provozní.

4.6 Distanční testování

S rozvojem distanční výuky se objevila potřeba převést i testování na novou distanční formu. Na rozdíl od ústního distančního zkoušení, které ze zkoušení prezenčního přebírá většinu metodologie a doplňuje ji jen o videokonferenční nástroj pro komunikaci, distanční testování se od prezenčního podstatně odlišuje. Do popředí se dostává důraz na věrohodnost a minimalizaci pokušení si v nepřítomnosti učitele vylepšit výsledky testu pomocí nepovolených pomůcek a zdrojů informací.

Pro uchování věrohodnosti hodnocení je třeba testování přizpůsobit distančním podmínkám. Jsou v principu dvě cesty, jak toho dosáhnout: proktorované testování a testování s „otevřenou knihou".

4.7 Proktorované testování

Jsou situace, kdy je potřeba vyzkoušet kandidáty, kteří nejsou všichni na stejném místě. Většinou se vyžaduje, aby ke zkoušce všichni přijeli do místa konání. To ale nemusí být vždy efektivní, nebo to dokonce nemusí být ani uskutečnitelné.

Klasickým řešením takové situace je, že se zkoušení shromáždí do několika center, kam za nimi přijede zkušební komise. Ve všech místech může zkouška probíhat současně, nebo i v různých časech. Vytvářejí se tedy paralelní testová sezení, případně se využívají i paralelní verze (varianty) testu.

Plně kvalifikovaný zkoušející může být zastoupený dohlížitelem, **proktorem**. Proktor není odborníkem ve zkoušené oblasti, nemůže tedy být hodnotitelem testu, nedokáže k testu poskytovat zpětnou vazbu a experta nenahradí ani v jiných rolích. Může však zajistit prostředí a podmínky pro regulérní provedení testu.

Proktorované zkoušení získalo oblibu nejprve při ověřování kompetencí armádních důstojníků, kdy bylo potřeba vyzkoušet lidi rozptýlené po celém světě, nebo v mezinárodních jazykových zkouškách. V typickém uspořádání se zkoušení dostaví do vybaveného zkušebního centra s proškoleným personálem, který ověří identitu zkoušeného, poskytne mu test a dohlíží, aby průběh zkoušky odpovídal vypsaným pravidlům.

Ani dohlížitel nemusí být na místě, kde se píše test, fyzicky přítomen. Může na testované dohlížet na dálku (**vzdálené proktorování**). Tato cesta umožňuje využít stávající typ testů a doplňuje je důsledným online dohledem. Popsaná metodika online dohledu nad distančním testováním se zhruba před deseti lety rozvinula do samostatné disciplíny – „online proktorování". **Proktorování** se snaží technickými prostředky eliminovat rizika nežádoucího chování účastníků testu. Pokrývá celý proces od identifikace zúčastněných, kontroly prostoru, v němž je testovaný, umístění kamer(y), kontroly spuštěných programů v počítači, až po odeslání a vyhodnocení výsledků.

On-line proktorované testování má dvě hlavní modality. Jednak testování ve velkém měřítku, kde se jako výhoda jeví úspory z rozsahu. Druhou modalitou jsou distanční zkoušky velkého významu, které mají vyšší bezpečnostní standardy a používají se například pro přijímací a výstupní zkoušky, pro certifikace apod. Oba typy se nabízejí i komerčně jako služba.

Proktorované testování velkého rozsahu se zpravidla organizuje pro stovky až tisíce účastníků. Pro doručování testu nejčastěji využívá mírně modifikované nástroje pro běžné elektronické testování (Moodle, BlackBoard, ...), případně rozšířené o moduly omezující např. otevírání dalších aplikací. Pro dohled se v takových velkokapacitních systémech často využívá umělá inteligence, která detekuje nestandardní chování účastníků. Při volbě software pro online testování je potřeba brát v úvahu, jak se chová při ztrátě spojení. Pokud se např. rozpadne spojení při testování v Moodle, jsou všechna data vyplněná testovaným ztracena. Pokud dojde k výpadku spojení při testování v Rogō, je ztracena pouze poslední (aktuálně zodpovídaná) položka.

4.7.1 Prevence a detekce podvádění při distančním testování

Prevence podvádění při proktorovaném testování se řídí stejnými pravidly jako prevence podvádění při testování, které probereme v samostatné kapitole o bezpečnosti testování.

Specifikou proktorovaného testování je nepřítomnost pedagoga na místě, což může pro některé studenty představovat výzvu k hledání neetických metod ovlivňování výsledku. Proto je při

distančním testování věnována velká pozornost technickým řešením, která nahrazují osobní dohled a omezují možnosti podvádění. Provedení dohledu může mít několik podob, lišících se mírou zabezpečení, počtem testovaných a cenou.

Živé proktorování v reálném čase

Ke kontrole slouží streamované video z webové kamery a mobilního telefonu v reálném čase. Tento druh vzdáleného dohledu vytváří obtížně řešitelný souběh požadavků na přenos obrazu. Na rozdíl od videokonferencí, kde ve vysokém rozlišení stačí přenos od přednášejícího, u dohledových systémů je požadavek na přenos obrazu ve vysokém rozlišení od všech testovaných osob. To by při sdílení plného videa vedlo k rychlému vyčerpání šířky přenosového pásma a tím k omezení maximálního počtu osob v jednom běhu testu. Používají se proto výkonné komprimace videa, obraz se přepíná mezi testovanými, nebo se přenášejí jen jednotlivé fotografie pořízené v náhodných časech.

Proktorování pomocí záznamu a dodatečného přezkoumání

Průběh testování se zaznamenává a záznam se posléze vyhodnocuje, či uchovává pro pozdější vyhodnocení. Výhodou je možnost testovat větší skupiny s menším množstvím dohlížitelů. Proktorování pomocí záznamu obrazu z webové kamery má prokazatelné příznivé efekty na zmírnění akademické nepoctivosti v online testech (Dendir a Maxwell 2020). Nicméně problémy se šířkou přenosového pásma jsou stejné, jako v případě živého proktorování.

Proktorování pomocí umělé inteligence

Dohled je dvojstupňový. V první úrovni sleduje zkoušené studenty umělá inteligence, která vyhodnocuje chování studenta v reálném čase. V případě incidentu upozorní živý dohled, který situaci řeší. Výhodou je opět možnost obsloužit velký počet testovaných s malým počtem dohlížejících. Pokud aplikace s umělou inteligencí běží na počítači testovaného, snižují se nároky na přenosové pásmo a otevírají se možnosti obsloužit skutečně velké počty testovaných.

Kromě dohledu se obvykle nasazují i technická řešení omezující možnosti nelegálního získávání informací z internetu v průběhu testování. Používají se k tomu speciálně vyvinuté internetové prohlížeče – např. Safe Exam Browser a další. Programy tohoto typu obvykle skvěle fungují na standardizovaných počítačích v učebnách, ale při jejich použití v domácím prostředí studentů mohou vznikat problémy s různými variantami operačních systémů a neslučitelnými kombinacemi jiných programů.

Zvláštní prohlížeč není jediným řešením zmíněného problému. Jako alternativa byl vyzkoušen javascriptový program PageFocus, který velmi citlivě a selektivně monitoruje pokus o otevření dalšího okna a upozorní testovaného na jeho nedovolené jednání (Diedenhofen et al. 2017).

4.7.2 Kontroverze proktorovaného zkoušení

Diskutovaným pedagogickým problémem proktorovaného zkoušení je, že *a priori* pohlíží na testovaného s nedůvěrou a svými technickými opatřeními předjímá nežádoucí soutěž – „kdo

s koho“. Nevýhod proktorovaného online zkoušení a námitek vůči němu je více (Nigam et al. 2021):

- Dochází k nežádoucí intruzi do soukromí studenta, například uchováváním obrazového záznamu z jeho domácnosti. Vznikají problémy s uchováváním osobních a biometrických údajů.
- Vytváří se prostředí vzájemné nedůvěry a podezřívání.
- Zvyšuje se testová úzkost.
- Mezi zkoušeného a zkoušejícího se staví řada technických prostředků, z nichž každý může selhat (ztráta spojení, zamrznutí systému...). Některé kroky testování jsou složitější (např. ověření identity, kontrola pracovního místa). Vznikají proto rozsáhlé pokyny a scénáře pro proktory i zkoušené. Proškolování všech zúčastněných může nežádoucím způsobem odvádět pozornost od samotné zkoušené oblasti, na niž by se měli studenti před testem soustředit.
- Online zkoušení je ovlivněno tím, do jaké míry je zkoušený seznámen s použitými technickými prostředky. Závisí na tom rychlost, kterou dokáže odesílat odpovědi, ale také jistota, se kterou odpovídá, a testová úzkost. Přechod na vzdálené zkoušení tak může vnést do hodnocení významné zkreslení.

V současné době není uzavřená diskuse, zda přínosy online proktorovaného zkoušení a proktorovaného zkoušení využívajícího umělou inteligenci vůbec mohou vyvážit jejich nevýhody a rizika (Nigam et al. 2021). Někdy však není vyhnutí a proktorované testování je třeba použít i s vědomím těchto problémů.

4.7.3 Alternativní přístupy

Velká část problémů spojených s proktorovaným zkoušením ve vysokoškolském vzdělávání souvisí se znalostním pojetím hodnocení. Vzdálené zkoušení ohrožuje ve větší míře, než je tomu u prezenčního zkoušení, jen několik hrozeb: záměna identity zkoušeného (tj. odpovídá někdo jiný, než by měl), nedovolená spolupráce (tj. zkoušenému někdo pomáhá či napovídá) a zejména použití nedovolených zdrojů a pomůcek.

Ačkoliv současné technické prostředky umožňují v ideálních podmínkách poměrně spolehlivě ověřovat identitu zkoušeného (např. i podle biometrických znaků), při časovém stresu, velkém počtu testovaných a špatné kvalitě spojení může být identita snadno podvržena. Zvláště pokud má podvádějící povědomí o použitých technických prostředcích a omezeních, jimž zkoušející čelí. V takovém případě je na místě zvážit nahrání a uložení identifikační procedury, což dává dodatečnou možnost řešit případné pochybnosti.

Velkým ohrožením validity výsledků je nedovolená spolupráce a používání nedovolených zdrojů a pomůcek během zkoušky. Pokud vyčerpáme technické a organizační prostředky pro jejich vyloučení, může tato rizika dále snížit vhodná konstrukce testu a testových úloh. Napovídání (vzájemná spolupráce) je snadné u uzavřených úloh. Efekt nedovolených pomůcek je podstatný zejména při zodpovídání znalostních otázek.

Vhodnou cestou se proto zdá konstruovat vzdálené zkoušky tak, aby ověřovaly především dosažení vyšších vzdělávacích cílů, nikoli jen samotné zapamatování a vybavení faktických znalostí. Konkrétní faktické údaje se totiž dají v průběhu testu nejsnáze najít pomocí

vyhledávače, v poznámkách nebo v literatuře, a v krátkém čase použít. Naproti tomu při řešení úloh zaměřených na hlubší porozumění, nebo dokonce určité dovednosti nepomůže hůře připravenému kandidátovi ani vyhledávání na internetu, ani opisování z knih či taháků – samotná izolovaná fakta pro zodpovězení úlohy nestačí. Obtížnější je také napovídání nebo spolupráce s další osobou.

Častým doporučením je, aby se vzdálené zkoušky co nejvíce koncipovaly jako „zkoušky s otevřenou knihou" (*open book*). V současné době sice ještě není dostatek studií, kterými by bylo možné toto doporučení jednoznačně podpořit, je však dobře podložené teoreticky.

V úvahu je třeba brát i vztahy mezi zkoušeným a zkoušejícími. Atmosféra důvěry a férovosti do značné míry potlačuje pokusy o podvodné jednání. Méně přísný dohled se proto jeví jako vhodné řešení hlavně tehdy, pokud vyučující/zkoušející pracuje se studenty dlouhodobě, jsou vytvořené kvalitní vztahy mezi učitelem a studenty i mezi studenty navzájem, a pokud se zkouší vyšší úrovně vzdělávacích cílů. Naproti tomu typickou situací, kdy se nelze obejít bez „tvrdého" přístupu a přísného dohledu, jsou vysoce kompetitivní zkoušky, u nichž se kandidáti a zkoušející navzájem neznají, např. přijímací nebo certifikační zkoušky.

4.8 Testování s otevřenou knihou

Testování s otevřenou knihou

Pozoruhodným přínosem pandemie COVID-19 byl prudký rozvoj **testování s otevřenou knihou**. Testy a zkoušky s otevřenou knihou umožňují pedagogům klást otázky, jejichž zodpovězení není možné jen na základě přístupu k informačním zdrojům. Vyžadují vyšší kognitivní dovednosti, vyhledání a zpracování informací a kritické myšlení místo memorování. V mnoha ohledech je zkouška s otevřenou knihou bližší běžné praxi. *Open-book* testování připravuje studenty na práci v digitálním světě (Sam et al. 2020).

Zdá se, že dohled a restrikce zajišťující při distančním testování rovné podmínky, nejsou jedinou možnou cestou ke spravedlivému on-line testování. Zvláště, když zesílená kontrola ze strany pedagogů vede k reakci ze strany části studentů, kteří pod tímto tlakem hledají stále sofistikovanější cesty, jak restrikce obejít. Tím se do ohniska vztahů mezi studentem a pedagogem dostává místo spolupráce nežádoucí soutěž „kdo s koho" v podvádění. Tato situace se zvláště vyhrocuje při testování v online prostředí, kde je při proktorovaném testování dozor velmi zřetelný a každá nedokonalost v zabezpečení testu může snadno vést k jeho znehodnocení.

Svoji roli hraje i společenský a technologický pokrok. Noví studenti jsou „digitální domorodci" a jsou s novými technologiemi zvyklí nativně pracovat. Nová komunikační a výpočetní zařízení jsou stále kompaktnější a sofistikovanější. Do budoucna bude téměř nemožné zabránit studentům v jejich používání při distančních zkouškách a bude stále obtížnější jim bránit i u zkoušek prezenčních. Použití online zdrojů během zkoušky se tak stane nekontrolovatelné a jejich zákaz bude prakticky nevymahatelný. Radikální restrikce, jako např. vypnout v den přijímacích zkoušek datové připojení a mobilní služby v celé zemi, se nezdají být v našich podmínkách správnou (ani žádoucí) volbou (Esemes.cz 2014).

Abychom zachovali rovnost podmínek a akademickou integritu, můžeme v nových podmínkách přesunout těžiště hodnocení od tradičních testů (se zavřenou knihou) k testům

s otevřeným přístupem k informacím. Netestovat již tolik samotné znalosti, ale přesouvat pozornost k testování dovedností. Klasické testy používáme ze setrvačnosti a často jen proto, že jsme dříve nebyli schopni dovednosti hodnotit – je načase to začít měnit.

Přechod na open-book testování znamená uzpůsobit celou testovou agendu nové situaci. Klasické testování umožňuje klást otázky zaměřené na vybavení jednotlivých informací. Pokud uvažujete o testování s otevřenou knihou, znamená to posun na vyšší úrovně Bloomovy taxonomie. Je třeba položit otázky, které vyžadují, aby studenti aplikovali své znalosti na nové situace a použili analytické a kritické myšlení. Aby takový přístup byl vůči studentům spravedlivý, doporučuje se nechat studenty procvičit si tyto pokročilejší kognitivní dovednosti dříve, než je budou potřebovat v testu.

Přepracovat testy na podobu „s otevřenou knihou" je nesmírně náročné. Prakticky je potřeba opustit všechny znalostní testové položky a vypracovat nové, na vyšších úrovních Bloomovy taxonomie, které by testovaly porozumění a dovednosti.

Přínosy

Jednou z nejtěžších, ale i nejpřínosnějších věcí, které musíte vyřešit při přechodu na open-book testování, je změna pohledu na akademické vzdělávání. Pokud budete upřímní, připustíte, že ve své vlastní praxi neustále vyhledáváte různé informace, vzorce, podrobnosti. Ty potom aplikujete na konkrétní situaci, provádíte jejich syntézu a postupně tvoříte svou vlastní práci. Naši studenti budou v budoucnu dělat totéž a měli bychom je na to připravit. Open-book hodnocení musíme strukturovat tak, aby měřilo jejich schopnost dělat tuto aplikaci a syntézu, místo abychom testovali zapamatování jednotlivých informacích, které zapomenou za měsíc.

Mohou tedy testy/zkoušky s otevřenou knihou pomoci řešit problém podvádění při distančním zkoušení? Zdá se, že mohou, neboť nedávné systematické přehledy prokázaly, že testy se zavřenou a otevřenou knihou dávají srovnatelné výsledky (Sam et al. 2020, Durning et al. 2016).

Ale nejen to. Open-book testy mohou pomoci zapojit studenty do procesů reflexivního a kritického myšlení. Také podporují jejich digitální gramotnost, kritické myšlení a procesy celoživotního učení, což jsou všechno významné ingredience pro budoucí uplatnitelnost absolventů.

Zagury-Orly a Durning, ale nejen oni, pokládají za pravděpodobné, že v budoucnosti se budeme setkávat s **hybridním modelem**, v němž studenti budou hodnoceni pomocí **kombinace zkoušek s otevřenou a zavřenou knihou**. První část zkoušky by mohla být se zavřenou knihou a hodnotila by studenty podle toho, co by měli vědět, aniž by nahlíželi do učebnic. Druhá části zkoušky (s otevřenou knihou) by byla zaměřena na vyšší kognitivní úrovně, na dovednosti, které jsou relevantní pro praxi založenou na důkazech (Zagury-Orly a Durning 2021, Johanns et al. 2017).

Rizika

Jedním z rizik používání zkoušek s otevřenou knihou je to, že učitelé nemusí na začátku vědět, jak navrhnout efektivní zkušební úlohy, které vyžadují kritické myšlení. Studenti mohou být ukolébáni falešnou představou, že si během zkoušky budou moci vše dohledat a řádně se na ni nepřipraví. Může vznikat mylný předpoklad, že zkouška bude snadná a všechny odpovědi půjde nalézt v učebnici nebo v jiných povolených zdrojích.

I při zkoušce s otevřenou knihou musí učitel vymezit okruhy povolených zdrojů informací, aby byla zachována rovnost šancí.

4.8.1 Doporučení pro tvorbu otázek do testů s otevřenou knihou

Pro testování s otevřenou knihou budou užitečné zejména otevřené typy položek, které poskytují studentům více prostoru, tedy například otázky s tvořenou odpovědí. Používejte úlohy založené na příběhu, které vyžadují, aby studenti uplatnili kritické myšlení v reakci na spouštěcí scénář. Předložte studentům údaje a ptejte se, co to může v rámci daného scénáře znamenat. Co dalšího to mohlo ovlivnit, jak se to dá ověřit atd.

Uveďme několik příkladů otevřených položek vhodných pro open-book testování (roztříděných podle úrovní Bloomovy taxonomie):

- Aplikace
  - Seřaďte ..., abyste tím demonstrovali ...
- Analýza
  - Identifikujte chybu v důkazu nebo výpočtu.
  - Vysvětlete tuto situaci pohledem teorie ...
  - Jaké jsou protiargumenty ...
  - Proč se výsledek A liší od výsledku B?
  - Jaký je vztah mezi X a Y
- Syntéza
  - *Popis experimentu.* Jaký očekáváte výsledek?
  - Popište další krok v tomto procesu ...
  - Která metoda je pro toto nejlepší?
  - Který argument je nejsilnější?
- Hodnocení
  - Posuďte situaci za tohoto stavu kritérií.
  - Vyhodnoťte, posuďte, doporučte, co by bylo lepší, ...
  - Jaké změny byste provedli?
  - Co by se stalo, kdyby ...

Používejte v otázkách formulace typu: „co je nejvhodnější“, nebo „co je nejdůležitější“, což studenty vede formulování úsudků a postojů.

Testování s otevřenou knihou a kompetence digitálního věku

- V době vysoké dostupnosti informací nabývá na významu studentova schopnost správně zformulovat otázku. Samotné informace jsou snadno dostupné, ale bez schopnosti posoudit situaci a zformulovat otázku, která z ní plyne, není možné je efektivně využít.
- Do popředí se dostává potřeba orientovat se ve světě, kde je informací spíš mnoho a jejich relevance je nejistá. Informace si mnohdy protiřečí, včetně vědeckých studií dodržujících „evidence based“ přístup. Orientovat se v této „džungli“ bude patřit ke klíčovým kompetencím digitálního věku.

5 HODNOCENÍ A KLASIFIKACE STUDENTŮ

5.1 Standardizace testování

Standardizované testování znamená, že testování je prokazatelně **objektivní, spravedlivé, reprodukovatelné a validní**. Tyto atributy přitom nevycházejí z dobré vůle jednotlivého učitele, ale jsou dosažené systematickým využitím doložitelných postupů a metod.

Používání explicitních a předem známých standardů umožňuje učitelům poskytovat studentům objektivní zpětnou vazbu o výsledcích učení a posilovat tak jejich motivaci. Studenty je standardizované hodnocení vnímáno jako spravedlivější než jiná hodnocení, která srovnatelnost otázek a podmínek neřeší.

Standardizace rovněž zajišťuje, že meze pro průchod testem budou nastaveny podle objektivních (hajitelných) kritérií, že pro testované budou zajištěny rovné podmínky a že výsledky budou navzájem porovnatelné nezávisle na termínu a konkrétních examinátorech.

Sám pojem „standardizace“ se v oblasti hodnocení a psychometrie používá v několika významech, což může být poněkud matoucí:

Standardizace jako **nastavení objektivní meze** (*cutscore*) pro průchod zkouškou. Používají se k tomu osvědčení metody, jako je Angoffova, Ebelova, metoda záložek a další. Jde o to, nastavit mez pro průchod zkouškou podle objektivních a doložitelných postupů, aby mez, která odděluje úspěšné od neúspěšných, nemohla být později zpochybněna.

Standardizace jako **zajištění rovnosti podmínek** při zkoušce. Správný výběr kandidátů, správné ohodnocení výsledků učení vyžadují, aby byl proces objektivní a rovný pro všechny zúčastněné. Musíme tedy zajistit, aby všichni studenti obdrželi rovnocenný test se stejným časovým limitem a veškerým dalšími podmínkami a aby nemohlo dojít k neoprávněnému zvýhodnění některých uchazečů.

Standardizace jako **zajištění souladu se standardy**. Aby mohly být hodnotící postupy jednotlivých škol a institucí vzájemně porovnatelné, případně aby jednotlivé instituce mohly vydávat platné certifikace o testování, musí samy dodržovat standardy, které jsou pro testování klíčové. Příkladem takových standardů mohou být Standardy pro pedagogické a psychologické testování.

Protože zajišťování rovnosti a reprodukovatelnosti podmínek, postupů a hodnocení není samozřejmé, používají se k tomu různé metodické pomůcky a nástroje. Například k zajištění reprodukovatelnosti testů realizovaných více pedagogy, na více školách či v delším časovém období bývá testovým týmem vytvářen **metodický materiál pro hodnotitele**, který se označuje např. jako *pokyny pro hodnotitele*, *příručka k testu*, *pokyny pro zkušební komisi*, *metodické pokyny pro hodnotitele*, *pokyny pro organizaci zkoušky* apod. (Dolejš et al. 2012, Baumgartnerová a Kapustová 2013, Čeština pro cizince 2010). Pedagog tak dostává přesný návod na přípravu, provedení a vyhodnocení testu, aby byla zajištěna reprodukovatelnost výsledků.

Přínosy standardizace

Jednou z hlavních předností standardizovaného testování je, že výsledky jsou dostatečně validní (správné) a reliabilní (spolehlivé) a mohou být objektivně dokumentovány a zreprodukovány. To je odlišuje od běžného školního hodnocení, které je závislé na konkrétním učiteli. Díky standardizovanému testování je možné nejen porovnávat výsledky respondentů napříč jednotlivými školami, ale je možno porovnávat jejich výkon v různých letech.

Standardizované testování poskytuje nejen informaci o znalostech jednotlivce, ale při agregaci výsledků celých testovaných skupin může poskytovat další užitečné informace – např. možnost poměrně přesně porovnat výsledky různých tříd, škol nebo jiných skupin v časové ose.

Rizika standardizace

V některých zemích se postupnou adorací ze standardizovaného testování stala ikona, kterou se zaštiťují i hodnocení, pro něž se tento formát vysloveně nehodí. Podle některých autorů „standardizované testy nemohou měřit iniciativu, tvořivost, představivost, koncepční myšlení, zvědavost, úsilí, ironii, úsudek, angažovanost, dobrou vůli, etické reflexe a celou řadu dalších hodnotných dispozic a atributů. To, co mohou měřit, jsou konkrétní dovednosti a znalosti, tedy nejméně zajímavé a nejméně významné aspekty vzdělávání“ (Ayers 2001). Kritici standardizovaných testů poukazují na uniformitu takového vzdělávacího modelu a produkování absolventů „jako na montážní lince“ (Davidson 2011). Tato uniformita ovšem není následkem standardizovaného testování, ale jeho nekritického používání. Další námitkou je, že nadužívání a zneužívání standardizovaných testů poškozuje výuku, neboť zužuje osnovy. Použití standardizovaného testování bez ohledu na cíle výuky totiž svádí k tomu, že co není testováno, se neučí. Způsob zkoušení se pak stává vzorem toho, jak předmět učit. Příznivci standardizovaného testování reagují, že nejde o kritiku standardizovaného testování, ale jeho nevhodné použití.

5.2 Určení hraničního skóre testu

Nalezení mezí pro průchod testem se často označuje jako „**standardizac**e“. Cílem je najít hranici mezi těmi, jejichž výkon je považován za adekvátní účelu, pro který je zkouška určena, a měli by tedy testem projít, a těmi, jejichž výkon je z tohoto pohledu shledán nedostatečným. Stanovení hraničního skóre bude, podobně jako každá lidská činnost, obsahovat určitou míru chybovosti, což může vést k falešně pozitivním a falešně negativním rozhodnutím. Cílem standardizace je minimalizovat tyto chyby.

Metoda nalezení takové hranice by měla současně být (Norcini 2003):

- hajitelná,
- důvěryhodná,
- podporovaná důkazy v odborné literatuře,
- snadno proveditelná,
- přijatelná pro zainteresované strany.

Napsaný test většinou nestačí jen obodovat. U většiny testů je třeba také říci, kteří studenti v testu uspěli a kteří nikoli, popřípadě je třeba k jednotlivým bodovým ziskům přiřadit i známky. Určení **hraničního skóre** (*pass mark*, *cutscore*) je někdy podceňovaným, a přitom mimořádně důležitým krokem. Při sestavování testu je totiž pro jeho autora či autory poměrně obtížné odhadnout, jak těžké jednotlivé úlohy pro studenty budou, a o to obtížnější je „trefit" se s obtížností vytvářeného testu do určité hodnoty. Přesto se mnohdy hraniční skóre stanovuje zkusmo, jen na základě odhadu jednoho nebo několika málo učitelů. Pokud má testování větší význam, je takový přístup zpochybnitelný – výsledky testu lze napadat tvrzením, že hodnocení bylo nepřiměřeně přísné, nebo naopak že test byl příliš benevolentní a do dalšího studia nebo do praxe pustil i studenty, kteří tam nemají co dělat. V případě standardizovaného testování mají proto i hraniční skóre být nastavena standardizovaným způsobem. Stanovené hraniční skóre je díky tomu podložené, odůvodněné a lze na něj mnohem větší měrou spoléhat. Způsobů, jak standardizovaným postupem hraniční skóre najít, je více. Jednotlivé přístupy se liší podle toho, jaký je účel testu, a dále pak jsou výrazné rozdíly v jejich složitosti a náročnosti na kvalifikované odborníky a jejich čas.

Relativní, absolutní a kompromisní metody

Hodnocení studentů může být založeno na porovnání výkonu studentů navzájem. Takové hodnocení nazýváme relativní. Nebo může být hodnocení založeno na splnění nějakých absolutních (na výkonu ostatních nezávisejících) kritérií. Takové hodnocení nazýváme absolutní. Případně může kombinovat prvky obojího – pak se mluví o metodě kompromisní.

Metoda relativního hodnocení vychází z předpokladu, že ve velkých skupinách je vždy (přibližně stejná) část respondentů připravena tak, aby testem prošla. Je v tom jistý optimismus, neboť pokud by všichni testovaní byli připraveni špatně, metoda stejně vybere nějakou část z nich jako vyhovujících. Proto se hodí zejména tam, kde nejde o konkrétní kompetenci uchazečů, ale o výběr nejlepších z dané skupiny. Typickým použitím této metody jsou např. přijímací testy.

Absolutní hodnocení naproti tomu vyžaduje od uchazečů prokázání konkrétních znalostí a dovedností, které je opravňují k postupu testem, či k výkonu nějaké činnosti. Příkladem takového hodnocení testu je například výstupní test v autoškole, testy u státnic, u atestací apod.

Při teoretických úvahách o hodnocení a klasifikaci studentů můžeme nahlížet tyto dvě odlišné koncepce hodnocení jako projev dvou odlišných pohledů na smysl vysokoškolského vzdělávání.

V prvním případě můžeme nahlížet vzdělávání jako mnohaletý test inteligence, který třídí jednotlivce podle jejich intelektuálních schopností a pracovních návyků. Tento přístup odráží

zájem potenciálních zaměstnavatelů vybrat nejvhodnější kandidáty na omezený počet prestižních míst a pomáhá zajistit, aby na klíčová místa byli vybráni ti nejschopnější. Při studiu tento přístup staví studenty proti sobě, nechává je mezi sebou soutěžit. Metodou hodnocení v tomto případě bude hodnocení relativní.

Druhý pohled je odlišný. Předpokládá, že smyslem vzdělávání je osvítit, posilovat a socializovat občany. Pedagog by se podle tohoto pohledu neměl tolik soustředit na rozřazování studentů podle schopností, ale pomoci jim najít správnou představu o světě a sobě samotných s cílem vybavit je znalostmi, nástroji a návyky, které z nich učiní užitečné a kulturně gramotné členy společnosti. Hodnocení studentů v rámci tohoto konceptu vychází ze splnění absolutních kritérií, a jedná se proto o hodnocení absolutní.

5.3 Relativní stanovení mezí pro průchod testem

Relativní standardizace je způsob vyhodnocení testu, při němž se výkon testovaného jedince porovnává s výkonem relevantní populace. Znamená to, že se zjišťuje, zda zkoušený jedinec dosahuje lepších nebo horších výsledků než ostatní testovaní. Testům, při nichž se výkon testovaného posuzuje v relaci k ostatním, se anglicky říká *norm-referenced tests*, (NRT). Tento přístup k hodnocení výsledku jednotlivce v kontextu výkonu ostatních používají například zkoušky SAT, používané jako rozhodující kritérium pro přijetí na mnohé vysoké školy v USA. V našem prostředí je relativní standardizace porovnávající výkon studentů mezi sebou běžnou součástí přijímacích zkoušek či různých rozřazovacích testů.

Relativní hodnocení je založeno na předpokladu, že výkonnost vzájemně srovnatelných studijních skupin (napříč prostorem a časem) je v zásadě stejná.

Výhody relativního hodnocení

Relativní hodnocení se neváže na obsah testu, ale hodnotí jednotlivé účastníky mezi sebou. Výhodou tedy je, že zabraňuje inflaci nejvyšších hodnocení, zřetelně odliší nejlepší studenty a není nutné individuálně standardizovat každý test zvlášť.

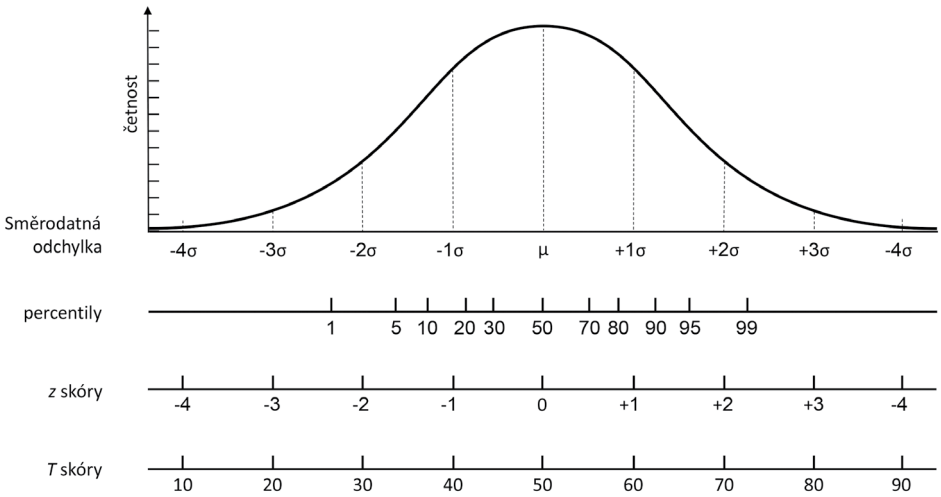
Nevýhody relativního hodnocení

Hodnocení studentů podle relativní standardizace odrazuje od spolupráce a týmové práce, protože si studenti uvědomují, že si navzájem konkurují o omezený počet nejvyšších hodnocení. Snižuje to i motivaci studentů ke studiu oslabením vztahu mezi jejich úsilím a výslednou známkou, protože ta závisí nejen na jejich vlastním výkonu, ale i na výkonu ostatních. Mezi nevýhody relativního hodnocení patří kolísání kvality úspěšných studentů podle kvality dané skupiny. Zejména u menších skupin se tedy může stát, že uspějí i studenti s úrovní znalostí, která neodpovídá našim požadavkům. A obráceně, část studentů nemusí v testu uspět, ani kdyby látku uměli sebelépe. Zvláště v menších a homogenních skupinách může relativní standardizace zveličit nepodstatné rozdíly. S ohledem na tato omezení bychom o užití relativního hodnocení měli uvažovat především ve velkých heterogenních skupinách, v nichž se nepředpokládá spolupráce. Relativní standardizace by se naopak neměla používat ve skupinách menších než 40 studentů.

Z pohledu studenta obsahuje tento způsob hodnocení zjevnou „nespravedlnost“, protože hodnocení nezávisí jen na výkonu samotného studenta, ale i na výkonech ostatních, s nimiž je porovnáván. Je tedy možné, že se stejnou mírou znalostí by student byl v jednom roce hodnocen lépe než v roce jiném. Pro minimalizaci tohoto rizika a zajištění meziroční porovnatelnosti se používá *vyrovnávání obtížnosti testů*, o kterém bude pojednáno v samostatné kapitole.

Relativní hodnocení prakticky

Při relativní standardizaci se skupina rozčlení podle dosaženého počtu bodů a oznámkuje se. Pro stanovení konkrétních známek se používá např. z-skóre nebo percentilové pořadí. Při čtyřstupňové klasifikační stupnici pak hranicím mezi jednotlivými klasifikačními stupni odpovídají např. z-skóre –2, 0, 2, jak je naznačeno na obrázku 5.3.1. Nastavení hraničního skóre při relativním hodnocení může mít arbitrární povahu, např. při přijímacích zkouškách danou kapacitou školy, pro níž přijímací testy realizujeme.



Obr. 5.3.1 Relativní standardizace porovnává výkon jednotlivce s ostatními zkoušenými. Celkové skóre se přitom převádí na odvozené hodnoty. K vyjádření studentova výsledku ve skupině lze použít některou z metod relativní standardizace testu: **Percentilová škála** zhruba udává, jaké procento testované populace dosahuje horších výsledků než daný student. **z-škála** popisuje, jak daleko (měřeno směrodatnou odchylkou dat) je výsledek daného studenta od průměru. **T-škála** používá stejnou metriku, ale vyjadřuje ji na stovkové stupnici.

5.3.1 Percentilová škála

Nejznámější metodou porovnávající vzájemně výkony testovaných je zobrazení jejich výkonů pomocí **percentilové škály**. K výsledku studenta se zjistí *percentil*, který zhruba říká, kolik procent studentů referenční skupiny mělo výsledek horší než daný student. Percentil tak přibližně určuje pořadí studenta přepočítané na interval 0 až 1 (resp. 0–100 %).

Při výpočtu percentilu dosaženého studentem se spočítá počet studentů, kteří měli výsledek horší než daný student, a přičte se polovina studentů, kteří měli stejný výsledek jako daný

student. Pak se určí, jak velkou část celkového počtu studentů tvoří tato skupina. Percentilové pořadí PRi pro osobu s *i*-tým nejhorším celkovým skóre lze odvodit prostřednictvím vztahu:

$$PR_i = 100 \cdot \frac{N_i - \frac{n_i}{2}}{n},$$

kde Ni je kumulativní četnost u daného výsledku, ni je četnost daného výsledku a n je počet testovaných studentů. Kumulativní četnost vyjadřuje počet studentů, kteří dosáhli daného nebo horšího výsledku.

5.3.2 z-skóre

Další metodou standardizace výsledku studenta je výpočet jeho z-skóre. Pro daného studenta jeho z-skóre ukazuje, **nakolik je jeho výsledek nad nebo pod průměrem** (měřeno v jednotkách **směrodatné odchylky**). Jednoduše tedy můžeme z-skóre vypočítat jako rozdíl studentova hrubého skóre X a průměru celé skupiny $\bar{X}$, vydělený směrodatnou odchylkou SD:

$$z = \frac{X - \bar{X}}{SD}$$

Pomocí z-skóre může vyučující snadno identifikovat žáky výtečné ($z > 2$) a naopak velmi slabé ($z < -2$). Snadno může také porovnat studentovy výsledky v různých částech testu.

Podrobnější rozbor dalších metod standardizace (např. C-škála a další) je k dispozici například v publikaci autorů Jeřábka a Bílka *Teorie a praxe tvorby didaktických testů* (Jeřábek 2010).

5.4 Absolutní stanovení mezí pro průchod testem

Absolutní hodnocení (standardizace) je způsob hodnocení testu, při němž se výkon studenta porovnává s absolutními kritérii – s požadavkem na nabytí vědomostí nebo dovedností, které musí mít, aby bylo možno považovat jeho znalost za dostatečnou pro úspěšné absolvování testu (a kurzu). Kritériem se přitom myslí dosažení konkrétní vědomosti a schopnosti, nikoliv dosažení určitého počtu bodů v testu. Např. stanovíme, že po absolvování kurzu první pomoci by měl frekventant znát doporučení týkající se kardiopulmonální resuscitace, jinak musí kurz absolvovat znovu. Jiným příkladem absolutně standardizovaného testu je test v autoškole: je důležité nevypouštět do ulic řidiče, kteří neznají základní předpisy, a to ani v případě, že by se v rámci konkrétní skupiny uchazečů řadili mezi relativně lepší. Absolutně standardizované testy se v angličtině nazývají *criterion-referenced tests* (CRT) a používají se například při národních licenčních zkouškách zdravotních sester v USA (National Council Licensure Examination, NCLEX).

V případě absolutního hodnocení testu je třeba správně zvolit hranici mezi úspěšným a neúspěšným studentem, tedy rozhraní mezi studenty, kteří danou oblast zvládli dostatečně, a těmi, kteří ji dostatečně nezvládli. Ke stanovení této hranice se bohužel občas používají

intuitivní či „tradiční“ postupy a arbitrárně nastavené meze (50 %, 60 %, 75 % apod.) bez dalšího zdůvodnění.

Existuje celá řada metod pro podložené nastavení absolutních mezí různých typů hodnocení studentů. Jejich přehled nalezne čtenář například v obsáhlém díle *Handbook of Test Development* (Downing a Haladyna 2006a).

Mezi nejznámější metody pro stanovení hraničního skóre na základě absolutních kritérií patří Angoffova metoda, Ebelova metoda, metoda záložek, metoda kontrastních skupin a metoda hraničních skupin. Další skupina „smíšených“ metod používá prvky obou přístupů absolutního i relativního – jmenuje z této skupiny např. metodu Cohenové.

Metody zlatého standardu (Angoffova metoda a Ebelova metoda) vycházejí z **expertního posudku** relevantních odborníků, kteří posuzují postupně jednotlivé položky testu a hledají shodu na názoru, s jakou pravděpodobností na ně mají studenti být schopni správně odpovědět. Tyto metody jsou považovány za nejspolehlivější, ale současně jsou velmi pracné a nákladné. Proto se často používají jednodušší a rychlejší metody a v případě pochybností se ověřuje shoda s metodami zlatého standardu.

Podívejme se nyní podrobněji na dvě nejdůležitější metody nastavení meze průchodu testem – Angoffovu a Ebelovu metodu.

5.4.1 Angoffova metoda

Angoffova metoda, resp. její modifikace podle Hambletona a Plakeové (1995), je založena na konceptu **minimálně kompetentního kandidáta**. Tím se myslí takový modelový kandidát, jehož znalosti a dovednosti jsou právě na spodní hraně přípustného minima. Jinými slovy, je to nejslabší student, který by ještě měl testem projít.

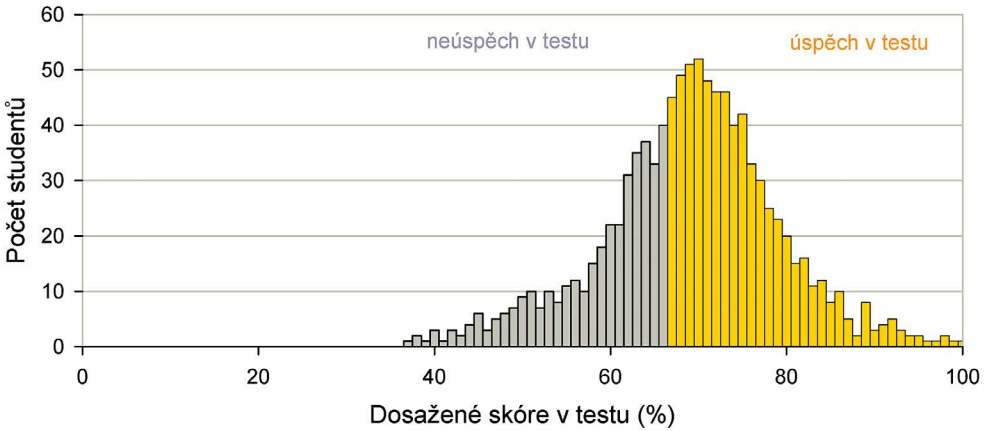
Položky testu posuzuje skupina 4–20 odborníků, kteří mají představu jak o tématu, tak o skutečných kompetencích studentů, tedy nejčastěji učitelů daného oboru. V rámci přípravného setkání by měla být skupina proškolena v metodice a seznámena s požadovanými výkonnostními standardy odpovídajícími kurikulu, aby se sjednotila představa o požadovaných kompetencích. V dalším kroku projdou panelisté samostatně jednu položku testu po druhé a do připravených formulářů si poznamenávají svůj odhad, s jakou pravděpodobností by ji měl onen *minimálně kompetentní kandidát* zodpovědět správně. Doporučuje se, aby několik prvních položek bylo z tréninkových důvodů posouzeno společně. Výsledky odborníků jsou pak zapsány do společné tabulky. Pokud se odhady pro některou položku liší více, než je předem dohodnutá maximální přípustná odchylka mezi odhady (obvykle 15 %), diskutuje se taková úloha v celé skupině a hledá se konsenzus, tj. shoda v hodnocení. Položky, u nichž nelze konsenzu dosáhnout, se z testu vyřazují, protože rozdílné názory odborníků obvykle indikují problém v položce samotné.

Požadované procentní skóre pro průchod testem se pak stanoví jako průměr pravděpodobností úspěšného řešení všech položek použitých v testu. Výhodou tohoto postupu je jeho objektivita a nezávislost na osobních preferencích. Nevýhodou je časová a odborná náročnost postupu (McKinley a Norcini 2014).

Tab. 5.4.1 Tabulka expertních odhadů pravděpodobnosti správného zodpovězení otázky minimálně kompetentním studentem

Položka číslo	Odborník 1	Odborník 2	Odborník 3	Odborník 4	Odborník 5	Odborník 6	Odborník 7	Průměr
1	0,70	0,70	0,65	0,65	0,80	0,60	0,70	0,69
2	0,50	0,50	0,60	0,60	0,55	0,50	0,60	0,55
3	0,80	0,75	0,70	0,70	0,70	0,80	0,70	0,74
4	0,70	0,60	0,70	0,70	0,75	0,60	0,60	0,66
...	...	...	...	...	...	...	...	...
Průměr								**0,66, tj. 66 %**

Poznámka: Pokud by byly v testu otázky typu ANO/NE, byl by samozřejmě nejnižší možný odhad úspěšnosti 0,5, neboť i student, který odpověď nezná, má 50% šanci odpovědět správně. Analogicky u výběrové otázky s pěti možnostmi a jedinou správnou odpovědí bude minimum 0,2.



Obr. 5.4.1 Mez úspěšnosti stanovená pomocí Angoffovy metody rozdělí soubor testovaných na úspěšné a neúspěšné. Tuto metodu podporují i některé testovací programy, například Rogō (o němž bude pojednáno v samostatné kapitole), z nějž můžete obdržet právě tento graf.

Pro nastavení správné meze pro průchod testem je v Angoffově metodě zásadní, že experti musí dobře znát cílovou skupinu a musí být schopni odhadnout, jak budou pro tuto skupinu obtížné konkrétní položky. Navzdory tomu, že se Angoffova metoda v praxi jeví jako funkční, někteří autoři diskutují, zda je představa minimálně kompetentního studenta dostatečnou kotvou pro nastavení standardu. Dle některých názorů by při určování požadovaných standardů výkonu mohlo být vhodnější brát v potaz, čeho je důležité dosáhnout (místo toho, jak obtížné je toho dosáhnout) a čeho by měli dosáhnout všichni kandidáti (nikoli jen čeho by dosáhla skupina úspěšných kandidátů). Existuje tedy podezření, že experti budou mít před očima spíše

průměrného kandidáta a budou přemýšlet, jak by prošel, než aby hledali minimálně přípustnou míru kompetence ve vztahu k požadovaným cílům učení (Burr et al. 2017).

Použití Angoffovy metody
Položky posuzuje skupina expertů a každý z nich u každé položky odhaduje, jaké procento minimálně kompetentních studentů by na danou otázku odpovědělo správně. Experti pracují samostatně, aby se vzájemně neovlivňovali. Výsledky se zapíší do tabulky, v níž řádky představují položky z testu a ve sloupcích jsou odhady jednotlivých expertů.

Po vyplnění tabulky se obvykle posuzuje, zda se experti ve svých odhadech shodli. Položky, v nichž je rozptyl odhadů větší než předem dohodnuté procento (typicky 15 %), je třeba prodiskutovat; často se odhalí nejednoznačná formulace či jiný problém.

Na závěr se vypočte průměr všech odhadů v tabulce. Tento průměr říká, kolik procent z celkového možného počtu bodů by měl dosáhnout minimálně kompetentní student. Jinými slovy tento průměr udává mez úspěšnosti pro daný test – tedy hranici mezi „prošel“ a „neprošel“. Mez úspěšnosti rozdělí soubor testovaných na úspěšné a neúspěšné.

Podpora Angoffovy metody je zahrnuta nejen v testovacím programu Rogō, ale volně (za registraci) ji včetně návodu na YouTube nabízí i společnost Assessment Systems, výrobce programů pro testování a analýzu testů (ASC nedatováno; Assessment Systems 2018).

Podmínky použití Angoffovy metody
Pro úspěšné použití Angoffovy metody je nutné, aby zúčastnění experti měli dostatek zkušeností v dané oblasti a poměrně přesně se shodli v představě, co studenti v daném kurzu musí zvládnout. Experti si tedy musí umět představit, co minimálně kompetentní student umí, resp. by měl umět.

5.4.2 Ebelova metoda

Pro nastavení meze pro průchod testem se vedle zlatého standardu – Angoffovy metody, používá i metoda Ebelova. Využívá také panelu expertů (nejčastěji učitelů), o kterých se předpokládá, že jsou důvěrně seznámeni jak s testovaným tématem, tak s úrovní studentů.

Metoda má tři fáze:
1. kategorizaci otázek,
2. odhad hodnotitelů, jaký podíl kandidátů by měl na úlohy zařazené v jednotlivých kategoriích správně odpovídat,
3. výpočet meze.

Fáze 1

Kategorizace otázek
V prvním kroku je třeba každou položku zařadit do dvou ortogonálních dimenzí, tj. odhadnout její obtížnost a důležitost.

Dimenze „obtížnost“ rozlišuje tři úrovně obtížnosti položek: „lehké“, „střední“ a „těžké“. Posuzovatelé individuálně odhadnou obtížnost každé položky a zařadí ji do příslušné kategorie (Lafave et al. 2008).

Dimenze „důležitost“ má stupně „zásadní“, „důležité“ a „užitečné“.

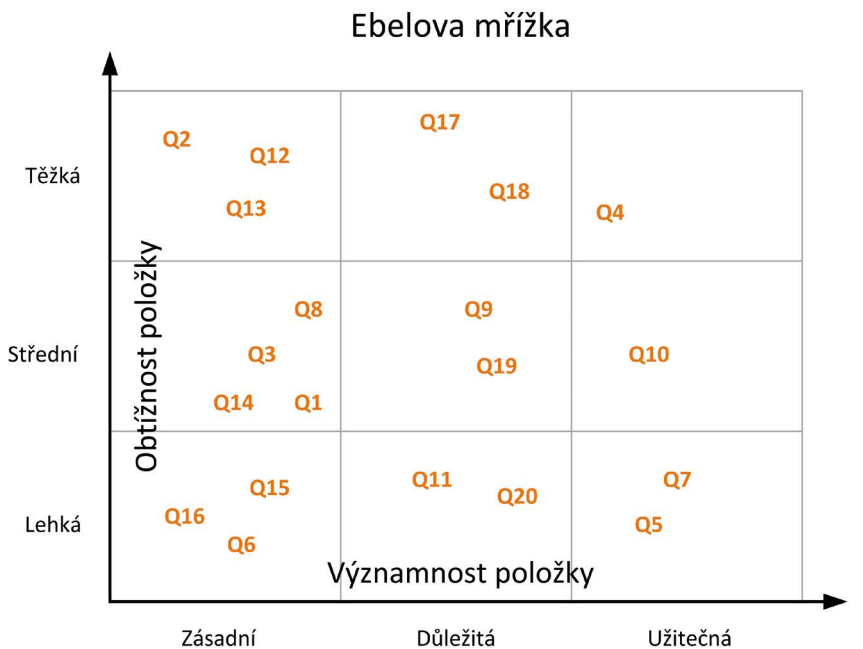
Každá úloha se na základě tohoto zařazení umístí do tzv. Ebelovy mřížky:

Fáze 2

Odhad úspěšnosti
V tomto kroku má každý expert odhadnout, jakou část úloh z každé kategorie by měl správně zodpovědět minimálně kompetentní student. I v tomto případě by experti měli dosáhnout předem dohodnuté shody, například krajní odhady pro každé z devíti polí Ebelovy mřížky by se od sebe neměly lišit o více než 20 %. V opačném případě se experti musí sejít, rozdíly prodiskutovat a odhad opakovat.

Fáze 3

Výpočet meze
Součiny těchto procentuálních odhadů úspěšnosti minimálně kompetentního studenta a počtů otázek v každé kategorii se poté sečtou. Součet se vydělí celkovým počtem položek, výsledek je hledaná hranice úspěšnosti.



Obr. 5.4.2 Ebelova mřížka pro zařazení položek do kategorií obtížnost a význam. Tuto mřížku najdete například v testovacím programu Rogō, který standardizaci pomocí Ebelovy metody nabízí.

Ebelova metoda má i modifikovanou variantu, při níž experti hodnotí pouze relevanci úloh a seskupí je od zásadních po užitečné. Soudci poté určí procento položek v každé ze tří kategorií, na které by měl být hraniční kandidát schopen správně odpovědět. Správná hodnota pro průchod studenta zkouškou je pak průměr napříč kategoriemi. Zatímco **tradiční Ebelova metoda** slouží pro stanovení expertního odhadu minimálního výsledku, kterého by měl student dosáhnout, aby ještě prošel testem (Cantor 1989; Violato et al. 2003), **modifikovaná Ebelova metoda** slouží spíše pro přípravu obsahově validního testu (Aziz 2005; Butterwick et al. 2006; Violato et al. 2002).

Příklad:

Tab. 5.4.2 Ebelova metoda – fáze 1
Experti rozdělí otázky podle dvou kritérií – **význam** a **obtížnost** (viz **Ebelova mřížka** výše). V tabulce jsou počty úloh v každé kategorii.

	Zásadní	**Důležitá**	**Užitečná**
Těžká	**3**	**2**	**1**
Střední	**4**	**2**	**1**
Lehká	**3**	**2**	**2**

Tab. 5.4.3 Ebelova metoda – fáze 2
Experti odhadnou úspěšnost zodpovězení otázek z každé kategorie minimálně kompetentním studentem

	Zásadní	**Důležitá**	**Užitečná**
Těžká	**50 %**	**50 %**	**30 %**
Střední	**70 %**	**70 %**	**50 %**
Lehká	**90 %**	**80 %**	**60 %**

Tab. 5.4.4 Ebelova metoda – fáze 3
Vypočtou se parametry pro jednotlivé kategorie:
počet otázek · odhad úspěšnosti minimálně kompetentního studenta

	Zásadní	**Důležitá**	**Užitečná**
Těžká	**3 · 0,5 = 1,5**	**2 · 0,5 = 1,0**	**1 · 0,3 = 0,3**
Střední	**4 · 0,7 = 2,8**	**2 · 0,7 = 1,4**	**1 · 0,5 = 0,5**
Lehká	**3 · 0,9 = 0,3**	**2 · 0,8 = 1,6**	**2 · 0,6 = 1,2**

Nakonec vypočteme hranici úspěšnosti: sečteme odhady pro jednotlivé kategorie z předešlé tabulky a výsledek vydělíme celkovým počtem otázek. V našem případě tedy 13 : 20 = 0,65, tj. za úspěšné absolvování testu považujeme získání nejméně 65 % z celkového počtu bodů.

Stejný výpočet, který byl výše rozepsaný do tří kroků můžeme provést i v jednom kroku, jak je ukázáno v následující tabulce:

Tab. 5.4.5 Tabulka pro výpočet meze úspěšnosti

Obtížnost	Důležitost (Relevance)	Počet otázek (n)	Proporce (P)	Součin (n · P)
Těžká	Zásadní	3	0,50	1,5
Střední	Zásadní	4	0,70	2,8
Lehká	Zásadní	3	0,90	2,7
Těžká	Důležitá	2	0,50	1,0
Střední	Důležitá	2	0,70	1,4
Lehká	Důležitá	2	0,80	1,6
Těžká	Užitečná	1	0,30	0,3
Střední	Užitečná	1	0,50	0,5
Lehká	Užitečná	2	0,60	1,2
Celkem		**20**		**13,0**
Hranice úspěšnosti				13 : 20 = 0,65, tj. **65 %**

Poznámka: Uvedený příklad je jen ilustrativní, v praxi bychom měli pracovat s většími počty úloh.

O významu Ebelovy metody svědčí i to, že Ebelova mřížka je součástí některých testovacích programů. Konkrétně je přímo zahrnuta v testovacím programu Rogō, o němž bude pojednáno dále.

Kontroverze

Hodnocení obtížnosti položek je nutně subjektivní a výsledek závisí na zkušenostech a proškolení hodnotitelů. Některé práce poukazují na problémy, které to způsobuje. Zdá se, že není jednoduché dosáhnout přijatelné shody mezi hodnotiteli položek. Rovněž se zdá, že existuje systematická tendence hodnotitelů podceňovat obtížnost obtížných položek a přeceňovat obtížnost lehkých (Homer et al. 2012). V experimentální studii nebyla skupina hodnotitelů schopna dosáhnout konzistentního odhadu obtížnosti položky, a když jim byly poskytnuty údaje o fungování položky v testu, snažili se revidovat své úsudky tak, aby odpovídaly těmto údajům, a to dokonce i v tom případě, že data byla podvržená (Clauser et al. 2009).

Ani metody zlatého standardu nemusí být tak robustní, jak se zdá. Mohou záviset na zkušenosti a školení hodnotitelů (Bourque et al. 2020).

Je proto vhodné i osvědčené standardizační metody občas zpětně zkontrolovat například pomocí kotvících položek a IRT analýzy, i kdyby jen proto, abyste se přesvědčili, že opravdu fungují tak, jak mají (Homer a Darling 2016).

5.4.3 Metoda záložek

Metoda záložek funguje tak, že se test doručí reprezentativnímu vzorku účastníků a na základě této skupiny se vypočítají hodnoty obtížnosti pro každou položku. Položky potom seřadíme podle obtížnosti a vyzveme odborníky, aby umístili záložku tam, kde si myslí, že by mělo být hraniční skóre. Vypočte se průměr za všechny posuzovatele a ten je ve skupině diskutován. Poté jsou všichni porotci požádáni, aby umístili druhou záložku, stejně nebo odlišně od té prvé, podle toho, zda se jejich názor při diskusi změnil. Na základě průměru, nebo mediánu druhého skóre je stanovena úroveň požadovaná pro průchod testem.

V dnešní době používáme pro tuto práci počítače, ale v minulosti se to opravdu často dělalo tak, že se položky tiskly do testových sešitů a odborníci do nich doslova vkládali záložku. Na rozdíl od Angoffovy metody musí být u *metody záložek* obtížnost položek zmapována předem na základě provedení testu na reprezentativním vzorku testovaných.

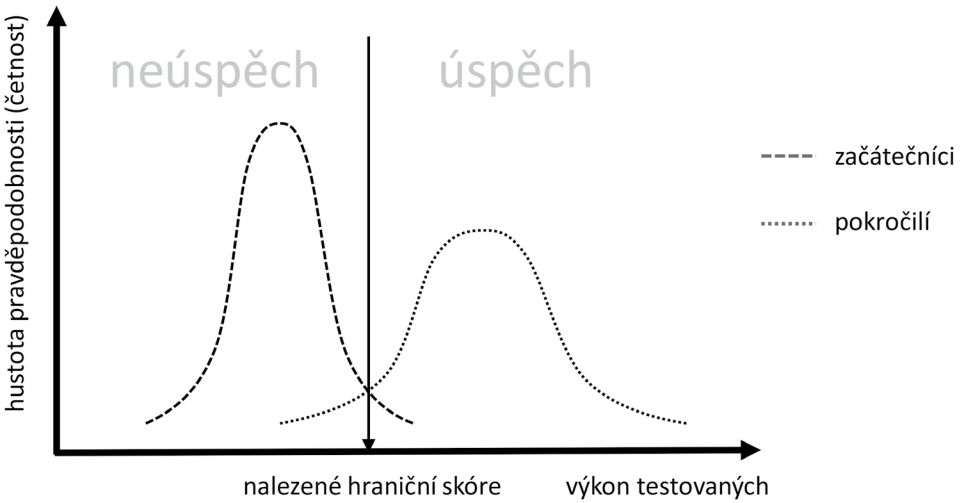
5.4.4 Metoda kontrastních skupin

Dalším postupem pro stanovení hajitelného hraničního skóre (*cutscore*) je metoda kontrastních skupin (Jørgensen et al. 2018). Na rozdíl od předchozích metod se při ní neposuzují jednotlivé položky testu, ale pracuje se jen s celkovými bodovými zisky.

Předpokladem této metody je, že máme k dispozici dvě *kontrastní skupiny*. Test předložíme například skupině začátečníků (tj. třeba studentům) a skupině pokročilých (například lidem z praxe). Pro každou skupinu se vytvoří křivka, která ukazuje rozdělení získaných skóre. Hraniční bod pro průchod testem se stanoví jako průsečík křivek obou kontrastních skupin (obr. 5.4.3).

V reálném případě, zvláště při malém počtu jedinců ve skupinách, nebudou křivky tak hladké a „normální“. Běžně se proto reálně nasbírané body prokládají hladkou křivkou a hledá se průsečík takto proložených křivek.

V každé metodě nastavení hraničního skóre musí být splněny implicitní předpoklady. Zatímco v Angoffově metodě implicitně předpokládáme, že odhady odborníků korelují s obtížností položek, v metodě kontrastních skupin předpokládáme, že výkon v testu koreluje s jinou dostupnou metodou hodnocení. Cílem metody kontrastních skupin vlastně je vyhodnotit, jak výsledky testů předpovídají nějaký „zlatý standard“ hodnocení testovaných, tj. rozdělení studentů do kontrastních skupin. Takovým zlatým standardem může být hodnocení učitele, nebo nějaká uznávaná metrika výkonu. Pokud vhodný „zlatý standard“ nelze nalézt, je třeba sáhnout po jiných metodách – např. metodě záložek, nebo modifikované Angoffově metodě.



Obr. 5.4.3 Nalezení hraničního skóre metodou kontrastních skupin. *(Obrázek ilustruje použití metody kontrastních skupin. Čárkovaná normální křivka představuje výkony skupiny začátečníků. Tečkovaná normální křivka představuje skupinu pokročilých. Svislá černá čára procházející průsečíkem obou křivek představuje mezní skóre vyhověl/nevyhověl.)*

5.5 Kompromisní metody stanovení mezí pro průchod testem

5.5.1 Hofsteeho metoda

Absolutní i relativní standardizace mají svá principiální omezení. Upozornil na to před časem například Hofstee, když prof. Wijnen z Maastrichtské univerzity navrhl zajímavou metodu relativní standardizace (Wijnenovu metodu) (Cohen-Schotanus a Vleuten 2010): Wijnen předpokládal, že průměrný student se pokusí vydat ze sebe to nejlepší a měl by obstát, takže můžeme průměrné testové skóre vzít jako výchozí parametr. Jako hraniční skóre potom můžeme použít arbitrárně zvolenou hodnotu mezi tímto průměrným skóre a jeho hodnotou sníženou o dvě standardní směrodatné odchylky. Výhodou tohoto řešení je, že koriguje vliv nespolehlivosti testu, protože směrodatnou odchylku používá jako měřítko. Prof. Hofstee nebyl s touto metodou spokojen a argumentoval, že tento postup nebere v úvahu absolutní výsledky testu a jako každá relativní standardizace odsuzuje víceméně pevné procento studentů předem k neúspěchu bez ohledu na jejich výkon.

Hofsteehoo metoda

Hofstee proto navrhl smíšenou standardizaci nabízející kompromis mezi absolutní a relativní standardizací. Hraniční skóre pro úspěch v testu se v této metodě nastaví pomocí expertních odhadů.

Předpokládá se, že každý z expertů je detailně seznámen s:

- testem,
- povahou testované skupiny,
- očekávanou úrovní znalostí kandidátů.

Experti musí odpovědět na dvě otázky:

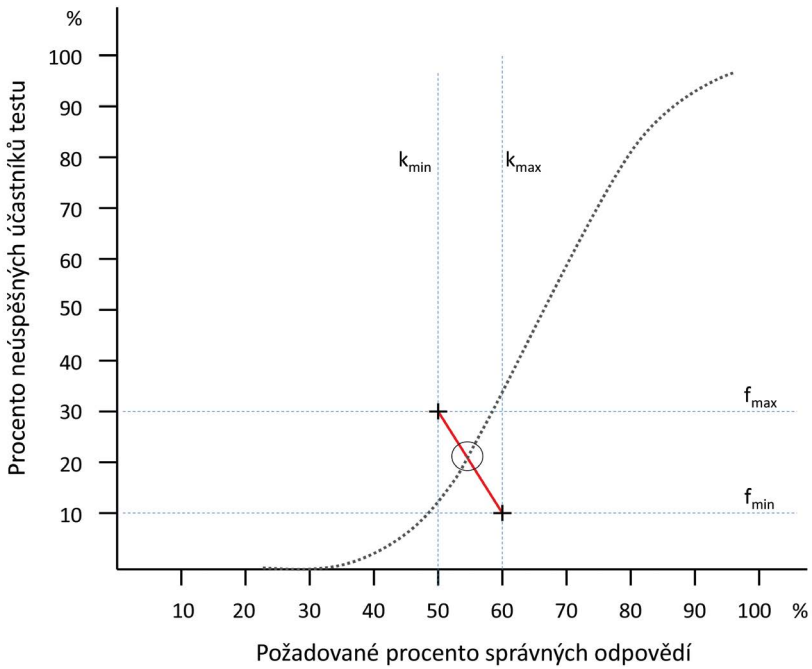
- V jakém rozmezí by se měl pohybovat počet studentů, kteří u daného testu propadnou (např.: *Tímto testem by nemělo projít 10–30 % studentů z dané skupiny*)?
- V jakém rozmezí by se mělo pohybovat minimální skóre pro úspěšné absolvování předloženého testu (např.: *Minimum pro úspěšné složení tohoto testu by mělo být někde mezi 50 a 60 %*)?

Analogicky k absolutní standardizaci jsou tedy experti tázáni na hranici úspěšnosti a současně jsou, podobně jako při relativní standardizaci, tázáni na žádoucí procento úspěšných. Po diskusi nad navrženými hodnotami, kde experti mohou své návrhy ještě upravit, získáme 4 hodnoty:

- minimální a maximální přípustný podíl neúspěšných f_{min} a f_{max},
- minimální a maximální přípustná hranice úspěšnosti k_{min} a k_{max}).

Všechny čtyři hodnoty se stanoví jako mediány návrhů jednotlivých expertů.

Hranice úspěšnosti se stanoví po obodování testu takto: Na základě provedeného testu se sestrojí distribuční křivka skóre v testu. Na vodorovné ose se vynese k_{min} a k_{max}, na svislé ose se vynese f_{min} a f_{max}. Sestrojí se přímka spojující průsečík f_{max} s k_{min} a průsečík f_{min} s k_{max}. Průsečík této přímky s distribuční křivkou se použije jako hranice úspěšnosti v testu (Norcini 2003).



Obr. 5.5.1 Hofsteeho metoda určení hraničního skóre pro průchod testem

Hofsteeho metoda bývá řazena mezi **kompromisní metody**, které se snaží vyřešit rozdíly mezi absolutní standardizací (posuzující procento správně zodpovězených položek) a relativní standardizaci (posuzující procento zkoušených, kteří v testu uspěli).

Hofsteeho metodě je v mnoha aspektech podobná metoda Beukova. Obě vyžadují, aby hodnotitelé stanovili hraniční skóre přímo, bez zkoumání jednotlivých položek, a zahrnují odhady soudců o výkonu celé zkoumané skupiny. Obě metody potřebují znát skutečné rozložení testového skóre, takže nemohou být provedeny, dokud není test dokončen a obodován. U obou lze potřebná doporučení expertů shromáždit ještě před podáním zkoušky. Beukova metoda dokáže na rozdíl od Hofsteeho metody navrhnout hraniční skóre i v případě, že odhady expertů jsou vyšší nebo nižší než body na distribuční křivce dosažených skóre (Bowers a Shindoll 2014).

5.5.2 Metoda Cohenové

Velmi elegantní **kompromisní metodou** pro stanovení hraničního skóre, kombinující kriteriální a relativní posuzování, je metoda podle Cohenové. Vychází z předpokladu, že ve skupině testovaných je vždy několik poměrně dobrých studentů, kteří mají napříč skupinami v podstatě porovnatelný výkon. Nejsou to ti úplně nejlepší, jejichž variabilita může být značná, ale „druzí nejlepší" – výteční studenti, jejichž výsledky jsou na 95. percentilu dané skupiny. Ukazuje se, že skóre těchto výtečných studentů je velmi dobrým a stabilním měřítkem obtížnosti testu. Prvním krokem metody podle Cohenové tedy je určení obtížnosti testu víceméně metodami relativní standardizace – seřazením studentů podle výsledků a určením, kolik bodů odpovídá 95. percentilu (tj. nad jakým bodovým ziskem se umístilo 5 % nejlepších studentů).

Ve druhém kroku určíme hraniční skóre, kterého je třeba dosáhnout pro průchod testem. Cohenová a Van der Vleuten devět let sledovali úspěšnost studentů v různých testech a navrhli, aby tímto hraničním skóre bylo 60 % počtu bodů, kterého dosáhli studenti na 95. percentilu. Do výpočtu se pak přidá ještě korekce na tipování, takže jeho finální podoba je:

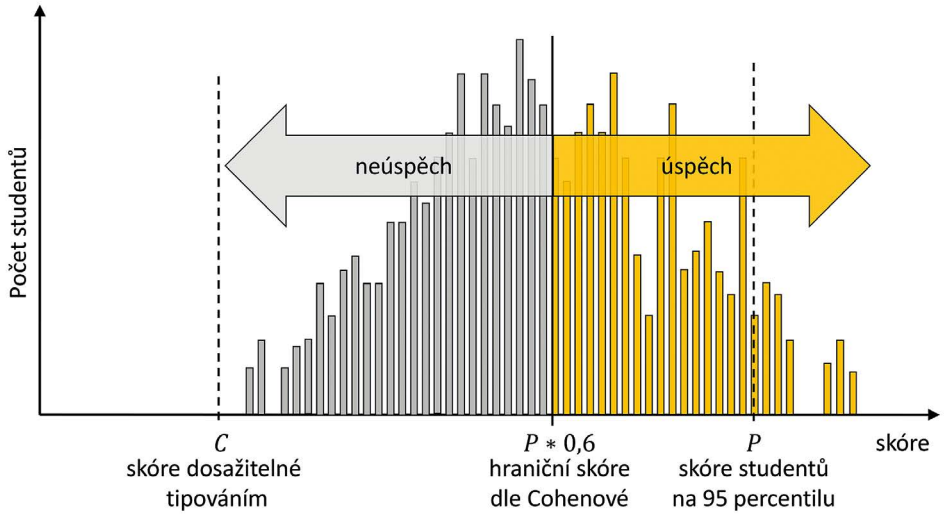
$$PM = 0{,}6 \cdot (P - C) + C$$

PM hraniční skóre (*passing mark*)
P skóre dosažené na 95. percentilu (tj. nad kolik bodů se dostalo 5 % nejlepších studentů)
C skóre, kterého lze dosáhnout tipováním

Příklad Test obsahoval 30 úloh typu „jediná nejlepší odpověď". Za každou otázku bylo možné získat 0 nebo 1 bod, celkově tedy bylo možné získat 0 až 30 bodů. V každé otázce se nabízelo 5 možností. Tipováním tedy bylo možné získat pětinu bodů, tj. 30 : 5 = 6 bodů. Test psalo 80 studentů. 5 % nejlepších z nich, tj. nejlepší čtyři studenti, dosáhli 27, 26, 26 a 24 bodů. Skóre dosažené na 95. percentilu je tedy 24 bodů. Metodou podle Cohenové navrhneme hraniční skóre PM

$$PM = 0{,}6 \cdot (24 - 6) + 6 = 16{,}8 .$$

Student, který získal 16 bodů, ještě neprojde. Student, který získal 17 a více bodů, napsal test úspěšně.



Obr. 5.5.2 Nalezení hraničního skóre podle zjednodušené metody Cohenové (tj. bez korekce na tipování). *Na histogramu zachycujícím distribuci skóre v testu jsou první čárkovanou čarou (vpravo) vyznačeny hodnoty skóre výtečných studentů, kteří se umístili na 95. percentilu. Prostřední čárkovaná linie ukazuje mezní hodnotu pro průchod testem, která je na 60% výkonu výtečných studentů. Poslední čárkovaná line (vlevo) ukazuje hodnotu skóre, které lze dosáhnout prostým tipováním.*

Metoda Cohenové se ověřovala na velkém souboru testů a ukázalo se, že má velmi dobré výsledky a vynikající shodu zejména s Angoffovou metodou. Překonává nevýhody široce používaných kriteriálních i relativních metod stanovení hraničního skóre. Její předností je jednoduchost, rychlost i nízké náklady (Cohen-Schotanus a Vleuten 2010). Nevýhodou může být, že bodovou hranici pro průchod testem nelze oznámit předem, ale až po napsání testu a jeho obodování. To může vyvolávat pochybnosti jak mezi studenty, tak především mezi lidmi odpovědnými za výuku, přestože je předem jasně a jednoznačně oznámený algoritmus, kterým se hraniční skóre stanoví.

5.6 Vyrovnávání obtížnosti testů

Součástí standardizace je i **vyrovnávání obtížnosti testů** (též *harmonizace testů*). Jejím cílem je zajištění vzájemné porovnatelnosti různých běhů nebo paralelních forem testu (například v jednotlivých letech, na jednotlivých školách apod.).

Vyrovnávání obtížnosti (*equating*) je technický postup, jak přepočítat hodnocení studentů z jednotlivých běhů (paralelních forem) testu tak, aby výsledky studentů dosažené v jednom běhu mohly být porovnávány s výsledky studentů v jiných běžích testu (Kolen et al. 2004) .

Vyrovnávání obtížnosti je důležitým aspektem kvality testování a přímo ovlivňuje jeho validitu. Je základním nástrojem při hodnocení vzdělávání, protože hraje zásadní roli při stanovení validity testu ve všech formách a letech.

Při porovnávání testů mezi sebou se používají dva postupy: **provázání testů** (*linking*) a **vyrovnání testů** (*equating*). Provázání dvou testů (*linking*) znamená, že mezi výsledky těchto testů vytvoříme relaci (prolinkování). Např. můžeme vytvořit tabulku odpovídajících skóre z obou testů, dosažených vždy studenty stejné úrovně v obou testech. Na základě této tabulky můžeme říci, že studenti, kteří v prvním testu získali skóre X získají v druhém testu s největší pravděpodobností skóre Y.

Tvrzení, že došlo k vyrovnání obtížnosti (*equating*), je mnohem silnější. Pokud by oba uvažované testy byly úspěšně vyrovnány, pak můžeme prohlásit, že studenti, kteří dosáhli skóre X v prvním testu a studenti, kteří dosáhli skóre Y ve druhém testu, mají velmi podobnou úroveň znalostí a dovedností měřených těmito testy.

Jinými slovy tvrzení, že dvě formy testu jsou vyrovnané (rovnocenné), znamená, že měří stejný obsah a podporují stejné závěry o tom, co studenti znají a umějí. Řekneme-li naproti tomu, že mezi oběma testy existuje provázání (prolinkování), jde o mnohem slabší tvrzení, které pouze znamená, že existuje statisticky měřitelná souvislost mezi skóre v obou testech. Je to dáno tím, že skutečnost, že studenti, kteří dosáhli v prvním testu skóre X a ve druhém skóre Y, ještě neznamená, že oba testy měří opravdu totéž (stejný konstrukt). Provázání testů tedy není dostatečným argumentem, abychom mohli jeden test nahradit druhým. K tomu by bylo třeba ověřit, že testy jsou i rovnocenné, tedy získat i potvrzení odborníků, že oba testy pokrývají stejnými prostředky stejnou doménu.

Vyrovnávání obtížnosti testů může testu buď předcházet (*pre-equating*), nebo jej následovat (*post-equating*). Předběžným vyrovnáváním úrovní testu se myslí sestavování nového testu tak, aby formátem, obsahem a svými charakteristikami odpovídal výchozímu testu. Při dodatečném vyrovnávání obtížnosti testu může být test rovněž sestaven podle pravidel pro předběžné vyrovnávání, ale konečné vyrovnání se provádí až pomocí dat získaných z analýzy proběhlého testu.

Pro vyrovnání obtížnosti dvou testů potřebujeme nějaké srovnatelné údaje. Jednou z možností je zadat oba testy dostatečně velké skupině lidí a porovnat výsledky. Aby se omezil vliv pořadí testů, může být skupina rozdělena a každá polovina dostane testy v opačném pořadí. Nevýhodou tohoto přístupu je nepraktičnost a velká časová náročnost administrace dvou testů. Roste rovněž bezpečnostní riziko, protože expozice dvou testů zvyšuje riziko vynesení jejich položek.

Pro omezení těchto negativních aspektů můžeme použít tzv. *kotvení testu*, kdy se do testu zařadí určitý počet úloh, které jsou ve všech verzích stejné. Tyto tzv. *kotvící položky* pak slouží ke vzájemnému porovnání různých verzí testu. Kotvící položky by měly být reprezentativní, měly by pokrývat rozsah obtížnosti testu a jejich počet by měl dosahovat minimálně 20 % z délky testu (Jelínek a Květon 2011). Výběr témat kotvících položek by měl kopírovat obsah celého testu. Sadu kotvících položek můžeme považovat za „miniverzi“ celého testu (Kolen et al. 2004).

Kotvící položky mohou být buď „vnitřní“, nebo „vnější“, podle toho, jestli se započítávají, nebo nezapočítávají do skóre testu. Mohou být „vložené“, pokud jsou rozptýleny v testu, nebo „připojené“ jako samostatný blok položek na konci testu.

Metod pro vyrovnávání testů je celá řada.

Lineární vyrovnávání je nástroj pro stanovení ekvivalentních skóre mezi dvěma paralelními formami testu v rámci klasické testové teorie. Lineární vyrovnávání je založeno na předpokladu, že se testy liší jen hodnotou svého průměru hrubých skóre a variabilitou výsledků (tedy velikostí směrodatné odchylky). Za těchto předpokladů můžeme přepočítat skóre z jednoho testu na druhý pomocí lineární transformace. Můžeme tedy nejprve transformovat průměrné skóre druhého testu na průměrné skóre testu prvního a potom transformujeme hodnotu skóre druhého testu pro jednu směrodatnou odchylku nad a pod průměrem. Výsledkem je lineární transformace skóre z druhého testu na bodovou škálu prvního testu. Metoda má několik omezení:

- Lineární vyrovnání nebude fungovat v případech, kdy vztah mezi výsledky testů není lineární (např. při asymetrickém rozdělení skóre).
- Transformace platí jen pro tu sadu testovaných, pro které byla spočtena.
- Transformace funguje nejlépe pro skóre vzdálené od průměru o méně než směrodatnou odchylku.

Výhodou lineární transformace je, že je snadno pochopitelná a výpočetně jednoduchá.

Pokud bychom chtěli použít robustnější přepočet, který funguje i pro studenty na okrajích zkoumaného pásma schopností, můžeme použít např. ekvipercentilní vyrovnání.

Ekvipercentilová metoda zajišťuje větší přesnost vyrovnání výsledků podél celé škály možných výsledků. Při tomto vyrovnávání výsledků určíme nejprve v obou testech percentilové pořadí dosažených skóre. Percentilová pořadí mezi oběma testy se poté pomocí tabulky spárují. Druhá možnost je, že se nejprve hrubá skóre převedou na percentily a ty se pak oskórují (už pro oba testy společně). Řada počítačových programů nabízí možnost vypočítat ekvivalentní skóre nebo stanovit percentilové pořadí pro všechna dosažená skóre. Percentilové pořadí se rovněž často používá pro sdělování výsledků studentům. Mezi nevýhody patří, že podobně jako lineární vyrovnání testů je i ekvipercentilní závislé na konkrétním výběru studentů a není bez dalšího použitelné pro jiné skupiny. Obě dosud zmíněné metody jsou v mnohém podobné. Někdy bývá lineární vyrovnání označováno za aproximaci ekvipercentilového (Hambleton et al. 1991).

Metody vyrovnávání založené na IRT. V praxi se více používají metody založené na teorii odpovědi na položku, které se ukázaly být přesnější a stabilnější než metody odvozené z klasické testové teorie a neobsahují závislost na konkrétní skupině testovaných.

Metody vyrovnání testů založené na IRT můžeme rozdělit do dvou skupin:

- metody vyrovnání pozorovaných skóre,
- metody vyrovnání skutečných skóre.

V prvním případě se srovnávají skutečná skóre ve dvou testových formách. Na základě znalosti chování kotvících položek přítomných v obou testových formách transformujeme skóre druhého z testů tak, aby obtížnosti kotvících položek v obou testech splynuly. Ve druhém případě se odhadovaná rozdělení součtových skóre ve dvou formách odvozují z modelu IRT, kam vyneseme do jednoho grafu charakteristické křivky dvou nebo více porovnávaných testů a vyrovnáme je pomocí metodiky ekvipotenciálního vyrovnání (Council of Chief State School Officers 2021).

Jedním z omezení metod vyrovnání testů založených na IRT je potřebný počet testovaných, který by neměl klesnout pod 500. Odhad parametrů v podmínkách malého vzorku není uspokojivý a zhoršuje se s komplexností IRT modelu.

Pro vyrovnávání obtížnosti testů na základě IRT je k dispozici volně dostupný software IRTEQ (Han 2009), nebo je možné využít balíček R equate (Albano 2016).

5.7 Klasifikace studentů

Klasifikace obecně znamená třídění, zařazování subjektů do tříd podle nějakých kritérií. Ve výuce je klasifikace chápána jako zhodnocení výsledků studentů. V případě, že pracujeme s konkrétním testem, který měří nějakou znalost či dovednost, hledáme objektivní, reprodukovatelný a spravedlivý postup, jak výkon studentů v testu ohodnotit pomocí známek.

V tomto smyslu je klasifikace pokračováním či spíše rozšířením standardizace ve smyslu nastavení mezí průchodu testem. Konstrukce klasifikační stupnice, respektive nastavení relace mezi výkonem v testu a klasifikačním stupněm, je **jediný subjektivní prvek, který do celého testování vstupuje**. Je mu tedy třeba věnovat náležitou pozornost (Jeřábek a Bílek 2010).[3]

Pro nastavení relace mezi výkonem v testu a klasifikačními stupni je nutno si ujasnit, jaká má být úloha testu. Úvaha je podobná jako při volbě mezi relativními nebo absolutními metodami stanovení hraničního skóre. Zvažujeme tedy, jestli spíše necháme studenty mezi sebou soutěžit a rozřadíme je do „výkonnostních skupin“ srovnáváním mezi sebou, nebo jestli klasifikujeme podle toho, nakolik daný student dosáhl cílových kompetencí.

5.7.1 Klasifikace založená na porovnání výkonu ve skupině

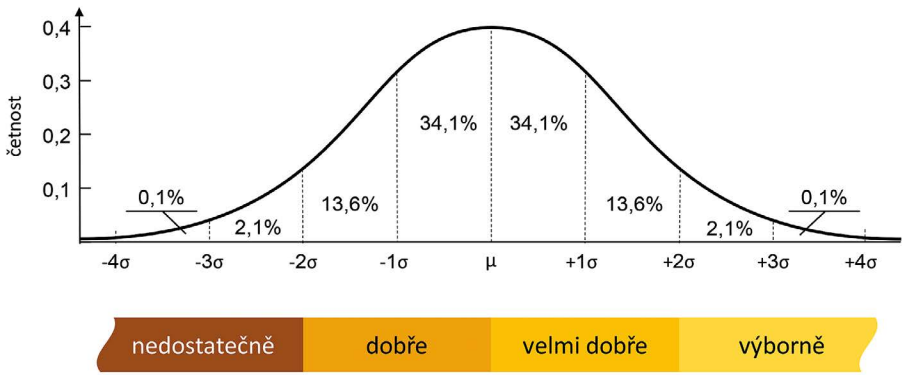
Klasifikace porovnávající výkonnost ve skupině (**relativní klasifikace** – *norm-referenced*) je založena na **výkonu studenta v kontextu skupiny**. Odvozuje hodnocení od pořadí studenta v rámci určité skupiny. Studenty seřadíme podle jejich výkonu v testu a pak jim podle předem dohodnutých mezí přiřadíme klasifikační stupně (známky). Relativní klasifikace je založena na předpokladu, že výkonnost různých studijních skupin (napříč prostorem a časem) je v zásadě stejná. Z pohledu studenta obsahuje tento způsob klasifikace zjevnou nespravedlivost, protože hodnocení nezávisí jen na jeho vlastním výkonu, ale i na výkonech ostatních. Je tedy možné, že kdyby byl v jiné skupině (třeba kdyby test psal v jiném školním roce), se stejnou mírou znalostí by dostal lepší známku.

[3] Poznámka: V textu věnovaném klasifikaci (známkování) studentských výsledků v testu se omezíme na diskusi o testech s uzavřenými položkami s výběrem odpovědi. Klasifikace jiných typů testových položek je diskutována v odborné literatuře (např. (Jeřábek 2010) nebo (McLachlan 2000)).

Chceme-li použít hodnocení založené na relativní standardizaci, je třeba učinit dvě rozhodnutí. Nejprve je třeba stanovit, jaký klasifikační stupeň přiřadíme k průměrnému výkonu. V často používaném čtyřstupňovém klasifikačním systému A, B, C, D můžeme intuitivně zvolit hranici mezi B a C jako klasifikaci odpovídající průměrnému výkonu; není to však jediná možnost.

Dále je nutné předem rozhodnout o hranicích oddělujících jednotlivé klasifikační stupně. Pro stanovení konkrétních známek v závislosti na výkonu se používá např. z-skóre, nebo percentilové pořadí podobným způsobem, jako je popsáno v kapitole o relativním nastavení mezí pro průchod testem.

Při čtyřstupňové klasifikační stupnici pak hranicím mezi jednotlivými klasifikačními stupni odpovídají např. z-skóre –2, 0, 2, jak je naznačeno na obrázku 5.7.1:



Obr.5.7.1 **Příklad klasifikace výsledku studenta v testu pomocí z-skóre** (relativní klasifikace). Skupina se rozdělí podle skóre v testu tak, že vzniklé podskupiny jsou od průměru vzdáleny o dohodnutý počet směrodatných odchylek. Takto vzniklé podskupiny oznámkujeme podle příslušné klasifikační stupnice – v tomto případě čtyřstupňové. Povšimněte si, že nejvyšší hodnocení „výborně“, dostane v tomto případě jen 2,2 % účastníků testu, průměrná hodnocení „velmi dobře“ a „dobře“ dostane velká většina studentů (každá z těchto skupin zahrnuje 47,7 %) a opět jen zcela zanedbatelný počet studentů (2,2 %) dostane nejnižší hodnocení „nedostatečně“.

Rozdělení do podskupin na základě směrodatných odchylek od průměru (z-skóre) není jediná možnost. Alternativou je rozdělit skupinu podle dosažných skóre na stejně velké podskupiny, a těmto podskupinám dát stejnou známku. Rozdělení může být například takové, že nejlepších 25 % dostane hodnocení „výborně“, dalších 25 % dostane hodnocení „velmi dobře“ atd.

Výhody a nevýhody

- Klasifikační systémy založené na porovnání studentů jsou jednoduché a snadno se používají.
- Dobře fungují v situacích, kdy je třeba studenty seřadit, například v rámci vstupních a přijímacích testů do úseku studia, v němž je omezený počet míst.

- Jsou vhodné ve velkých kurzech, které nepodporují spolupráci mezi studenty, ale obecně kladou důraz na individuální úspěch.
- Zjevnou nevýhodou je, že pro hodnocení jednotlivce jsou určují nejen jeho výsledky, ale také výsledky ostatních studentů.
- Hranice hodnocení lze stanovit až poté, co test proběhl. Dopředu se tedy nelze vyjádřit k tomu, jak bude test obtížný (byť je dopředu známo, jak se hranice určí).
- Relativní hodnocení bude spíše použitelné ve velkých neselektivních skupinách, které budou reprezentativní pro celou populaci studentů. V malých třídách (pod 40) nemusí být tato skupina reprezentativním vzorkem. Jeden student může dostat výborné hodnocení, protože je ve skupině se slabým prospěchem, zatímco jeho spolužák se stejným výsledkem v lepší skupině dostane nižší ohodnocení.
- Druhou námitkou proti hodnocení v relaci k ostatním je, že podporuje soutěživost spíše než spolupráci. Tento způsob hodnocení nastavuje mezi studenty vztah přímé konkurence. Když jsou studenti postaveni proti sobě kvůli několika málo nejlepším hodnocením, která mají být rozdána, je méně pravděpodobné, že budou při studiu navzájem spolupracovat.

Kompromisním řešením pro malé skupiny je použít při relativním hodnocení tzv. „kotvení". Známkování se upraví podle toho, jaká je celková (průměrná) úroveň studentů ve skupině (Jacobs a Chase 1992). Pokud učitel použil podobný test v různých letech opakovaně, může nashromážděné výsledky testů použít jako kotvu. Současnou skupinu pak porovnává s touto nasbíranou velkou skupinou. Podobně lze jako kotvu využít dobře sestavený pretest, ve kterém pomocí absolutních kritérií odhadneme schopnost celé skupiny. Modifikace relativního klasifikačního systému pomocí kotvení pomáhá zmírnit pocity konkurence mezi studenty, protože pak již nesoutěží jen mezi sebou.

5.7.2 Klasifikace založená na kritériích

Absolutní klasifikace (*criterion-referenced*) měří úspěšnost studenta **ve vztahu ke kritériím** vyžadovaným pro dosažení toho kterého klasifikačního stupně. Obvykle jsou kritériem počty bodů nebo procento z celkového počtu bodů, jichž musí student dosáhnout, aby mohl dostat příslušné ohodnocení. Nejjednodušším způsobem klasifikace je stanovit, kolik procent z celkového počtu bodů je třeba k dosažení určité klasifikace. Například pro klasifikaci známkou A požadovat 90 % a více, pro B 80–90 % a tak dále. Problém tohoto přístupu je v arbitrárním nastavení hranic jednotlivých klasifikačních stupňů. Pokud dopředu stanovíme podobné bodové hranice, autor testu se do nich musí „strefit". Bude-li test nebo některá jeho varianta obtížnější nebo naopak snazší, než předpokládal tvůrce klasifikační stupnice, bude také výsledné hodnocení vnímané jako nepřiměřeně přísné, nebo naopak benevolentní. Proto je vhodné u důležitějších zkoušek nastavit hranice pomocí odhadů většího počtu expertů, například podle Angoffovy nebo Ebelovy metody.

Při absolutní klasifikaci, na rozdíl od relativní klasifikace, není oznámkování studenta ovlivněno výkonností ostatních a není založeno na vzájemném srovnávání studentů ve skupině. Pokud bychom zkoušeli výrazně nadprůměrnou skupinu studentů, mohou všichni dostat dobré známky, a naopak, pokud by se náhodou sešla skupina slabých studentů, nemusí dobré známky dostat nikdo. Studenti mezi sebou nesoutěží, a je tedy pravděpodobnější, že budou spolupracovat. To může být výhodné i pro jejich aktivní zapojení do výuky, která je často

na spolupráci založená. Klasifikace jednotlivého studenta přitom není ovlivněna celkovým výsledkem třídy.

Absolutní a relativní klasifikace jsou ve skutečnosti do jisté míry provázané. Většina vyučujících nastavuje kritéria na základě svých zkušeností s obvyklým výkonem studentů. Tím se do absolutní klasifikace dostávají relativní prvky. Podobně si někdy učitelé ponechávají jistou flexibilitu v absolutní klasifikaci tím, že studentům předem sdělí, že kritéria v prvním běhu testu mohou být zmírněna, pokud by dobrých známek dosáhlo příliš málo studentů. Například hranice 90 % pro získání klasifikace A může být snížena na 85 %. Pokud by test byl pro studenty obtížnější, než si vyučující představoval, může takto snížit kritéria pro hodnocení. Opačný postup, kdy by vyučující zpřísnil kritéria, protože příliš studentů dosáhlo dobrého hodnocení, se nedoporučuje.

Dalším způsobem, jak klasifikovat studenty podle kritérií, je stanovit cíle předmětu a přidělovat známky podle toho, do jaké míry jich student dosáhl (např. A = student dosáhl všech hlavních i vedlejších cílů předmětu, B = student dosáhl všech hlavních a několika vedlejších cílů atd.).

Propracovanější forma absolutní klasifikace rozlišuje mezi různými typy či úrovněmi znalostí a dovedností, které student prokazuje na různých úkolech. Větší důraz se klade na ty z nich, které odrážejí vyšší úrovně osvojení látky. Tento přístup bere v úvahu jak množství látky, tak úroveň její kognitivní komplexity. Můžeme například vzdělávací cíle svého kurzu rozdělit do dvou skupin: na základní a pokročilé. Základní cíle se týkají minimálních nezbytných znalostí a dovedností, které si studenti musí osvojit. Pokročilé cíle naproti tomu představují vyšší úrovně dovedností, jako použití kritického myšlení, řešení komplexních problémů a podobně.

Pro zjištění, nakolik se podařilo dosáhnout základních a pokročilých výukových cílů, může být přinejmenším pro začátek jednodušší použít dva zcela oddělené testy. Zjednoduší se tím vyhodnocení zkoušky a uchovávání záznamů o ní. Při oddělení testů je také snazší se zaměřit na jednotlivé cíle výuky a zpracovat pro ně testové otázky. Pro posouzení základních cílů výuky to bývá poměrně snadné. Posuzování, nakolik se podařilo dosáhnout pokročilých cílů výuky, bývá obvykle obtížnější, neboť je těžší vymyslet testové úlohy postihující i schopnost nabyté vědomosti aplikovat.

Pro absolvování obou druhů testu lze nastavit odlišné požadavky na výkon studentů, jak je naznačeno v tabulce 5.7.1.

Tab. 5.7.1 Příklad možného nastavení absolutní standardizace pro klasifikaci základního a pokročilého testu v pětibodové klasifikační stupnici.

Klasifikační stupeň	Základní test	Pokročilý test
A	90 % nebo více	85 % nebo více
B	90 % nebo více	75–84 %
C	80 % nebo více	60–74 %
D	80 % nebo více	50–59 %
F	méně než 80 %	méně než 50 %

V uvedeném příkladu požadujeme, aby studenti prokázali zvládnutí alespoň 80 % základních vzdělávacích cílů a 50 % pokročilých cílů. Pokud požadujeme, aby nastavení hranic úspěšnosti bylo objektivnější, můžeme použít některou z metod expertního odhadu popsaných výše.

Kriteriální hodnocení je z hlediska výuky na vysoké škole prioritní. Je sice pro učitele náročnější, vyžaduje promyšlení očekávaných výsledků učení, ale pro studenty je transparentní a odvozené známky by měly být obhajitelné z přiměřeně objektivního hlediska – studenti by měli být schopni vysledovat své známky podle konkrétních výkonů při řešení stanovených úkolů. Kriteriální hodnocení svou transparentností vytváří důležitý rámec pro zapojení studentů do procesu učení.

Při absolutním hodnocení je současně vhodné sledovat rozložení známek ve studijní skupině – jinými slovy, sledovat výsledky kriteriálního modelu známkování z pohledu relativního modelu hodnocení. Pokud narazíme na to, že příliš mnoho studentů dostává špatné, nebo naopak dobré známky, nebo je rozložení nějakým způsobem deformované, pak to může naznačovat, že něco není v pořádku a že je třeba prověřit proces hodnocení. Může jít například o problém s celkovou obtížností hodnotících úloh (například málo náročné zkušební otázky nebo příliš málo otázek, případně úkoly, které nerozlišují mezi studenty s různou úrovní znalostí a dovedností). Osvědčené postupy klasifikace ve vysokoškolském vzdělávání vycházejí z převážně z kriteriálního hodnocení, mírně modifikovaného relativní korekcí a zpětnou vazbou (James 2002).

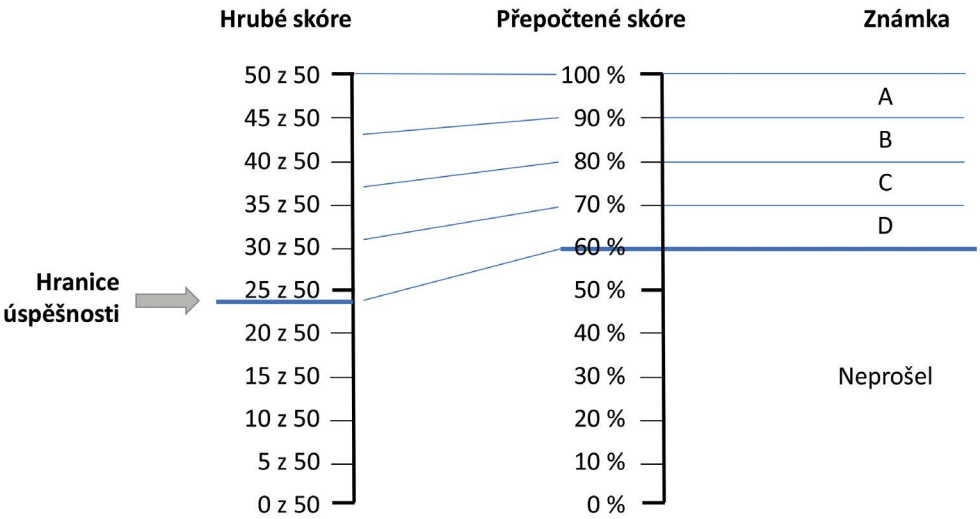
5.7.3 Klasifikační schémata a přepočítávání skóre

Mnohdy je užitečné, aby výsledky hodnocení studenta byly vyjádřené způsobem, který umožňuje srovnávání mezi předměty, případně i mezi obory nebo vysokými školami. Vznikají proto standardní klasifikační schémata (*marking scheme*, *grading system*, *academic grading*, ...) (Wikipedia 2021). Tato schémata umožňují srovnávání v rámci jednotlivých univerzit, ale někdy i v rámci celých států. Výsledky konkrétních testů, popřípadě celých souborů písemných i jiných prací, se přepočítávají na standardní škálu, podle níž se pak udělují známky.

Jako příklad můžeme vzít klasifikační schémata Univerzity v Edinburgu (CSPC 2021). Např. pro pregraduální lékařské obory je relevantní schéma CMS3 (CMS3: Bachelor of Medicine and Bachelor of Surgery):

Přiřazení známky a přepočtení skóre si můžeme ukázat na následujícím příkladu: Uvažujme test, v němž mohli studenti získat od 0 do 50 bodů. Pomocí standardizačních metod tvůrci testu určili, že pro úspěch v testu je třeba získat alespoň 24 bodů z 50 možných (tzv. *pass mark*). Klasifikační schéma označuje nejhorší známku, která odpovídá splnění testu, písmenem D a hranici úspěchu přiřazuje číselnou hodnotu 60 %. V tomto případě tedy hrubé skóre 24 bodů z 50 možných odpovídá přepočtenému skóre 60 %.

Poté, co jsme stanovili přepočet pro hraniční skóre, určíme, jak se budou přepočítávat vyšší bodové zisky. Obvykle se používá jednoduchý lineární přepočet hrubého skóre na přepočtené. V tomto případě bude přepočtenému skóre 70 % (minimum pro známku C) odpovídat hrubé



Obr. 5.7.2 Přepočet testových skóre na klasifikaci.

Metodami standardizace byla určena hranice úspěšnosti v konkrétním testu na 24 bodů z 50. V klasifikačním schématu, které používá daná instituce, tomuto meznímu hrubému skóre odpovídá přepočtené skóre 60 % (hranice pro známku D). Výsledky lepší než 24 bodů z 50 se pak rovnoměrně rozdělí k jednotlivým známkám.

Tab. 5.7.2 CMS3 schéma

Počet bodů	Známka	Popis
90–100	A	Výborně
80–89	B	Velmi dobře
70–79	C	Dobře
60–69	D	Splnil (*pass*)
50–59	E	Podmíněné selhání (může být přehodnoceno)*
0–49	F	Selhání

* *Podmíněné selhání je forma výzvy studentovi, aby si známku ve stanoveném čase opravil. Pokud tak neučiní, počítá se horší známka.*

skóre 30,5 bodu z 50, přepočteného skóre 80 % dosáhne student s 35 body z 50 atd. Jinými slovy, nejprve jsme určili, kteří studenti v testu uspějí, a pak jsme je mechanicky rozdělili do jednotlivých klasifikačních stupňů. Přepočet lze matematicky vyjádřit následovně:

$$Z = 60 + \frac{40}{100 - P} \cdot (p - P)\,,$$

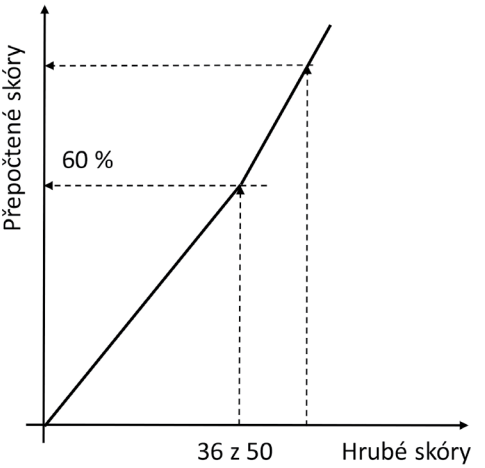
kde Z je přepočtené skóre (z kterého dle CMS3 schématu určíme známku), P je minimální hrubé skóre nutné k úspěšnému absolvování daného testu (*pass mark*) a p je hrubé skóre dosažené daným studentem.

Podobně pak můžeme přepočíst i hrubé skóre nižší než hranice úspěšnosti (*pass mark*). Přepočítávání v celém rozsahu možných bodových zisků pak může být dvojitě lineární.

Tato transformace má podobu dvou navazujících úseček, ale v linii je mírný ohyb. V anglické literatuře se proto označuje jako „dogleg“ (Thompson 2019).

5.7.4 Škálování

Složitějším způsobem převodu skóre na jinou stupnici, vhodnější pro reportování, je škálování. Na rozdíl od předchozího přístupu, kdy byl přepočet založen na třech bodech (0 %, hranice úspěšnosti a 100 %), je škálování podrobnější. Slouží především k tomu, abychom mohli srovnatelným způsobem sdělit výsledky různých paralelních forem (variant) testu, které se mohou mírně lišit svou obtížností (viz též Vyrovnávání obtížnosti testů).



Obr. 5.7.3 Dvojitě lineární (linear dogleg) přepočet hrubých skóre na přepočtená skóre. V tomto testu bylo metodami pro standardizaci určeno, že hraniční skóre má být 36 bodů z 50. Tomu má odpovídat přepočtené skóre 60 %. Vyšší bodové zisky se lineárně přepočtou tak, aby hrubé skóre 50 bodů z 50 odpovídalo přepočtenému skóre 100 %. Obdobně nižší bodové zisky se lineárně přepočtou tak, aby hrubé skóre 0 bodů z 50 odpovídalo přepočtenému skóre 0 %.

Řada důležitých testů jako jsou ACT, SAT, GRE a MCAT, je vykazována na stupnicích, které jsou zvoleny záměrně tak, aby nesly určitou informaci. SAT a GRE mají historicky nastaven nominální průměr 500 a směrodatnou odchylku 100, zatímco ACT má nominální průměr 18 a směrodatnou odchylku 6. Jedná se vlastně o stejné škály, protože nejsou ničím jiným než přepočteným z-skóre.

„Průměrné hodnoty“ byly vybrány arbitrárně, a poté byly nastaveny hranice rozsahu skóre pomocí násobku směrodatných odchylek. Díky tomu se hodnocení v testech SAT a GRE pohybují v rozmezí od 200 do 800 a v testech ACT v rozmezí od 0 do 36. Pro lepší pocit zkoušeného jsou stupnice nastavené tak, že za „odevzdání formuláře“ u zkoušky SAT obdržel 200 bodů. Výsledek 300 bodů se může jevit jako povzbudivé číslo, ale je to jen 100 bodů nad minimem, což odpovídá pouhému 3. percentilu.

Často se vůbec neuvádí hrubé skóre dosažené v testu, ale výhradně některé přepočtené skóre. Pokud existuje více verzí testů, které se srovnávají, škálování vyrovná skutečnost, že se verze liší obtížností. Zvolená bodovací škála by měla být alespoň tak široká, jako je počet položek v testu, jinak by se ztratila část rozlišení, které výsledky testu přinášejí.

Při škálování se nejprve definuje rozsah, ve kterém mají ležet sdělované výsledky. Začíná se obvykle nalezením střední hodnoty a směrodatné odchylky hrubých skóre v testu, a ty se poté převedou na jinou, přepočtenou střední hodnotu a směrodatnou odchylku. Již zmiňované lineární a dvojitě lineární přepočty nemusí stačit, používají se proto i složitější transformace. Pro vyrovnávání paralelních forem testu je vhodná např. ekvipercentilní transformace (viz kapitola Vyrovnávání obtížnosti testů).

6 ANALÝZA TESTU A JEHO POLOŽEK

Sumativní didaktický test lze chápat jako nástroj na měření míry znalostí a dovedností, které si student při výuce osvojil. Výsledky rozhodného testování mohou mít pro účastníky testu zásadní důsledky – např. přijetí či nepřijetí do dalšího studia, certifikaci pro určité povolání či udělení titulu. Pokud by byly testy neadekvátní svému účelu a neměřily kvality, které očekáváme, že budou měřit, mohlo by docházet při rozhodování k podstatným chybám a tím ke snížení efektivity a ohrožení věrohodnosti celého systému. Je proto důležité kvalitu testů i testových úloh měřit a průběžně sledovat.

Část vlastností testů (a úloh) je popsatelná pomocí intuitivně pochopitelných kategorií **obtížnosti** a **citlivosti**. Obtížnost můžeme chápat jako pravděpodobnost, s níž testovaný na daný test nebo úlohu neodpoví správně. Citlivostí se myslí míra, s níž test nebo položka rozlišují mezi lépe a hůře připravenými studenty.

Kromě těchto intuitivních metrik používáme pro popis vlastností testu ještě pojmy **reliabilita** a **validita**. Reliabilita (spolehlivost) vyjadřuje přesnost a opakovatelnost testu. Pomocí reliability vlastně zjišťujeme, zda přezkoušení studenta jinou verzí téhož testu povede k potvrzení předchozího výsledku. Validita (správnost) říká, zda test nebo položka měří znalost, kterou měřit chceme.

Mimo tyto tradiční metriky se v posledních letech věnuje značná pozornost **férovosti** (spravedlivosti) testů. Ověřujeme, zda test nějakým způsobem neznevýhodňuje některé skupiny testovaných.

Položková analýza umožňuje vyhodnotit na základě analýzy proběhlého testu vlastnosti jednotlivých úloh (položek testu), zejména jejich obtížnost a citlivost. Součástí položkové analýzy může být i **analýza distraktorů**, která podrobněji zkoumá kvalitu nabízených možností v uzavřených (výběrových) úlohách. Zabývá se např. tím, jak testovaní volili jednotlivé navržené odpovědi v závislosti na celkovém výkonu testovaného.

Výsledky položkové analýzy poskytují pro každou úlohu řadu psychometrických údajů, které umožňují konstruovat nezávislé testy s obdobnými vlastnostmi.

Součástí analýzy testu by měly být jeho **popisné statistiky** a grafické zobrazení výsledků, nejčastěji ve formě **histogramů**. Porovnání grafů z jednotlivých běhů testu nám pomůže posoudit, zda nedošlo například k vynesení některých úloh použitých v testu apod.

Podívejme se nejprve na vlastnosti testu jako celku, především na jeho reliabilitu a jeho validitu.

6.1 Reliabilita

Výsledek testu by měl v ideálním případě záviset jen na tom, co chceme testovat, tj. skóre získané v testu by mělo záviset jen na schopnostech testovaného v oblasti, kterou testem zkoušíme (tzv. *skutečné skóre*). V reálném životě se ale výsledek testu (hrubé skóre) od skutečného skóre liší v důsledku víceméně **náhodných chyb**. Každý test má tedy určitou **spolehlivost a přesnost**, kterou vyjadřujeme jako **reliabilitu** (spolehlivost, preciznost, reprodukovatelnost) (Schindler 2006).

Reliabilita říká, do jaké míry při opakovaném nezávislém hodnocení týchž jedinců dostaneme podobné výsledky. Vliv na ni má třeba, jak dobře testovaný rozumí zadání úloh, zejména pokud jsou komplikovaně formulovány a on je z jiného kulturního nebo jazykového prostředí. Výsledek testu také závisí na pozornosti testovaného, ovlivní jej prostředí v místnosti a vyrušování během testu, nebo to, zda testovaný pracuje ve stresu. Reliabilitu snižuje i případné hádání odpovědí atd.

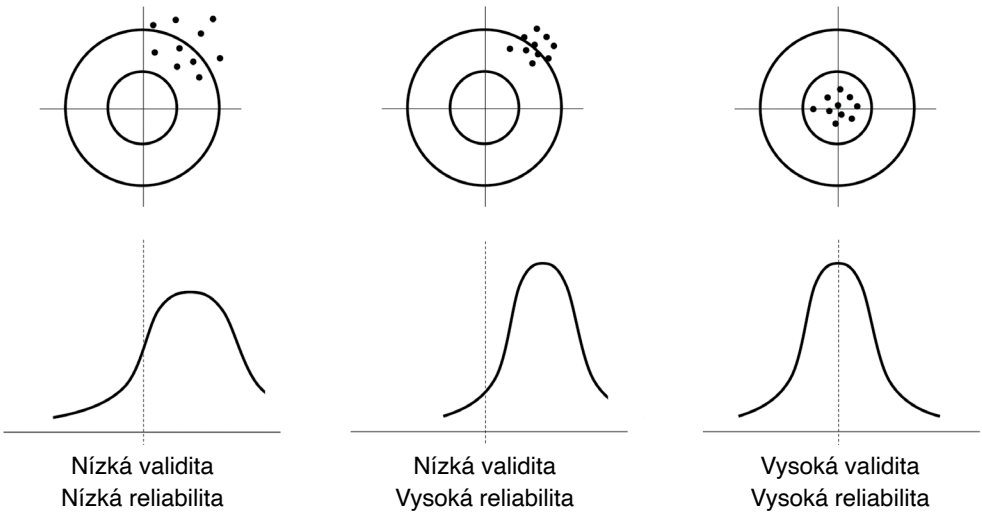
Reliabilita nabývá hodnot mezi 0 a 1 (0 % a 100 %). Zjednodušeně si můžeme reliabilitu představit jako míru potlačení náhodných chyb vyjádřenou v procentech. Reliabilita 50 % znamená, že přibližně za polovinou variability pozorovaného skóre (hrubého skóre) je variabilita skutečného skóre (tj. měřených schopností testovaného) a druhá polovina jde na vrub náhodných chyb. Reliabilita 0,8 znamená, že variabilita pozorovaného skóre je z 80 % tvořena variabilitou skutečné schopnosti a 20 % tvoří chyby.

Minimální výše reliability testu, kterou lze považovat za uspokojivou, závisí na kontextu, např. na počtu úloh v testu a počtu testovaných. Pokud jde o počet testovaných, bylo uveřejněno několik doporučení, která se shodují, že pro rozumný odhad reliability by neměl klesnout pod několik stovek (Kline 1995). Pokud by byl počet účastníků výrazně nižší, je možné místo reliability pracovat s *celkovou užitečností*, jak ji zavedl Van der Vleuten (Vleuten 1996; Vleuten a Schuwirth 2005).

Pro účely pedagogické diagnostiky jedinců, např. při rozhodování o přijetí k dalšímu studiu, se zpravidla požaduje koeficient reliability minimálně 0,8 (a vyšší). Pro ostatní školskou praxi postačuje koeficient reliability pohybující se v blízkosti hodnot 0,6–0,7 (Schindler 2006). U testů s malým počtem úloh (10 a méně) zpravidla reliabilita nepřesahuje hodnotu 0,6–0,8. Nižší hodnota reliability nemusí nutně znamenat, že test je vysloveně špatný, ale musí se s ním zacházet opatrně a neměl by sloužit jako samostatný podklad pro rozhodování. Naopak velmi vysoký koeficient reliability (blízký 1) může znamenat, že úlohy v testu jsou tak vnitřně konzistentní (navzájem si tak podobné), že jsou vzájemně zastupitelné a v testu by jich mohlo být méně, aniž by to výrazně zhoršilo jeho vlastnosti.

Reliabilita popisuje technickou kvalitu a vnitřní konzistenci testu, avšak nikoli jeho správnost. Test může být spolehlivý – mít vysokou reliabilitu, ale přitom nemusí měřit to, co by měl, takže může mít současně nízkou validitu. Reliabilita testu je ovšem nutným předpokladem jeho validity.

Koncepty reliability a validity a vztah mezi nimi lze dobře ilustrovat na příkladu:



Obr. 6.1.1 Schéma přibližující vztah reliability a validity

6.1.1 Odhady reliability

Reliabilitu z principu nelze spočítat přímo, ale můžeme ji zkusit odhadnout. Při odhadu reliability se snažíme určit, do jaké míry je variabilita výsledků testů způsobena variabilitou skutečných skóre a do jaké míry chybami v měření. (Připomeňme, že chyby měření mohou mít náhodnou a systematickou složku.) Cílem je navrhnout testy tak, aby zdroje chyb byly minimalizovány.

Pro odhad reliability testů se podle situace používají čtyři hlavní přístupy (Murphy a Davidshofer 2005; American Educational Research Association 2014):

- Reliabilita jako shoda mezi posuzovateli (*inter-rater reliability*): Tato tzv. *klasifikační konzistence* se používá k posouzení míry, do jaké různí hodnotitelé poskytují navzájem shodné odhady stejného jevu. Využívá se zejména tam, kde do skórování testu vstupují subjektivní faktory. Podmínkou objektivity je srovnatelné proškolení hodnotitelů, čímž se sjednotí požadovaná kritéria. Je třeba mít na paměti, že vysoká shoda mezi hodnotiteli ještě neznamená, že by testovaný dosáhl stejného výkonu při opakování testu. Shoda mezi posuzovateli a konzistence jejich hodnocení je tedy podmínkou, ale ještě nestačí pro zaručení vysoké reliability skóre testovaných osob (American Educational Research Association 2014).
- Test-retest reliabilita (spolehlivost testu při jeho opakování): Používá se k posouzení, jak jsou navzájem konzistentní výsledky stejného testu při opakovaném použití na stejné skupině. Jejich konzistenci lze posoudit pomocí výpočtu jejich korelace. Zatímco u jevů, kde opakovaná měření téže veličiny jsou na sobě nezávislá (měření délky, váhy, ...), dává tato metoda výtečné výsledky, pro didaktické testování je obtížně použitelná. Jednotlivé běhy testu totiž nelze považovat za nezávislé. Při krátké pauze mezi testy si mohou účastníci

pamatovat, jak odpovídali při prvním běhu testu, a výsledná reliabilita bude nadhodnocena. Doporučuje se proto rozestup minimálně 3 měsíce, i když i tam hrozí zkreslení, neboť studenti se mezi tím mohou látku naučit. Při opakování testu s větším časovým odstupem mohou zase studenti látku již zapomenout a bude se nutně lišit i dosažený výsledek. To skutečnou reliabilitu „opticky" snižuje.

- Reliabilita paralelních verzí testu: Používá se k posouzení konzistence výsledků dvou testů vytvořených podle stejného předpisu, stejným způsobem, ze stejného tématu. Posuzování reliability paralelních verzí testu (výpočtem jejich korelace) odstraňuje sice problémy s nezávislým opakováním testu, které jsme viděli při metodě test-retest, ale přináší nové obtíže s tvorbou ekvivalentních forem testu. Paralelní formy by měly být vytvořeny podle přesně stejného plánu testu a jejich položky by měly mít stejné psychometrické charakteristiky. Někdy se objevuje snaha vytvářet „paralelní" položky změnou číselných hodnot v příkladech, změnou jmen a názvů v textu úloh apod. V praxi se však ukazuje, že nově odvozené položky mívají zpravidla vyšší obtížnost, takže je třeba vytvářet dvojice položek už při jejich psaní, a pak je do testů losovat.
- Spolehlivost jako vnitřní konzistence: Používá se k posouzení konzistence výsledků napříč položkami v rámci testu. V předchozím odstavci jsme diskutovali posouzení reliability paralelních forem testu, tedy korelaci mezi testem a paralelním (opakovaným, leč nezávislým) testem. Protože vytvořit paralelní nezávislý test bývá obtížné, používá se jako přiblížení (náhrada paralelního testu) náhodné rozdělení jednoho testu na dvě poloviny. Vzniklé poloviny pak uvažujeme jako dva nezávislé paralelní testy. Korelace mezi těmito dvěma polovinami (korigovaná o délku testu) je dobrým odhadem korelace „skutečný" test – opakovaný test. Problémem tohoto přiblížení je, že neznáme vliv náhodného rozdělení testu na poloviny. Možná by jiné rozdělení na dvě poloviny přineslo jinou korelaci, a tedy jiný odhad reliability test-retest. Mohli bychom sice vystřídat všechna možná dělení na poloviny a pak vzít střední korelaci jako měřítko spolehlivosti, ale to by při testu s více položkami mohlo být velmi pracné. Jednodušší je rozdělit test na nejmenší možné části (jednotlivé položky) a vypočítat korelace mezi nimi. Tento přístup je dobrým měřítkem vnitřní konzistence a základem pro hojně používané Cronbachovo alfa (Schuwirth a Vleuten 2011). Cronbachovo alfa lze brát jako průměr odhadů reliability u tesů rozdělených na všechny možné poloviny (Crocker a Algina 1986).

6.1.2 Cronbachovo alfa

Cronbachovo alfa bylo vyvinuto Lee Cronbachem v roce 1951 s cílem poskytnout měřítko vnitřní konzistence testu, tedy míry, nakolik všechny položky v testu měří stejný konstrukt a jaký je rozptyl měření v testu. Pokud jsou položky v testu vzájemně korelované, hodnota alfa se zvyšuje. Hodnota koeficientu alfa je ovlivněna také délkou testu. Pokud je test krátký, hodnota alfa se snižuje. Hodnota alfa je vlastnost konkrétního provedení testu – závisí na složení konkrétní testované skupiny.

Při interpretaci Cronbachova alfa je třeba mít na paměti, že koncept reliability předpokládá, že test je homogenní v tom smyslu, že testové položky zkoumají stejný latentní rys na stejné škále. Pokud je tento předpoklad porušen, může být skutečná reliabilita testu odhadem podhodnocena. U vícedimenzionálních testů by mělo být alfa vypočteno pro každý měřený konstrukt zvlášť. Pokud nejsme o jednodimenzionalitě testu přesvědčeni, musíme se na Cronbachovo alfa dívat

jako na dolní hranicí odhadu reliability. Odhady přijatelných číselných hodnot Cronbachova alfa se pohybují v širokých mezích (od 0,70 do 0,95) (Tavakol a Dennick 2011b).

Nízká hodnota alfa může být ovlivněna nízkým počtem otázek, heterogenitou měřeného konstruktu, nebo malou korelací mezi položkami. Nejjednodušší metodou, jak zjistit příčinu nízkého alfa, je vypočítat korelace jednotlivých položek s celkovým skóre testu. Položky s nízkou korelací (blížící se nule) nesouvisejí se zbytkem testu a je možné je odstranit.

Pokud je Cronbachovo alfa příliš vysoké, může to naznačovat, že některé položky jsou v testu již nadbytečné a nepřinášejí žádnou další informaci navíc. Maximální doporučená hodnota alfa je 0,90 (Streiner 2003).

Použití Cronbachova alfa můžeme demonstrovat na následujícím příkladu:

Představme si, že chceme zkoušet sčítání čísel od jedné do deseti. Snadno sestavíme test, ve kterém bude větší množství (řekněme padesát) doplňovacích úloh typu „3 + 4 = ...“. Ten, kdo sčítat umí, odpoví správně na všechny otázky, nebo nanejvýš udělá jen ojedinělé nahodilé chyby. Naopak ten, kdo sčítat vůbec neumí, se jen ojediněle strefí do správného řešení. Takto sestavený test můžeme označit za vnitřně konzistentní – testuje jediný koncept (sčítání v daném oboru čísel). Cronbachovo alfa se bude blížit jedné.
Pokud bychom nyní v testu vyměnili polovinu úloh za příklady typu „12 : 3 = ...“, situace se změní. Dáme-li takto změněný test žákům prvních či druhých tříd základní školy, budeme testovat dva koncepty: sčítání a dělení. Lze si představit, že část žáků bude umět dobře sčítat, ale zcela pohoří v dělení. Test již nebude tak konzistentní, jako v předešlém případě; nemůžeme už také říci, že kterékoliv dvě úlohy testují totéž. Cronbachovo alfa se sníží.

Mluvíme-li o *vnitřní konzistenci testu*, měli bychom si uvědomit, že nezávisí jen na samotných úlohách, ale také na cílové skupině. Pokud bychom totiž dali onen upravený test s jednoduchými početními úlohami gymnaziálním studentům, pravděpodobně by se jevil opět jako vnitřně konzistentní a Cronbachovo alfa by se blížilo jedné: z pohledu takovéto pokročilejší skupiny testovaných totiž zkoušíme opět jediný koncept – základní početní úkony. Zda je konkrétní úloha věnovaná sčítání nebo dělení, bude v tomto případě lhostejné.

Z uvedených příkladů vyplývá, proč by Cronbachovo alfa konkrétního testu nemělo být ani příliš nízké, ani příliš vysoké. Je-li test nekonzistentní, budou se nám špatně interpretovat jeho bodové výsledky. Představme si, že náš test s úlohami na sčítání a dělení dáme žákům druhých tříd. Podle dosaženého počtu bodů asi poměrně snadno rozpoznáme skupinu těch, kteří umí dobře sčítat i dělit, a skupinu žáků, kteří sčítat ani dělit neumí vůbec. Mezi nimi budou žáci, kteří sčítají i dělí, ovšem s mnoha chybami, ale také ti, kteří výborně sčítají, neumí však vůbec dělit. Z výsledku takového testu nepoznáme, zda konkrétní žák obstál v obou činnostech srovnatelně, nebo byl v jedné výborný a v druhé propadá; pravděpodobně by bylo vhodné namísto jednoho testu použít dva samostatné, každý zaměřený na jinou dovednost.

Pokud se naopak Cronbachovo alfa blíží jedné, znamená to, že mnoho studentů z dané skupiny odpovědělo buď na všechny otázky správně, nebo na všechny otázky špatně. Jinými slovy, odpověděl-li student správně na několik prvních otázek, odpovídal správně i na všechny ostatní a obráceně. V uvedeném testu sestaveném pouze z příkladů na sčítání by asi bylo zbytečné

dávat žákům padesát otázek – pokud bychom test zkrátili, dostali bychom pravděpodobně zcela srovnatelné výsledky. Test s velmi vysokým Cronbachovým alfa navíc nemusí dostatečně jemně rozlišovat mezi různými úrovněmi znalostí.

Ačkoliv je Cronbachovo alfa široce používané, je třeba mít na paměti všechna jeho omezení.

6.2 Validita

Validita (správnost, pravdivost, věrnost, platnost) popisuje, do jaké míry test měří to, co chceme, aby měřil. Validita testu se týká míry, v jaké jsou závěry založené na jeho výsledcích smysluplné a užitečné. Tedy jestli je test správně navržen a zda jeho výsledek není příliš ovlivněn **systematickými chybami**.

Podle definice je validita testu míra, ve které shromážděné důkazy a teorie podporují navrhovanou interpretaci testových skóre při doporučeném způsobu použití testu (American Educational Research Association 2014). Z definice je patrné, že validita (na rozdíl od reliability) je konstrukt, který nelze měřit přímo. Lze na něj pouze usuzovat ze souvislostí s dalšími pozorováními.

V praxi se musíme ptát, zda náš test měří skutečně to, co by měřit měl. Výslednou validitu přitom ovlivňuje celý řetězec předpokladů, které je třeba mít na paměti. Například pokud použijeme test z profilových předmětů na střední škole k výběru uchazečů o studium medicíny, pak bychom měli zvážit:

1. Zda test měří znalosti a schopnosti studenta, které mohl nabýt na střední škole.
2. Zda schopnost zvládnout předměty vyučované na střední škole predikuje schopnost absolvovat vysokou školu.
3. Zda absolvování vysoké školy predikuje schopnost absolventa být dobrým lékařem.
4. Zda výsledek testu neovlivňují nějaké vedlejší faktory (např. dostupnost přípravných materiálů).

Je zřejmé, že přesné vyjádření validity naráží na některé principiální problémy. Je například obtížné popsat, kdo je *dobrý lékař*. V zahraniční se to někdy obchází tím, že se zkoumá míra akademické a profesní úspěšnosti absolventů. Jde ale o zjednodušení, neboť dobrým lékařem může být i zcela neambiciózní absolvent, který odejde dělat obvodního lékaře do pohraničí. Při odhadu validity přijímacích testů se proto často spokojíme s *mírou úspěšnosti* vyjádřenou jako schopnost úspěšně absolvovat školu v čase k tomu vymezeném. Aby kompromisům nebyl konec, nelze v praxi často čekat na ověření validity celou dobou řádného studia a spokojíme se s akademickou úspěšností např. po prvním roce studia. Tím do našeho řetězce předpokladů přibude další, kde předpokládáme, že úspěšné absolvování prvních ročníků studia predikuje v přijatelné míře úspěšnost v celém studiu. Takový předpoklad může mít ve skutečnosti jen omezenou platnost, například proto, že první roky studia na lékařských fakultách se věnují teoretickým oborům a vyšší ročníky klinickému studiu.

6.2.1 Validace testu

Protože validitu testu nelze měřit přímo, soustředíme se v praxi na jeho *validaci*, tj. shromažďování důkazů, že je test validní. Validace testu přestavuje shromáždění empirických dat a logických argumentů, které prokazují, že závěry jsou skutečně vhodné. Důkazy, kterými se snažíme doložit validitu, mohou mít různou povahu. Jednotlivé typy důkazů se navzájem nenahrazují, spíš se prolínají a doplňují.

6.2.1.1 Obsahová validace

Obsahová validace se zabývá vztahem mezi obsahem testu a cílovými kompetencemi, jichž má testovaný dosáhnout. Během přípravy testu (zejména při plánování a recenzi testu) se několik zkušených pedagogů zabývá otázkou, zda a nakolik úlohy obsažené v testu pokrývají zkoušené znalosti a dovednosti a obráceně, jestli všechny úlohy spadají do zkoušené oblasti a nezkoušejí něco jiného. Zkoumá se také, jestli je zastoupení úloh věnujících se jednotlivým tématům proporčně vyvážené. Posouzení obsahové validity je svým způsobem kontrola, zda byl dodržen plán testu (tj. jeho blueprint – specifikační tabulka).

Vždy při tom závisí na účelu testu. Například je-li cílem testu hodnocení vzdělávacího programu, mohou být jeho předmětem i témata, která nebyla probírána, a testem se vlastně zjišťuje, jak si studenti s novou problematikou poradí. Naproti tomu, pokud je test určen k posouzení, zda testovaný může postoupit do dalšího ročníku, musí obsah testu striktně vycházet z obsahu vyučované látky (American Educational Research Association 2014).

Při obsahové validaci je třeba rovněž sledovat, zda interpretace dosažených testových skóre nezvýhodňuje některou z podskupin testovaných.

6.2.1.2 Kriteriální validace

Obsahová validace zmíněná výše slouží k ověření, zda připravovaný test odpovídá cílům zkoušeného oboru. Neprokazuje ale, jak takový test odpovídá objektivním kritériím (např. studijnímu úspěchu), s nimiž bychom náš test rádi porovnali. K tomu slouží *kriteriální validace*, která zkoumá vztah mezi výsledkem testu a objektivním nezávislým kritériem nebo kritérii (známkami, postupem do dalšího studia, úspěšným absolvováním školy, ...).

Obecně rozlišujeme dva typy studií, které souvislost testu s kritériem studují, studie **souběžné** a **prediktivní**.

Při zkoumání *souběžné validity* (*concurrent validity*) srovnáváme validovaný test a kritérium současně a porovnáváme, zda jde skutečně o alternativní způsoby měření stejného konstruktu (American Educational Research Association 2014). V principu může být souběžným kritériem jiný, již ověřený test. Zjišťujeme pak, do jaké míry se shodují výsledky zkoumaného nového testu s tímto ověřeným testem. Míru shody můžeme vyjádřit např. pomocí korelačního koeficientu.

Prediktivní validita popisuje, do jaké míry náš test **předpovídá budoucí hodnoty nějakého kritéria**. Predikční validita je klíčovým parametrem všech přijímacích testů. Účelem přijímacích testů je vybrat studenty s nejlepšími dispozicemi pro budoucí studium. Je proto na místě zkoumat, zda používané testy skutečně predikují úspěšnost ve studiu. V praxi to znamená, že se zjišťuje korelace výsledků přijímacích zkoušek s úspěšností studia, nebo že se z dat odhaduje regresní model, kterým lze úspěšnost ve studiu předpovídat.

Kromě toho nás může zajímat, zda daný test **přináší novou informaci nad ty, které získáváme jinými způsoby**, tedy jaká je jeho **validita inkrementální** neboli přírůstková. V případě zmíněných přijímacích testů nás může například zajímat, zda přijímací testy přidávají novou informaci o budoucím studiu uchazeče nad tu, kterou nám poskytuje jeho středoškolský prospěch. Např. studie (Štuka et al. 2012) na základě dat studentů přijatých na 1. LF UK ukázala, že středoškolský prospěch vysvětlí zhruba 15 % variability úspěšnosti ve studiu. Výsledek z přijímací zkoušky zvýší procento vysvětlené variability úspěšnosti na 22 %, přidání informace o úspěšně absolvovaných profilových předmětech na střední škole na 25 % a informace o roku maturity dokonce na 30 %. Všechny zmíněné efekty byly v modelu signifikantní (tedy statisticky průkazné), prokázala se tak jejich přírůstková validita (Štuka et al. 2013).

Zájemci o validaci testů mohou najít podrobnější informace v řadě pramenů (American Educational Research Association 2014; Byčkovský a Zvára 2007; Zvára 2008).

6.2.1.3 Konstruktová validace

Konstruktová validita testu vyjadřuje, jestli test měří požadovaný psychologický konstrukt. Patří k nejdůležitějším průkazům validity. Testem se snažíme posoudit schopnosti studenta, které nelze žádným způsobem změřit přímo – jsou latentní. Snažíme se proto vytvořit abstraktní konceptuální konstrukt (model), který nám pomáhá tuto latentní schopnost pochopit a popsat.

Jako příklad si představme test z matematiky. Latentní schopností může být schopnost řešit určitý typ slovních matematických úloh. Pokud má test tuto latentní schopnost hodnotit, ale testové úlohy jsou psané dlouhými kostrbatými souvětími, může se stát, že ve skutečnosti měříme spíše schopnost orientovat se v komplikovaném a dlouhém textu – tedy úplně jiný koncept. Výkon je pak ovlivněn faktorem, který nemá souvislost s měřeným konstruktem, z hlediska testu je to tedy konstruktově irelevantní rozptyl.

Prokázání konstruktové validity vyžaduje shromáždění více zdrojů důkazů. Je zapotřebí důkaz, že test měří to, co má měřit (v tomto případě znalost základní matematiky), a také důkaz, že test neměří, co nemá měřit (čtenářské dovednosti). To se označuje jako *konvergentní a diskriminační důkazy validity*.

Konvergentní důkazy validity spočívají v poskytnutí důkazů, že dva testy, o nichž se předpokládá, že měří úzce související dovednosti nebo typy znalostí, spolu silně korelují. To znamená, že dva různé testy nakonec hodnotí studenty podobně. *Diskriminační důkazy validity* podle stejné logiky spočívají v poskytnutí důkazu, že dva testy, které neměří úzce související dovednosti nebo typy znalostí, spolu silně nekorelují (tj. poskytnou rozdílné pořadí studentů).

Jak konvergentní, tak diskriminační validita poskytují důležité důkazy pro konstruktovou validitu. Jak již bylo uvedeno dříve, test základní matematiky by měl měřit především konstrukty související s matematikou, a nikoli konstrukty týkající se čtení. Aby bylo možné určit konstruktovou validitu konkrétního testu z matematiky, bylo by třeba prokázat, že korelace výsledků tohoto testu s výsledky jiných testů z matematiky jsou vyšší než korelace s testy ze čtení.

6.2.1.4 Zobecnění průkazu validity

Pro praktické použití vztahu mezi testem a kritériem v nových podmínkách (např. stejný kurz další akademický rok) je třeba provést důkaz, že ověření validity získané v předchozích

podmínkách lze použít k předpovědi míry validity v novém, ale podobném prostředí. Tomuto kroku, který je protikladný k *hypotéze situační specifičnosti*, se říká **zobecnitelnost validity** a obvykle se ověřuje pomocí metaanalýz. Snažíme se při nich posoudit, zda jsou přiměřeně porovnatelné parametry předchozích studií posuzujících kriteriální validitu. Výsledky zpravidla podporují argumenty pro zobecnění validity, což naznačuje, že není nutné provádět nový důkaz validity v každém novém případě, pokud se podmínky a parametry studie významně neliší (Schmidt a Hunter 1977).

6.2.1.5 Souhrn důkazů validity

Celková validace integruje jednotlivé důkazy validity zamýšlené interpretace testových skóre, včetně zahrnutí technické kvality testu, férovosti testu a reliability testových skóre.

6.3 Popisné statistiky a grafy

Prvním krokem testové analýzy bývá shromáždění popisných statistik a jejich grafická prezentace. Po provedeném testu vás určitě bude zajímat, jak dopadl. Popisná statistika poskytuje číselný popis testu a přehledně sumarizuje jeho výsledky. Poskytuje informaci, jaký byl celkový počet testů, kolik bylo maximum a minimum dosažitelných bodů, jaký byl nejlepší a nejhorší dosažený výsledek či jaký byl průměrný počet bodů.

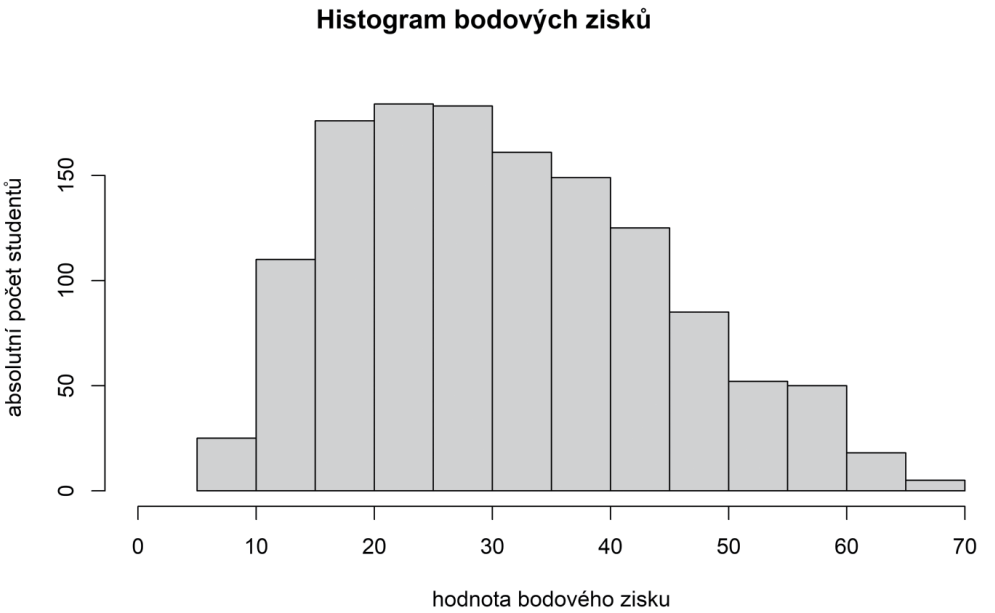
Z pohledu formálního členění zde najedeme popisné charakteristiky polohy (např. průměr, medián a modus) a charakteristiky variability (např. směrodatnou odchylku, minimální a maximální hodnoty proměnných, charakteristiky špičatosti a šikmosti).

Podobnou souhrnnou popisnou statistiku uvidíte i ve výstupech, které poskytují komerční softwarové nástroje pro analýzu testů (Iteman a další). Tabulka popisných statistik o výsledcích konkrétního testu:

Tab. 6.3.1 Tabulka popisných statistik

počet účastníků testu	1354
počet žen	964
počet mužů	390
minimální možný počet bodů	0
maximální možný počet bodů	70
dosažené minimum	25
dosažené maximum	70
průměr	28,6
medián	37,5
směrodatná odchylka	12,4

Protože intepretace některých charakteristik testu z číselných údajů nemusí být zcela intuitivní, používá se názorné grafické vyjádření. Např. pro znázornění rozdělení celkových počtů bodů studentů v testu (hrubých skóre) se s výhodou používá histogram.



Obr.6.3.1 Histogram bodových zisků v reálném testu

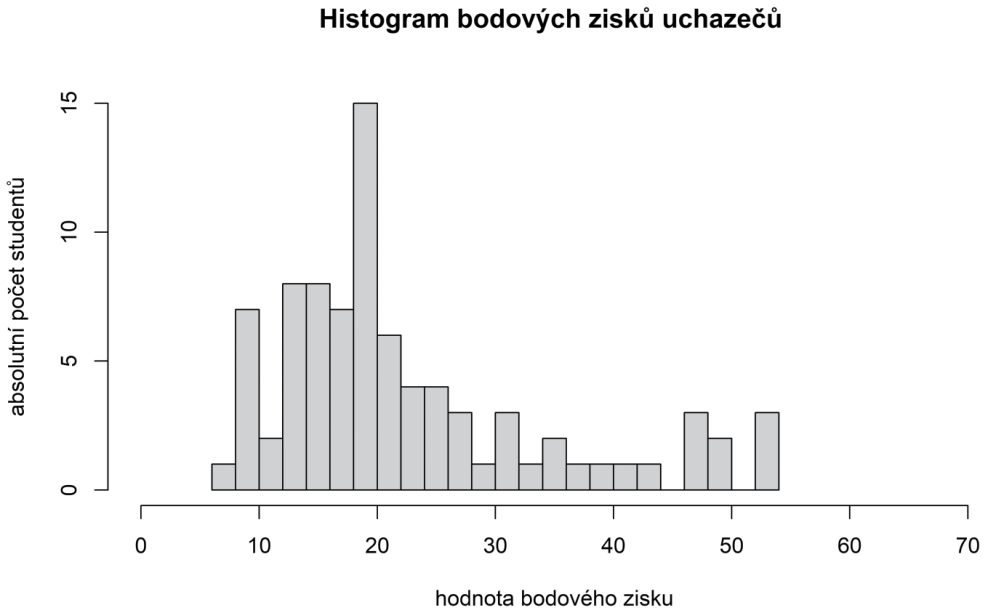
Histogram je grafické znázornění distribuce dat pomocí sloupcového grafu, v němž výška sloupců vyjadřuje četnost sledované veličiny v daném rozsahu hodnot a šířka sloupce reflektuje rozsah tohoto intervalu. V ideálním případě, pro velký soubor hodnot a zjemňující se členění intervalů, by se distribuce měla blížit normálnímu rozdělení. V praxi však bývá distribuce složitější a odráží specifické podmínky testu.

Například asymetrický histogram na obrázku 6.3.1 naznačuje, že test byl pro danou skupinu testovaných poměrně obtížný, protože většina pozorování je soustředěna do sloupců s nízkým počtem bodů. V tomto konkrétním případě to ovšem bylo záměrné a žádoucí, protože bylo potřeba vyselektovat jen malou skupinu nejlepších uchazečů, a pro ty byl test nastaven.

Druhý histogram (obr. 6.3.2) ukazuje, jak se zmenší jeho vypovídací hodnota při malém počtu testovaných (a současně bylo pro tvorbu histogramu nastaveno příliš podrobném dělení intervalů na vodorovné ose, čímž do jednoho sloupce padlo méně případů a graf je tak zatížen větší náhodnou chybou). Informace je zašuměná a můžeme se jen dohadovat, zda dvouvrcholové rozdělení je reálné a indikuje, že mezi testovanými byla podskupina s nezvykle dobrými výsledky. Pokud bychom tento jev vyhodnotili jako závažný, bylo by možné metodami **forenzní testové analýzy** (viz kapitola o bezpečnosti) dále zkoumat příčinu tohoto jevu.

6.4 Položková analýza

Po dokončení ostrého běhu testu pravděpodobně nejprve budeme chtít test vyhodnotit, abychom zjistili, jak si v něm studenti vedli. Avšak odpovědi studentů neobsahují jen informaci



Obr.č. 6.3.2 Histogram bodových zisků v reálném testu při malém počtu účastníků testu a nevhodně podrobném členění

o jejich znalostech a schopnostech, do výsledků testu se promítají i vlastnosti testových úloh. Tak jako vyhodnocením testu můžeme získat informaci, jak si vedli jednotliví účastníci testu, můžeme položkovou analýzou zkoumat (psychometrické) vlastnosti položek. Položková analýza je důležitá i pro autory a recenzenty úloh, protože jim poskytuje objektivní zpětnou vazbu o tom, jak se jimi vytvořené či recenzované položky v praxi chovají. Zatímco recenzenti dobře dokážou posoudit např. obsahovou validitu, jejich odhady obtížností úloh bývají často velmi subjektivní. Proto nás položková analýza zajímá jako zdroj objektivní reflexe našich položek, nástroj pro jejich průběžné vylepšování a pro edukaci autorů a recenzentů úloh (Maierová et al. 2015).

Základním předpokladem položkové analýzy je, že analyzovaný test je konzistentní, tj. že ho psali kvalifikovaní učitelé, a že se tedy skládá z úloh měřících jednu oblast znalostí nebo schopností. Kvalita jednotlivých položek se posuzuje porovnáním odpovědí studentů na položku s jejich celkovým skóre v testu.

Hlavními charakteristikami úloh jsou jejich **obtížnost** a **citlivost**.

6.4.1 Obtížnost položky

Jednou ze základních charakteristik testové úlohy je, jestli na ni alespoň někteří účastníci testu dokážou správně odpovědět, jestli není pro testované příliš obtížná.

Obtížnost položky můžeme odhadnout podle toho, jaký podíl účastníků testu na ni dokázal správně odpovědět. Tomuto podílu se říká index obtížnosti a značí se P:

$$P = \frac{n_S}{n}.$$

kde n_S je počet testovaných, kteří na danou položku odpověděli správně a n je počet všech testovaných.

Index obtížnosti nabývá hodnot mezi 0 a 100 % (respektive 0 a 1). Čím víc studentů na položku odpovědělo správně, tím je hodnota indexu blíže ke 100 % (respektive 1). Je to trochu matoucí, neboť mluvíme o obtížnosti a tento index je nejvyšší, když je položka nejsnazší.

Proto se zavádí doplňková veličina, **hodnota obtížnosti**. Hodnota obtížnosti udává poměr testovaných, kteří na danou úlohu odpověděli nesprávně, jde tedy o doplněk indexu obtížnosti:

$$Q = \frac{n_N}{n} = 1 - P$$

Pro složitěji bodované úlohy se indexy počítají pomocí aritmetického průměru bodových hodnocení všech testovaných v dané položce a nejvyššího dosažitelného počtu bodů za ni.

Při sumativním testování přinášejí největší užitek, nejlepší diskriminaci, úlohy, jejichž hodnota obtížnosti není ani příliš velká, ani příliš malá (typicky 20–80 %). Je to logické, protože položky, které jsou příliš obtížné, nerozliší mezi slabšími a lepšími účastníky testu, neboť příliš těžkou úlohu prostě nikdo nevyřeší. Podobě na opačném konci obtížností nepřinese téměř žádnou informaci příliš snadná položka, protože příliš snadnou úlohu vyřeší i velmi slabí účastníci testu. U položek s okrajovými hodnotami obtížnosti se tedy zákonitě snižuje jejich diskriminační schopnost.

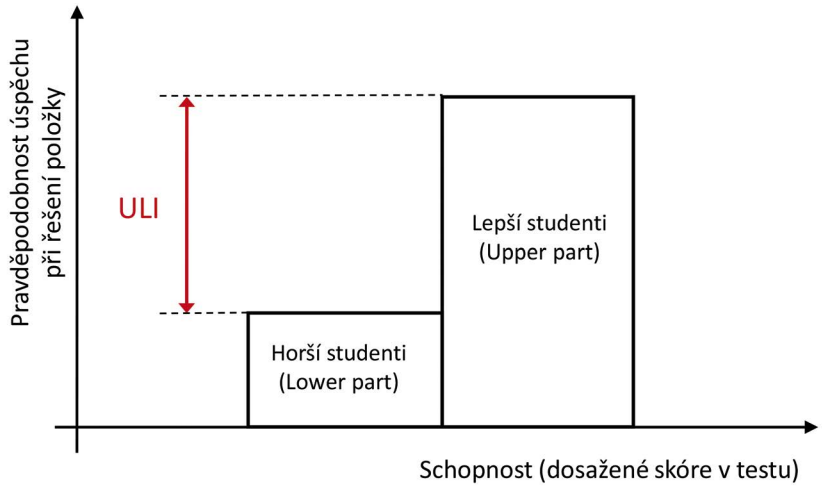
Povšimněme si, že tento odhad obtížnosti položek (zavedený v rámci klasické testové teorie, CTT) je závislý na testovaných. Pro každou skupinu vyjde hodnota jinak, a budou-li se skupiny navzájem výrazněji lišit, může obtížnost téže úlohy vycházet pro každou skupinu úplně jinak. Překonání této provázanosti mezi obtížnosti a testovanými umožňuje teorie odpovědi na položku, v níž je schopnost testovaných jedním z parametrů.

6.4.2 Citlivost položky

Citlivost úlohy, neboli její diskriminace, popisuje její schopnost rozlišovat mezi různě výkonnými studenty. Představme si, že skupinu studentů rozdělíme na lepší a horší, např. podle jejich celkového výsledku v testu. Rozdíl mezi průměrnou úspěšností obou skupin při řešení konkrétní úlohy vyjadřuje schopnost této položky rozlišovat mezi lepšími a horšími studenty a označuje se jako **upper-lower index (ULI)**.

ULI spočítáme jako rozdíl úspěšnosti mezi skupinou lepších (U – *upper*) a horších (L – *lower*) studentů při řešení konkrétní položky.

$$ULI = P_U - P_L,$$



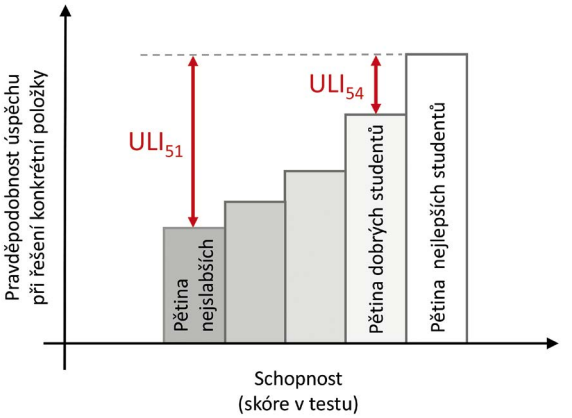
Obr. 6.4.1 Index ULI – rozdíl v pravděpodobnosti správné odpovědi na položku mezi lepšími a horšími studenty

Pro testy, které mají rozlišit mezi nejlepšími a druhými nejlepšími, např. při přijímacích testech s velkým převisem zájemců, nás může zajímat, jak položka rozlišuje právě v okolí dělícího skóre mezi přijatými a nepřijatými. V takovém případě lze použít index ULI zaměřený na předěl mezi určitými percentily, mezi něž padá dělící skóre.

Index ULI může teoreticky nabývat hodnot mezi −1 a 1, ale záporné hodnoty jsou příznakem velmi hrubé chyby v položce (nebo chyby v klíči) a v praxi jsou vzácné. ULI rovné jedné znamená, že všichni lepší studenti položku zvládnou, zatímco všichni horší nikoli. Byčkovský a Zvára (2007) uvádějí, že:

- pro položky s obtížností mezi 0,2 a 0,3, nebo obtížností mezi 0,7 a 0,8 by citlivost ULI měla být alespoň 0,15,
- v případě úloh s obtížností mezi 0,3 a 0,7 by rozlišovací schopnost ULI měla být aspoň 0,25.

Pokud je hodnota ULI nižší, je třeba úlohu považovat za podezřelou. V praxi se položky kolem uvedených hranic považují za sice nikoli ideální, ale tolerovatelné. Pokud je však hodnota ULI příliš nízká (ULI < 0,1), je třeba úlohu zkontrolovat, jestli je dobře zkonstruována a zda neobsahuje nějakou závažnou chybu. Pokud pracujeme s jemnějším dělením intervalu schopností (jako v případě ULI_{54}), může být hodnota indexu kolem 0,1 naprosto v pořádku. Nicméně jakmile je hodnota libovolně pojatého ULI blízká nule, nebo dokonce záporná, znamená to, že úloha nefunguje. Záporná hodnota ULI znamená, že horší studenti odpovídali lépe než lepší. V položce tedy může být něco, co lepší studenty zavede na nesprávnou stopu, např. v úloze hledají chyták. Záporným ULI se také projeví chyba v klíči, podle kterého se úloha boduje. Takovou položku je třeba buď opravit, nebo z testování rovnou vyřadit. Zajímavý problém představuje metodika rozdělení intervalu schopností na menší díly. Může se stát, že interval nejde „strojově“ rozdělit úplně ideálně, např. proto, že na pomezí mezi



Obr. 6.4.2 Index ULI_{54} – rozdíl v pravděpodobnosti správné odpovědi na položku mezi pětinou nejlepších a pětinou dalších studentů

skupinami je velká skupina se stejnými výsledky. V praxi se ukazuje, že pro představu o citlivosti položky je způsob dělení na hraně intervalu málo významný. I když spornou skupinu na hraně intervalů rozdělíte arbitrárně, výsledné ULI dává většinou velmi dobrou představu o chování položky.

Některé práce používají jiné dělení intervalu schopností. Zkoušející například rozdělí studenty na tři skupiny podle výsledků v testu. Často se používá rozdělení testovaných na „horní třetinu" a „spodní třetinu", ale studie ukázaly, že když jsou studenti rozděleni na skupiny, které mají v „horní" a „dolní" skupině po 27 % studentů, hodnota diskriminace se zvyšuje (Swerdlik et al. 2012). Je zřejmé, že 46 % procent studentů se středním skóre v testu se při výpočtu indexu diskriminace neprojeví. Této praxe se přidržuje např. i testovací systém Rogō, který počítá ULI na základě dolních a horních 27,5 % studentů.

6.4.3 Vizualizace výsledků položkové analýzy

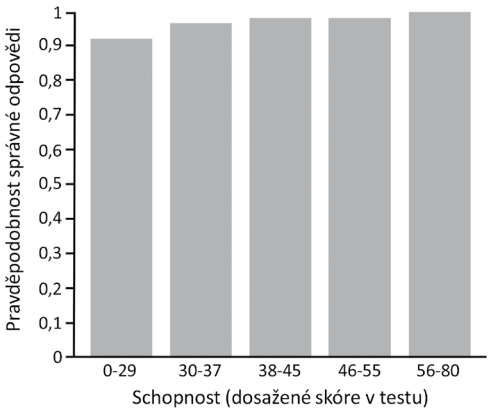
6.4.4 Příklady položek a grafické vyjádření jejich vlastností

Ukažme si několik příkladů chování položek použitých v přijímacích testech a jejich grafické vyjádření.

Příklad první:

Člověk slyší zvuk v rozsahu:
a) 16 až 20 000 Hz,
b) do 100 000 Hz,
c) méně než 16 Hz,
d) více než 20 000 Hz.

Při recenzi položky bychom mohli diskutovat řadu chyb, které položka vykazuje. Například navržené distraktory c) a d) nemají povahu „rozsahu", o němž se mluví v zadání. Podívejme se však, jak dopadlo použití této položky v reálném testu.

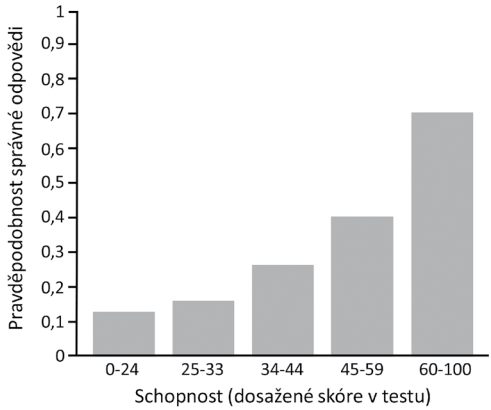


Obr. 6.4.3 Vizualizace chování položky. Položka „Člověk slyší zvuk v rozsahu ...“, je tak snadná, že prakticky nerozlišuje mezi různě schopnými studenty.

Studenti byli rozděleni na pětiny podle celkového výsledku v testu. Pro tyto pětiny byly spočteny pravděpodobnosti správné odpovědi. Vidíme, že již nejslabší studenti dosáhli v této položce více než 90% úspěch. Studenti v lepších skupinách již téměř 100 %. Položka je tak snadná, že prakticky nerozlišuje mezi lepšími a horšími studenty.

Příklad druhý:

Energie fotonu je:
a) nepřímo úměrná frekvenci,
b) přímo úměrná vlnové délce,
c) přímo úměrná frekvenci,
d) nezávislá na vlnové délce.



Obr.6.4.4 Vizualizace chování položky. Položka „Energie fotonu je...“, je spíše obtížná. Pro nejslabší studenty je velmi obtížná a nerozlišuje mezi nimi, ale velmi dobře rozlišuje mezi studenty výtečnými a nejlepšími. Vidíme také značný rozdíl mezi nejslabšími a nejlepšími studenty. Položka může být v testu velmi užitečná.

Metodika je stejná jako v předchozím příkladu, opět jsme studenty rozdělili na pět stejně početných skupin podle jejich celkového výkonu v testu. Povšimněte si, že poslední pětina pokrývá rozsah 40 bodů ve stobodovém testu. Je tady patrné, že test jako celek byl poměrně těžký. Podobně se chová i tato konkrétní úloha. Její maximální rozlišovací schopnost je mezi čtvrtou a pátou pětinou. Položky, které rozlišují na „obtížném" konci spektra bývají poměrně ceněné a napsat je nebývá snadné. U této úlohy bylo takové chování překvapením, protože odborníci při recenzi odhadovali, že bude snadná.

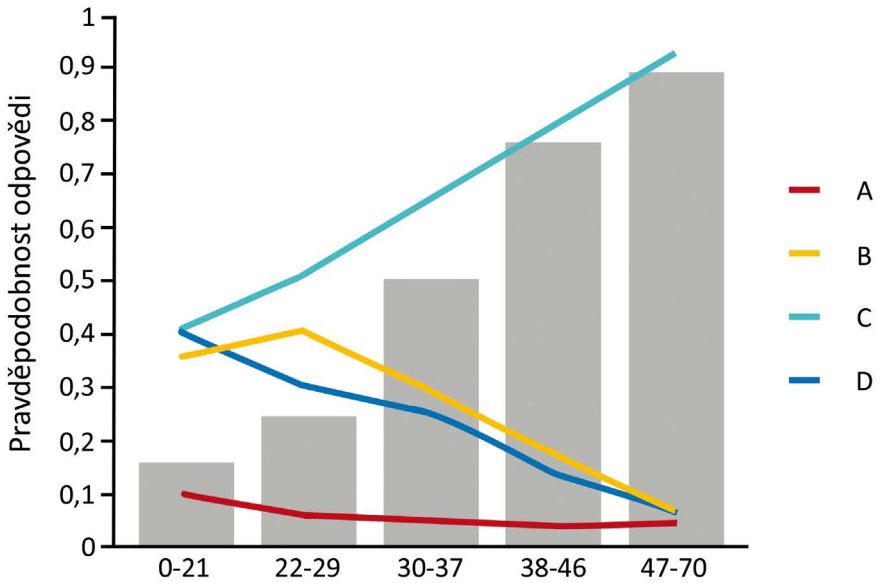
6.4.5 Analýza distraktorů

Jako analýza distraktorů se označuje rozbor, jak ke kvalitě výběrové úlohy přispívají nabízené možnosti – tedy správná odpověď (klíč) a především pak nesprávné možnosti (distraktory). Snažíme se zjistit, zda jsou distraktory pro studenty dostatečně atraktivní a jaký podíl z celkového počtu studentů si distraktory vybral.

Podívejme se vizualizaci analýzy distraktorů na konkrétním příkladu. Studenti byli v testu o 70 úlohách tázáni, jak vzniká metanol:

Jakou reakcí může vzniknout methanol?
a) Oxidací oxidu uhelnatého.
b) Oxidací methanalu.
c) Redukcí formaldehydu.
d) Oxidací methylaldehydu.

Jako správnou odpověď označili autoři testu možnost c).



Obr.6.4.5 Distraktorová analýza

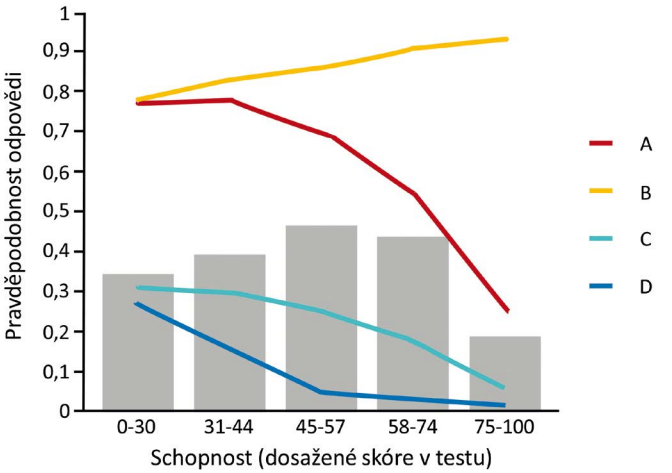
Studenti byli rozděleni do pěti skupin podle toho, kolik celkem získali bodů za celý test. Šedé sloupce uvádějí, jaký byl podíl správných odpovědí pro každou z těchto pěti skupin. Vidíme tedy, že nejslabší skupina studentů – šedý sloupec nejvíce vlevo – odpovídala správně mnohem méně často než nejlepší skupina podle celkového dosaženého skóre (sloupec zcela vpravo). Rozdíl výšky posledního a prvního šedého sloupce $ULI_{51} = 0,7$ ukazuje, že položka dobře rozlišuje mezi nejlepšími a nejhoršími studenty, přestože můžeme diskutovat, zda je opravdu dobře sestavená. I rozdíl výšky pátého a čtvrtého sloupce $ULI_{54} = 0,14$ je uspokojivý a ukazuje na dobrou diskriminaci mezi nejlepšími a druhými nejlepšími studenty. Položka jako celek tedy velmi dobře funguje.

Podívejme se nyní na fungování nabízených možností. Jejich chování popisují barevné lomené čáry, které pro každou skupinu (podobně úspěšných) studentů ukazují, s jakou pravděpodobností by si tito studenti nabízenou odpověď vybrali. Červená čára (distraktor A) je pro všechny skupiny studentů prakticky nepřijatelná volba. Jen v nejslabší skupině tuto možnost volí asi 12 % studentů, ale potom už prakticky nikdo další. Modrozelená čára správné odpovědi (klíč C) spojitě roste v celém intervalu schopností. To vypovídá o tom, že tato odpověď je správně vytvořena. V nejslabší skupině volí studenti odpověď C se stejnou pravděpodobností jako oba další distraktory, takže – vyjma neatraktivního distraktoru A – studenti nejslabší skupiny vlastně hádají. To je opět příznak dobře rozlišující položky. Zatímco distraktor D (tmavě modrá čára) monotónně klesá v celém intervalu schopností, což ukazuje na jeho správné fungování, distraktor B (žlutá čára) s rostoucí schopností studentů nejprve trochu roste a pak teprve začne klesat. Z nejlepších studentů jej nevolí prakticky nikdo. Nicméně to, že pokles není monotónní, znamená, že studenti druhé nejslabší skupiny o něm uvažují způsobem, který autor nepředpokládal. V této úloze autoři použili tři různé názvy pro tutéž látku – formaldehyd, methanal a methylaldehyd. První dva jsou poměrně běžné. Ve druhé nejslabší skupině bylo pravděpodobně hodně studentů, kteří sice věděli, že metanol lze vytvořit jednoduchou reakcí z formaldehydu čili methanalu, pak už ale jen tipovali, jestli je onou reakcí oxidace, nebo redukce.

Podívejme se nyní na analýzu distraktorů v případě nefunkční položky. Studenti byli v testu o 100 položkách tázáni na vzácné plyny:

> Vzácné plyny
> a) Jsou v přírodě málo zastoupené a netvoří téměř žádné sloučeniny.
> b) Alespoň jeden se využívá v lékařství.
> c) Jsou netečné, ale jinak normální plyny s dvouatomovou molekulou, jako má např. vodík.
> d) Jsou vždy těžší než vzduch.

Podíváme-li na výšku šedivých sloupců, vidíme, že na položku odpovídají nejhůře právě nejlepší studenti. Správnou odpověď má představovat současná volba nabízených odpovědí A) a B). Zatímco pravděpodobnost, že student zvolí možnost B), roste s jeho schopností, u možnosti A) tomu tak není. Tuto odpověď volí studenti ve dvou nejhorších skupinách, ale potom pravděpodobnost jejího výběru strmě klesá. Nabízená odpověď A) obsahuje zásadní problém, který tuto položku zcela znehodnocuje. Prozkoumáme-li jej, vidíme, že obsahuje chyb hned několik. Jde nikoli o jedno, ale o kombinaci dvou tvrzení: „Vzácné plyny jsou



Obr. 6.4.6 Distraktorová analýza

v přírodě málo zastoupené.“ a „Vzácné plyny netvoří téměř žádné sloučeniny.“ Problematické jsou relativizující termíny „málo“ a „téměř žádné“, které způsobují, že rozhodnutí, zda je možnost správná, záleží na čistě subjektivním pohledu. Ještě větší problém představuje vymezení „v přírodě“, protože autor měl patrně na mysli biosféru, zatímco nadaní studenti si pod „přírodou“ zřejmě představili spíše vesmír. A při tomto pohledu tato odpověď pravdivá není. Zbývající dva distraktory (c, d) fungují správně, ale to už položku nezachrání. Pokud už se stane, že takovou položku autor napíše, neměla by projít recenzí. Analýza distraktorů je pak posledním okamžikem, kdy díky objektivnímu pohledu můžeme napravit opomenutí autora a recenzentů a položku z testu vyřadit před jeho obodováním.

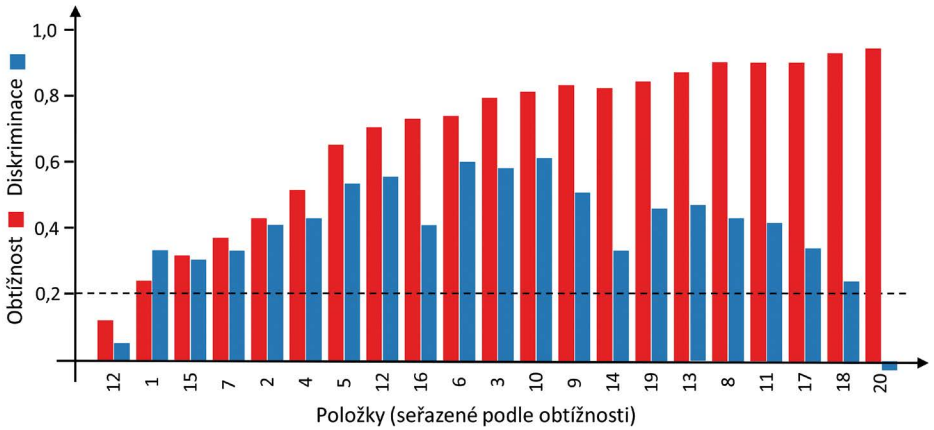
Aby se daly distraktorové analýzy dobře interpretovat, je třeba mít testová data nad dostatečně velkým souborem studentů. Zatímco samotná úspěšnost jednotlivých skupin v položce (šedé sloupce) je poměrně stabilní, protože se na ní projeví data všech studentů skupiny, jednotlivé distraktory již celá skupina nevolí, a jsou proto výrazně citlivější na ovlivnění náhodným „šumem“. Pokud si má zobrazení chodu distraktorů uchovat rozumnou vypovídací schopnost, musí být (při rozdělení na pět podskupin) v celé testované skupině více než několik set lidí. Pokud jsou počty menší, lze použít rozdělení na menší počet podskupin, v krajním případě jen na dvě (dva šedé sloupce). Tím sice přijdeme o jemnost detailního pohledu, ale výsledek bude méně ovlivněn náhodnými jevy.

Distraktor je považován za funkční (plausibilní), pokud si jej zvolí nejméně 5 % z testované skupiny. Navrhnout dostatečně atraktivní distraktory může být poměrně obtížné, mimo jiné proto, že učitel již si nemusí být schopen představit, co je pro studenty obtížné a co ne. Při tvorbě nových položek může učitel pro návrh distraktorů použít předchozí, nejlépe formativní testování, v němž studentům předloží podobnou položku jako úlohu s krátkou tvořenou odpovědí. Distraktory pro výběrovou úlohu pak vytvoří podle nesprávných odpovědí.

6.4.6 Grafický náhled na výsledky celého testu

Dvoubarevný graf

Pro rychlou orientaci, jak se podařilo sestavit test, můžeme v položkové analýze s výhodou požít dvoubarevný graf. V literatuře bývá též nazýván jako *difficulty-discrimination plot*, nebo zkráceně „DD-plot“. Na vodorovné ose jsou položky seřazené podle obtížnosti, od nejlehčích k nejtěžším. U každé úlohy je červeným sloupečkem vynesena její obtížnost a modrým její citlivost. Na tomto grafu na první pohled rozpoznáme „podivně“ se chovající položky, jejichž citlivost je malá, nebo dokonce záporná, a můžeme se zabývat jejich podrobnější analýzou, abychom zjistili příčiny anomálií.



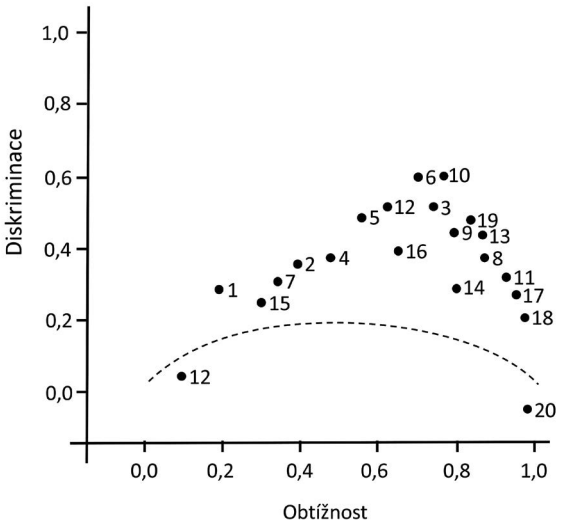
Obr. 6.4.7 **Dvoubarevný graf** (*difficulty-discrimination plot*, **DD-plot**) ukazuje testové položky seřazené podle obtížnosti (výška červeného sloupce). U každé položky je vynesena její diskriminace (modré sloupce). Vodorovná přerušovaná linka ukazuje mez (20 %), pod niž by neměla diskriminace fungující položky klesnout. Položka č. 12 je velmi snadná a její diskriminační schopnost je velmi malá. Položka č. 20 je velmi obtížná a její diskriminační schopnost je velmi malá a navíc záporná, tj. lepší studenti odpovídají hůře než slabší. Tato úloha pravděpodobně obsahuje nějaký další problém, kterého si autor nebyl vědom. Tuto položku je nutno z testu vyřadit.

Jinou, možná ještě názornější podobu grafu dostaneme, pokud vyneseme diskriminaci položek (na svislé ose) v závislosti na jejich obtížnosti (na vodorovné ose).

6.4.7 Indexy Rit a Rir

Pro posouzení citlivosti položky je rovněž možno použít korelační koeficient mezi bodovým ziskem za položku a bodovým ziskem za celý test, který se označuje Rit (*correlation item-test*), případně korelační koeficient mezi položkou a zbytkem testu, Rir (*correlation item-rest*).

Koeficient Rit se počítá jako bodově biseriální korelační koeficient mezi skóre položky a celkovým skóre z testu. Říká nám, do jaké míry daná položka přispívá k výběru správně



Obr. 6.4.8 Pro stejný test o 20 položkách je vynesena diskriminace položek (na svislé ose) proti obtížnosti položek (na vodorovné ose). Test je spíš obtížný a diskriminace úloh v něm spíše podprůměrná. Obě podezřelé položky (č. 12 a č. 20) se v tomto zobrazení jasně vydělují.

odpovídajících respondentů ze všech účastníků testu. Jinými slovy, reflektuje rozlišovací schopnost položky a vypovídá o výkonu položky oproti testu jako celku. Kladné hodnoty blízké 1 znamenají, že studenti úspěšní při řešení dané položky byli rovněž úspěšní při řešení celého testu. Záporné hodnoty ukazují, že studenti, kteří správně vyřešili danou testovou úlohu, dosáhli spíše nízkého celkového skóre ve zbytku testu. Korelace ukazuje, zda položka měří stejný konstrukt jako zbytek testu. Pokud je test zaměřen na více témat, je třeba to brát při intepretaci tohoto koeficientu v úvahu. Hodnota Rir je podobná Rit, ale přesnější, protože se nebere v úvahu příspěvek ke korelaci od samotné položky. Rir je vždy o něco nižší než Rit.

Pro číselné hodnoty korelačního koeficientu Rit existují doporučení podobně jako pro index ULI:

- Vyhněte se otázkám s hodnotou Rit pod 0,20.
- Vždy se dívejte na Rit v kombinaci s obtížností P.

Přestože posuzování diskriminace pomocí ULI je běžnější, například CERMAT používá při analýze úloh v testech velkého významu právě Rit (CERMAT 2018).

6.5 Klasická testová teorie

Klasická testová teorie (CTT) je v psychometrii nejpoužívanější, je nejstarší a asi nejsnáze pochopitelná. Vychází z ústředního předpokladu, že pozorované skóre (O) je kombinací takzvaného skutečného skóre (T) a skóre chyby (e):

$$O = T + e$$

To samozřejmě není jediný předpoklad, který je pro použití CTT nutný, dalším důležitým předpokladem je lokální nezávislost jednotlivých pozorování, tj. že všechna měření jsou na sobě navzájem nezávislá. Skutečné skóre je hypotetické skóre, které by student získal pouze na základě své kompetence. Ale protože každý test má chybu měření, pozorované skóre nemusí být nutně stejné jako skutečné skóre.

Model skutečného skóre a jeho předpoklady vedly ke zkoumání statistických vlastností testovaných položek, které by mohly zlepšit spolehlivost testu. Byly identifikovány tři důležité charakteristiky položek:

1. obtížnost položky – podíl správně odpovídajících testovaných,
2. citlivost položky – rozdíl v obtížnosti položky pro (v testu) dobré a špatné studenty,
3. analýza distraktorů – analýza podílu nesprávně zvolených odpovědí u výběrových položek.

Ukázalo se, že nejspolehlivější testy byly složeny z položek, které měly obtížnost kolem 0,5, citlivost větší než 0,3 a s distraktory zvolenými tak, že je volilo rozumné procento studentů (Ryan a Brockmann 2009).

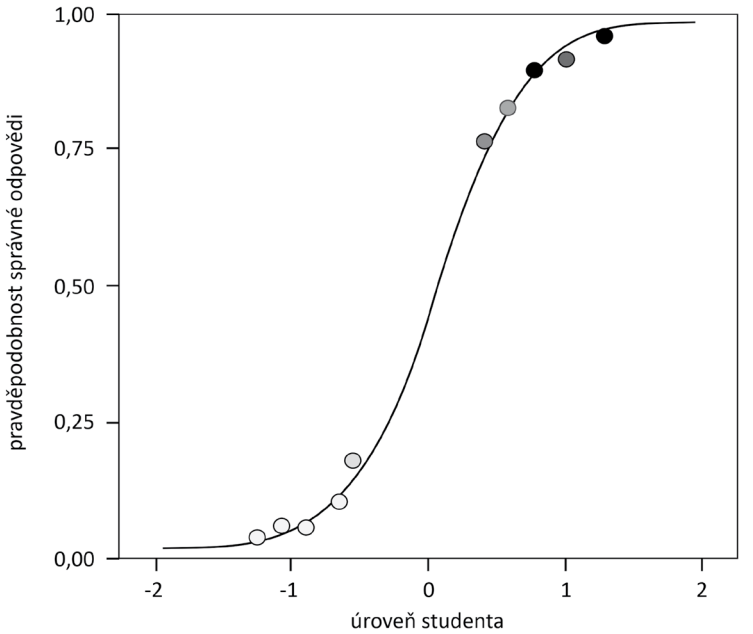
Klasická testová teorie (CTT) má řadu výhod. Je jednoduchá, srozumitelná a široce přijímaná. Po dlouhou dobu byla CTT hlavním nástrojem zkoumání testů a dodnes je hojně používá, zejména při položkové analýze. Umožňuje analýzu testů a položek i při malém počtu testovaných, což je její hlavní výhodou proti teorii odpovědi na položku (IRT).

Naopak hlavní nevýhodou CTT je závislost charakteristik položky na testované skupině a naopak, měřená schopnost testovaných závisí na konkrétním testu. Testovaný se může jevit jako dobře připravený, pokud je test snadný, a naopak. Jak ale rozlišíme mezi vlivem obtížnosti konkrétního testu a připraveností studenta? Už sama definice obtížnosti položky ukazuje provázání mezi touto charakteristikou a zkoumanou skupinou. Zda je úkol těžký, nebo snadný, závisí na schopnostech zkoumaných účastníků, a současně výsledek měření schopností účastníků testu závisí na tom, zda jsou položky těžké, nebo snadné (Hambleton et al. 1991).

Povšimněte si, že v rámci samotné CTT by nešlo vytvářet kvalitní položkové banky, protože parametry položky závisí na konkrétní testované skupině. Prakticky by také nešlo vytvářet paralelní formy testů (Hambleton et al. 1991).

6.6 Teorie odpovědi na položku

Klasická testová teorie dává dobré výsledky, pokud mají testovaní srovnatelnou úroveň znalostí a schopností. Představme si, že tomu tak v nějakém konkrétním případě není. Například, že skupina testovaných je složena z těch, kdo už absolvovali řidičský kurz, a těch kdo, s ním právě začínají. Pokud jim předložíte stejnou otázku o přednosti vozidel na křižovatce, může být tato otázka pro jedny lehká a pro druhé obtížná. Vidíme tedy, že s konceptem obtížnosti úlohy postaveným na klasické testové teorii v tomto případě nevystačíme. Řešením by bylo skupinu rozdělit a měřit obtížnost úlohy na opět na homogenních podskupinách. Dostali bychom tak dvě různé hodnoty obtížnosti, odpovídající dvěma úrovním znalostí.



Obr. 6.6.1 Pravděpodobnost správné odpovědi v závislosti na úrovni znalostí studenta, (odvození IRT)

Pokud bychom skupinu členili podrobněji, např. podle délky školení, mohli bychom nakonec získat (téměř) spojitou informaci o obtížnosti zkoumané položky. Tato spojitá křivka popisuje chování položky pro různé úrovně znalostí a dovedností studentů a nazývá **charakteristická funkce položky** (*item characteristic function*, ICF).

Slabší studenty budeme tedy hledat v levé části křivky (světlá kolečka v našem grafu) a lepší studenty v pravé části (tmavá kolečka). Na tomto konceptu je založena celá teorie odpovědi na položku (Item Response Theory – **IRT**).

6.6.1 Vlastnosti IRT modelů

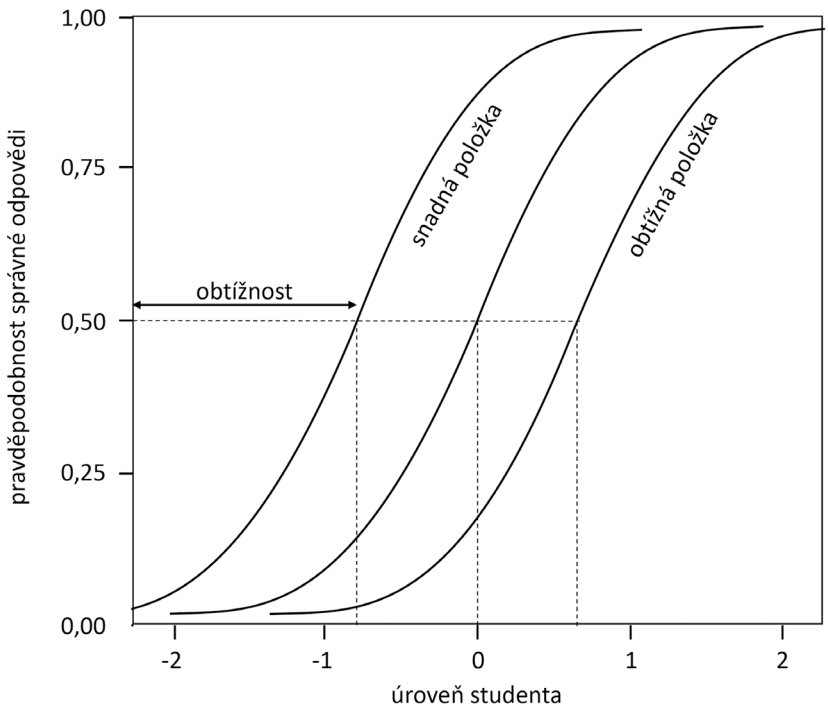
Předpokládejme, že z proběhlých testů známe pravděpodobnosti správných odpovědí pro různé úrovně studentů. Pokud máme takových měření dost, mohli bychom se jimi pokusit proložit křivku a odhadovat pravděpodobnost úspěchu pro další možné testované. Proložená **charakteristická funkce** položky má většinou typický esovitý tvar, který se dá matematicky popsat jako logistická funkce. Esovitý tvar je společný i pro jiné charakteristické funkce (mimo oblast psychometrie), např. funkce zčernání fotografické emulze v závislosti na osvícení a další. Esovitost charakteristické křivky vyjadřuje skutečnost, že převod mezi podnětem a reakcí je efektivní jen v omezeném rozsahu podnětů. Představme si, že skupině různě starých jedinců předložíme test rozpoznávání tvarů. Pro předškoláky bude asi příliš obtížný a pro maturanty příliš snadný. Plochost charakteristické křivky pro okrajové hodnoty úrovně schopností odpovídá tomu, že v těchto skupinách test nebude dobře rozlišovat schopnější od méně schopných.

Charakteristická funkce popisující chování položky stojí v základu řady matematických modelů, které se snaží popsat, jak vyšetřovaní reagují na položky. Proto se tomuto přístupu říká **teorie odpovědi (odezvy) na položku** (*item response theory*, **IRT**).

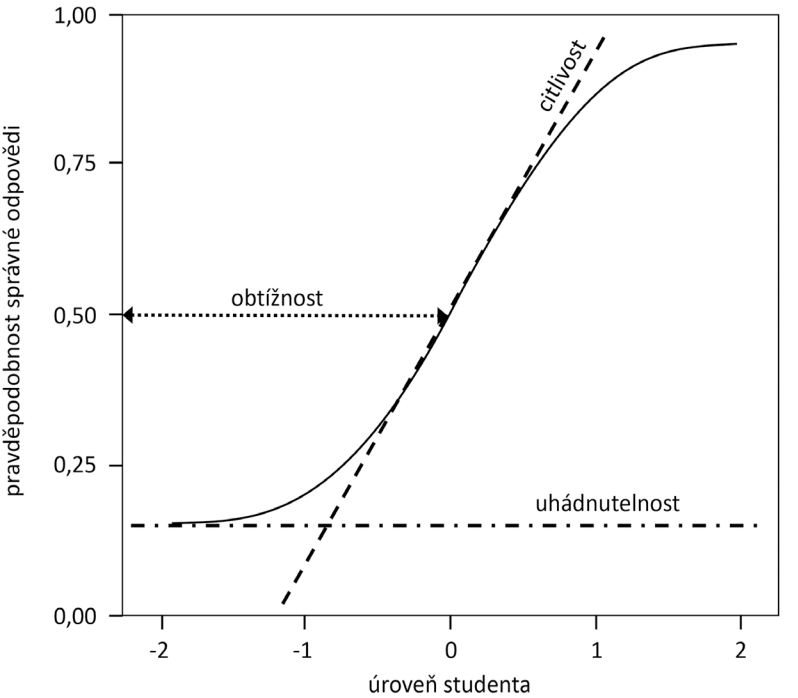
IRT, nebo též teorie latentních vlastností, je psychometrická teorie, která byla vytvořena s cílem lépe porozumět tomu, jak jednotlivci reagují na jednotlivé položky v psychologických a vzdělávacích testech. Pojem latentní znak se používá v IRT proto, že charakteristiky jednotlivců nelze přímo pozorovat; musí být odvozeny pomocí určitých předpokladů o procesu reakce, které pomáhají odhadnout tyto parametry. Parametr θ na vodorovné ose IRT grafu reprezentuje úroveň latentního rysu jedince, kterým může být lidská schopnost nebo vlastnost měřená v testu. Tou může být kognitivní schopnost, fyzická schopnost, dovednost, znalost, postoj atd.

Teorie odezvy na položky překonává klasickou teorii testů (CTT) v řadě aspektů. Poskytuje efektivnější popis toho, jak položky skutečně fungují, odstraňuje problém se závislostí vlastností položek na vzorku testovaných, dovoluje vytvářet testy se srovnatelnými vlastnostmi a vyrovnávat různé verze (běhy) testu, umožňuje odhadnout vliv hádání odpovědi a umožňuje využít detailní znalost vlastností položek pro adaptivní testování.

Nejjednodušší IRT model počítá s jednou proměnnou – obtížností. Různě obtížné položky jsou reprezentovány charakteristickými křivkami stejného tvaru, jen posunutými vlevo (pro lehčí položky), nebo vpravo (pro těžší položky) (Tavakol a Dennick 2013).

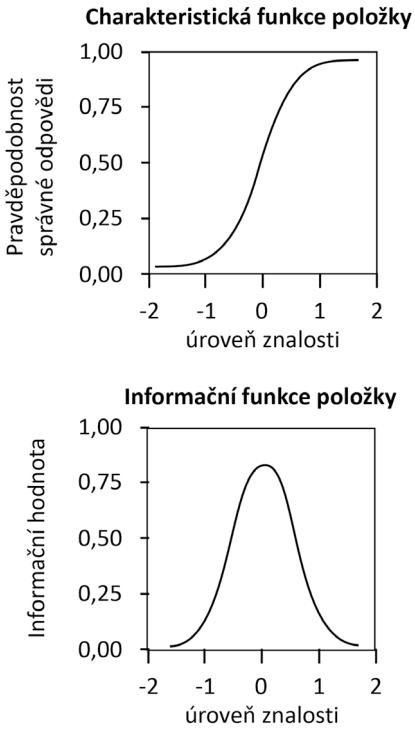


Obr. 6.6.2 Charakteristické křivky různě obtížných položek v jednoparametrickém IRT modelu



Obr. 6.6.3 Schematické znázornění parametrů v tříparametrickém modelu IRT. *Obtížnost položky souvisí s polohou charakteristické křivky (respektive vzdáleností od svislé osy), citlivost souvisí se sklonem charakteristické křivky v daném bodě (představte si, že bychom u koeficientu ULI zjemnili dělení na vodorovné ose nade všechny meze, pak jako měřítko citlivosti rovněž dostaneme sklon v daném bodě.) Třetím parametrem je uhádnutelnost položky, která je v grafu reprezentována čerchovanou vodorovnou čárou (asymptotou). Pokud máme položku s výběrem z šesti nabízených odpovědí, pak i zcela neznalý student má šanci 0,17, že správnou odpověď uhádne.)*

Jednoparametrický IRT model se někdy též označuje jako **Raschův model**. Je to trochu zjednodušení, protože, ač jsou si oba modely vnějškově velmi podobné, vycházejí z jiných předpokladů a přístupů. IRT má více deskriptivní povahu, protože si klade za cíl přizpůsobit model datům. V porovnání s tím, Raschův model klade důraz na zapadnutí dat do modelu. Co se tím myslí? Jedním z předpokladů Raschova modelu je „jednorozměrnost" testu, tedy že test měří jen jeden základní konstrukt. Pokud položka měří jiný konstrukt, musí být z testu vyloučena. Součástí práce s Raschovým modelem je proto identifikace nadbytečných rozměrů testu a eliminace položek, které jejich vznik způsobují. Dalším předpokladem je nezávislost položek. Tedy, že pravděpodobnost správné odpovědi na jednu položku by měla být nezávislá na odpovědi na ostatní položky. Předpoklad nezávislosti není naplněn, pokud mají položky vysokou pozitivní korelaci. Pro dodržení nezávislosti položek by měla být vždy jedna ze vzájemně závislých položek z testu vynechána. V tomto smyslu se tedy „upravují" data, aby odpovídala modelu. S daty se dále pracuje podobně jako v IRT analýze. Zájemce o toto téma odkazujeme na rozsáhlou literaturu (Tavakol a Dennick 2013; Stemler a Naples 2021; Kean et al. 2018; Boone et al. 2017). V praxi je důležité, aby bylo vždy deklarováno, se kterým modelem se pracuje, aby nemohlo dojít k nedorozumění.



Obr. 6.6.4 Souvislost informační funkce položky a charakteristické funkce položky

Realitu věrněji popisují komplexnější IRT modely, které kromě obtížnosti pracují i s citlivostí položky. Příkladem může být dvouparametrický logistický model. Zatímco obtížnost je, stejně jako v jednoparametrickém modelu, reprezentována polohou křivky, citlivost je reprezentována jejím sklonem. Dává to dobrý smysl, čím je charakteristická křivka strmější tím ostřeji bude test rozlišovat mezi podobně nadanými jedinci. Citlivost je jistě žádoucí vlastnost položky, ale snadno nahlédneme, že velmi citlivá položka bude fungovat jen v omezeném rozmezí úrovní schopností testovaných.

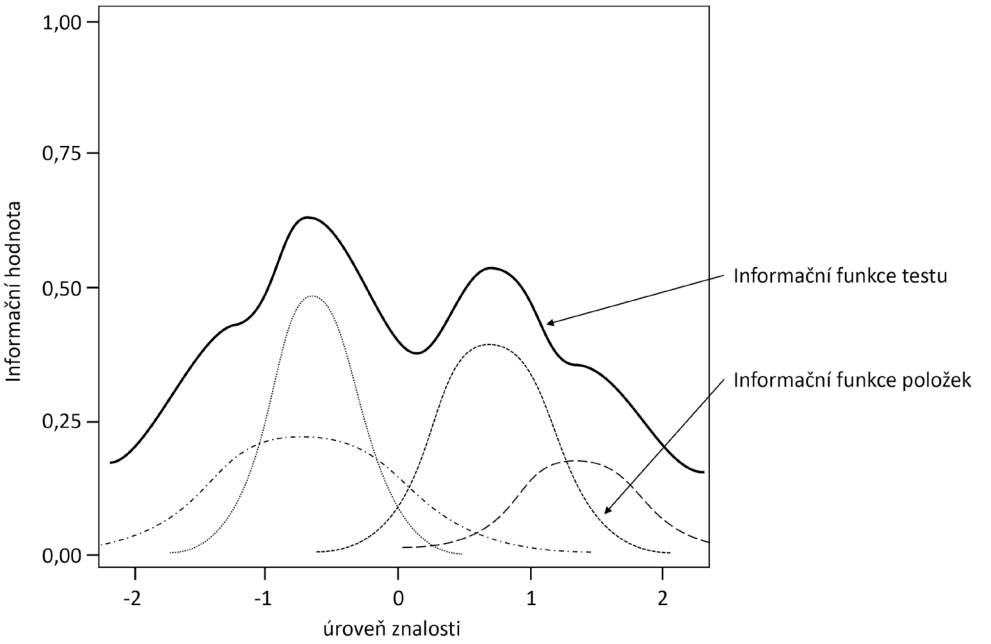
6.6.2 Informační funkce položky

Podíváme-li se na typickou charakteristickou funkci položky standardního esovitého tvaru, pak vidíme, že položka dobře rozlišuje jen v určitém okolí svého inflexního bodu, kde sklon charakteristické funkce zajišťuje, že se posun na ose latentní vlastnosti (úroveň znalosti) promítne do změny pravděpodobnosti správné odpovědi. S rostoucí vzdáleností je charakteristická křivka položky stále plošší a položka pro tyto hodnoty schopnosti testovaného přestává rozlišovat mezi lepšími a horšími účastníky testu.

Informace, kterou z použití položky můžeme vytěžit, je nejvyšší v okolí inflexního bodu charakteristické funkce a pak rychle klesá. Funkce popisující informační přínos položky má zvonovitý tvar, nazývá se **informační funkce položky** a můžeme ji získat derivováním charakteristické funkce položky.

6.6.3 Informační funkce testu

Informační funkci celého testu získáme jako součet informačních funkcí jednotlivých položek (předpokládáme, že odpovědi na položky jsou na sobě pro konkrétní hodnotu latentní schopnosti nezávislé).
Z tvaru charakteristické funkce položky se odvíjí tvar informační funkce položky. Vysoce rozlišující položky (se strmou charakteristickou křivkou) mají vysokou a úzkou informační křivku. Taková položka má vysokou informační hodnotu, ale jen v úzkém rozsahu obtížnosti. Položky s plošší charakteristickou křivkou, a tedy nižší hodnotou informační funkce, mohou mít pro danou úroveň latentního parametru nižší rozlišovací schopnost, ale zase mohou být přínosem v širším rozsahu obtížností. Pokud známe informační funkce položek, můžeme při plánování testu sledovat pokrytí intervalu latentní schopnosti informační funkcí testu, aby nedocházelo k nadbytečné redundanci podobně fungujících položek a na druhé straně, aby byl pokryt celý interval schopností, který nás zajímá.



Obr. 6.6.5 Schéma ilustrující, jak se informační funkce testu skládá z informačních funkcí jednotlivých položek. Přerušované křivky představují informační funkce položek. Plnou čarou nad nimi je znázorněna informační funkce celého testu.

6.6.4 Software pro výpočet IRT modelů

Zatímco odhady obtížnosti a citlivosti v rámci klasické testové teorie jsou výpočetně poměrně jednoduché, v případě IRT je situace nepoměrně složitější. Neznámou latentní schopnost studenta odhadujeme hledáním maxima funkce věrohodnosti, jak odhadované parametry

popisují chování položek. Tyto optimalizační procedury jednak vyžadují sofistikovaný softwarový nástroj a za druhé, aby byl odhad dostatečně robustní je zapotřebí velký počet testovaných. Nejméně stovky, lépe však tisíce. Čím přesnější (víceparamteričtější) model, tím větší je požadavek na počet testovaných.

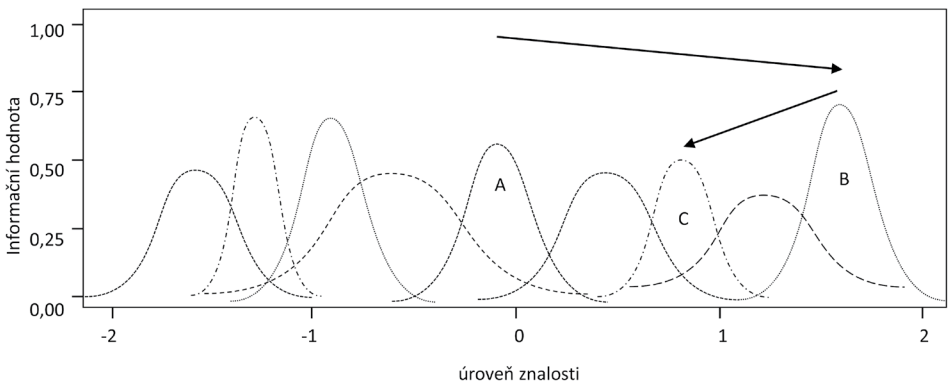
Vzhledem k tomu, že matematické modely IRT mohou být pro běžného smrtelníka poněkud nepřehledné, je pro další studium možno zvolit literaturu, která se drží v přijatelných mezích obtížnosti. Můžeme doporučit např. přehled literatury, který podává Hynek Cígler v časopise Testfórum (Cígler 2014).

Pro výpočet modelu v rámci teorie odpovědi na položku je k dispozici řada programů. Můžete zvolit pronájem některého komerčního programu, jako je například Stata (ve verzi 14 a vyšší), IRTPRO, nebo Xcalibre. Nebo naopak sáhnout po příslušných knihovnách v open source prostředí R. Komerční software je zpravidla uživatelsky přátelštější, ale dražší, zatímco prostředí R je zadarmo, ale předpokládá, že se naučíte základy prostředí a budete používat programové kódy z knihoven, nebo si s jejich využitím budete vytvářet kód vlastní, což může být časově dost náročné.

Na pomezí těchto dvou světů je zdarma dostupná webová aplikace ShinyItemAnalysis od Patricie Martínkové a jejích spolupracovníků, kterou používáme a můžeme doporučit. Je postavená na prostředí R, ale její rozhraní je „klikací“, takže se snadno používá (Martinková a Drabinová 2019, Martinková et al. 2017a).

6.7 Adaptivní testování – využití IRT v praxi

Při běžném testu obdrží účastník řadu položek, z nichž některé pro něj nemusí být zcela relevantní. Mohou být těžší, nebo lehčí, než je jeho úroveň. Informační funkce testových položek pokrývají interval úrovní obtížnosti, v němž se pohybují schopnosti většiny testovaných



Obr. 6.7.1 Princip adaptivního testování. Známe-li psychometrické vlastnosti položek, můžeme ke stejné přesnosti odhadu parametru „znalosti“ dospět s použitím menšího počtu položek. Představme si, že jako první zadáme testovanému položku průměrné obtížnosti (A). Testovaný odpoví správně a algoritmus adaptivního testování mu vybere těžší položku (B), tu testovaný nezodpoví dobře a algoritmus mu nabídne snazší položku (C). Místo aby respondent musel odpovídat všechny položky testu, postačí v tomto ilustračním příkladu zodpovězení tří položek k dostatečně přesnému určení úrovně respondenta.

jedinců. Nechtěným vedlejším jevem je, že každý účastník testu odpovídá na řadu úloh, které jsou pro něj moc snadné, nebo naopak moc obtížné. Přitom obojí je demotivující a z pohledu testující instituce jde o plýtvání časem. Při elektronickém testování si proto lze představit algoritmus, který bude testovanému vybírat položky, jejichž obtížnost bude přizpůsobována jeho výkonu při řešení předchozích úloh.

Tento přístup se nazývá „počítačové adaptivní testování" (*computer adaptive testing*, CAT). Umožňuje změřit latentní schopnost studenta se stejnou přesností jako klasický test, ale s použitím menšího počtu položek.

Adaptivní testování tedy přizpůsobuje test testovanému, položku po položce, na základě jeho odpovědí. Správná odpověď vede k obtížnější položce, zatímco nesprávná odpověď vede k jednodušší položce. Obtížnost položek se průběžně přizpůsobuje schopnostem testovaného. Nadaný student obdrží obtížnější položky, zatímco průměrný student obdrží položky snazší. Počet použitých položek souvisí s požadovanou přesností měření. To znamená, že test se zastaví, když je dosaženo předem stanovené požadované přesnosti psychometrických kritérií. U adaptivního testování je test jen tak dlouhý, jak je skutečně třeba.

Metoda je založena na teorii odpovědi na položku (*item response theory*, IRT), o které pojednávala předchozí kapitola.

6.7.1 Výhody a nevýhody počítačového adaptivního testování (CAT)

CAT je moderní způsob testování, který využívá algoritmy k optimálnímu přizpůsobení testu pro každého zkoumaného. V tradičním pojetí jsou položky sestavovány do testové sady a jsou předkládány studentům v této sadě. Nejviditelnější nevýhodou je tohoto přístupu je neefektivita. Obtížnost testových položek nijak nereflektuje schopnosti testovaného. Představme si mimořádně schopného studenta, který správně zodpoví všechny nejtěžší otázky. Můžeme mu s jistotou přiřadit vysoké skóre bez ztráty času na zodpovídání všech jednoduchých otázek. Zatímco u jednoho studenta se tato úspora může zdát ještě malá, uplatníte-li stejnou metodu na celou testovanou skupinu, jsou úspory času markantní.

Dalším problémem je nestejná přesnost měření pro studenty s různou úrovní znalostí. V tradičních testech bývá obvykle zastoupeno nejvíce položek se střední obtížností. To má dobrý důvod: mezi testovanými bude pravděpodobně velké množství lidí se střední úrovní schopností. Lidé s průměrnými schopnostmi budou testem velmi přesně vyhodnoceni. Stane se tak ale na úkor malé přesnosti měření u studentů s nízkou, nebo naopak vysokou úrovní schopností. Ti jsou hodnoceni s mnohem menší přesností. Ze stejného důvodu mohou mít studenti s nadprůměrnými, nebo podprůměrnými schopnostmi špatnou zkušenost s testem. Slabí studenti se mohou cítit vyčerpaní a odrazení tím, že většina položek je příliš obtížná, zatímco nadprůměrní studenti mohou být demotivováni tím, že většina položek je pro ně příliš snadná.

Výhody CAT:
- kratší testy (až o 50 %),
- stabilní přesnost,
- příznivá zpětná vazba testovaných,
- lepší motivace testovaných,

- menší expozice testových položek,
- možnost využití pro měření pokroku studenta (jeho test na konci bude jiný).

Nevýhody CAT:
- nemožnost se v průběhu testu vracet k dříve zodpovězeným položkám,
- citlivost na *testovou úzkost*,
- potřeba předchozí kalibrace položek,
- u položek s výhodnými vlastnostmi může jejich příliš časté využití způsobit jejich vynesení,
- vyžaduje dostatek pilotních testerů (několik set),
- příprava vyžaduje velmi kvalifikované odborníky,
- náročnější na vysvětlení veřejnosti – vyšší náklady na public relations.

6.7.2 Požadavky na počítačové adaptivní testování (CAT)

CAT mají mnoho výhod, včetně zkrácení doby testování na polovinu, ale vyžadují zkušené psychometriky, rozsáhlé pilotní vzorky a specializovaný software. Uveďme na tomto místě základní přehled toho, co je potřeba zvážit při rozhodování o adaptivním testování.

1. Položky musí být hodnotitelné automaticky, protože podle výsledku položky předešlé se v reálném čase volí položka následující. Tím jsou vyloučeny některé jinak užitečné formy testových položek (otázky s tvořenou odpovědí, esej apod.).
2. Je třeba mít zdroje na vývoj bank s velkým počtem položek. Obvykle potřebují banky alespoň třikrát více úloh, než je zamýšlená délka testu (i když to často není více, než je potřeba pro tradiční formy testu).
3. Musí proběhnout rozsáhlé pilotní testy. IRT vyžaduje, aby pro pilotní testování byl použit vzorek alespoň 100–1000 testovaných. Požadovaný počet závisí na složitosti použitého modelu IRT. Komplexnější modely IRT vyžadují větší vzorky.
4. Je třeba mít odborníky na psychometrii. Pro úspěšné nasazení jsou třeba kvalifikovaní odborníci zejména na kalibraci položek a IRT analýzu, případně i pro simulaci adaptivního testování s danou testovou sadou.
5. K dispozici musí být analytický software. Pro kalibraci položek je potřeba software pro analýzu IRT (např. volně dostupný ShinyItemAnalysis nebo komerční ekvivalenty).
6. Nezbytná je položková banka podporující IRT, schopná ukládat IRT parametry položek a podporovat navrhování CATs.
7. Konečně je potřeba mít vhodný systém pro doručování testu. Ten musí být schopný adaptivního testování na základě IRT, přinejmenším s příslušnými kritérii ukončení a algoritmy výběru položek.

6.8 Využití IRT pro analýzu férovosti

Férovostí (spravedlivostí, objektivností) testu myslíme jeho schopnost měřit studovaný rys nebo konstrukt se stejnou validitou ve všech podskupinách testované populace. Férovostí jsme se zabývali při recenzi testových úloh a zmiňovali jsme, že patří mezi důkazy validity. Ale protože se nám pro její ex-post analýzu (z dat proběhlého testu) budou hodit nástroje teorie odpovědi na položku, vracíme se zde k tomuto tématu znovu.

Položku označujme jako „diferencující“ (differential item functioning – DIF), když lidé se stejnou latentní schopností, ale z různých podskupin, mají různou pravděpodobnost správné odpovědi. Samotný rozdíl v průměrném výkonu mezi skupinami ještě nemusí být nutně neférový. Neférovost nastává pouze tehdy, pokud rozdíl v měřeném výkonu neodpovídá skutečnému rozdílu latentní vlastnosti, kterou má test měřit.

> Představme si například, že zkoumáme spravedlnost testu porozumění čtenému textu. Přitom zjistíme, že studenti se zrakovým postižením dosahují horších výsledků. Znamená to, že test je vůči těmto studentům nespravedlivý? To zatím nevíme. Je možné, že studenti se zrakovým postižením mají skutečně nižší čtenářské dovednosti než ostatní studenti.
>
> Předpokládejme nyní, že test vytiskneme znovu s výrazně větší velikostí písma a zjistíme, že průměrný výsledek postižených studentů stoupne na úroveň studentů bez postižení. To naznačuje, že test čtení v původní verzi s malým písmem byl pro postižené žáky nespravedlivý (neobjektivní). Výsledek také naznačuje, že test je spravedlivý, pokud je prezentován ve verzi s velkým písmem. Malá velikost písma vnášela do konstrukce testu systematickou chybu.

Rozlišujeme proto tzv. „benigní“ DIF, kdy rozdíl v pravděpodobnosti správné odpovědi souvisí s měřeným latentním znakem, a „nepříznivý“ DIF, kde se do výsledku promítají artefakty v procesu měření, nestejné možnosti přípravy, na jazykovém prostředí závislá interpretace textu a podobně. Neexistuje žádná jednoznačná kvantitativní metoda, která by dokázala rozlišit tyto dva případy od sebe. Pokaždé, když narazíte na diferenciální fungování položky, je třeba položku pozorně přezkoumat v týmu odborníků (Breslau et al. 2003).

Diferenciální funkce položky pohledem IRT

Pro zkoumání férovosti položek lze použít analýzu založenou na teorii odpovědi na položku vztaženou na zkoumané podskupiny testované populace. Teoreticky by měly být problematické položky vyřazeny nebo opraveny již při recenzi položek, kdy se ověřuje obsahová a konstruktová validita, ale ani pečlivá recenze nemusí zachytit vše. Analýza odpovědí testovaných studentů zahrnující chování položek vůči různým podskupinám testovaných může pomoci zachytit problematické položky a zlepšit kvalitu a férovost testu v dalších kolech.

V praxi se zkoumá, zda obtížnost položky není rozdílná pro vybrané podskupiny testovaných (např. absolventi gymnázií vs. absolventi ostatních středních škol), které mají jinak stejné schopnosti (měřeno např. celkovým skóre). Pro dané skupiny proložíme naměřenými body charakteristické křivky podle IRT teorie a porovnáme tyto křivky mezi sebou. Jako index popisující rozdílné fungování položky pro obě skupiny pak bereme plochu mezi oběma křivkami.

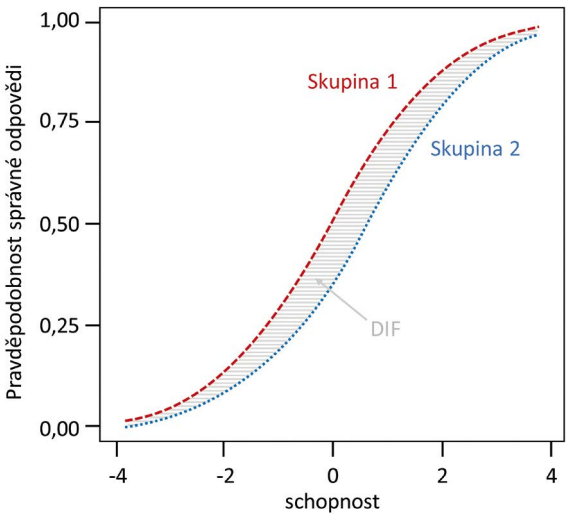
> Ve Spojených státech se v testech SAT se vyskytla otázka na verbální analogie:
> Najděte podobný vztah:
> běžec : maraton
> (a) vyslanec : velvyslanectví,
> (b) mučedník : masakr,

(c) veslař : regata,
(d) rozhodčí : turnaj,
(e) kůň : stáj.
Je snadné najít správnou odpověď („veslař“ a „regata“), pokud jste z prostředí, kde se pojmy „maraton“ a „regata“ používají. Při analýze testů se ukázalo, že na tuto otázku odpovídali prokazatelně hůře afroameričtí studenti (22 % správných odpovědí), než jejich bílí kolegové (53 % správných odpovědí), ačkoli v jiných otázkách tomu tak nebylo. Otázka předpokládala „samozřejmou“ znalost sportu rozšířeného jen mezi jednou ze subpopulací (Culbertson 1995).

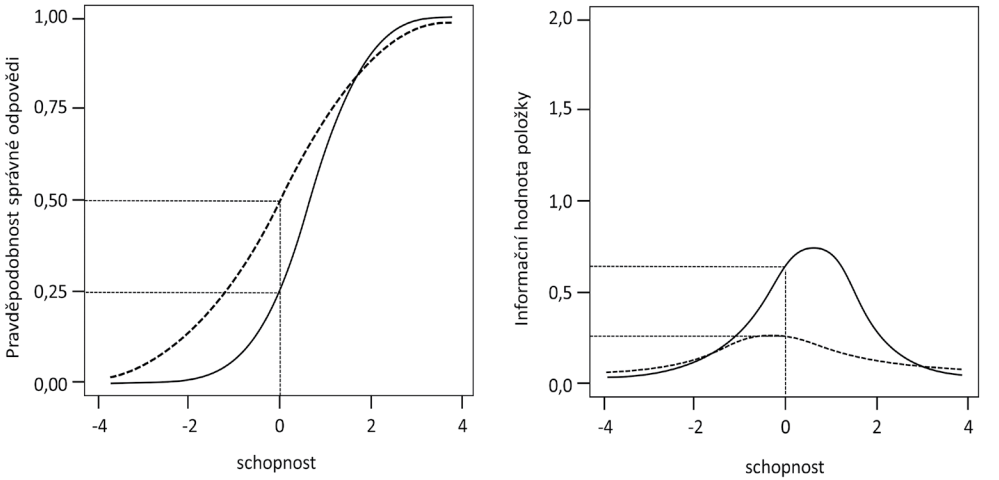
Využití IRT pro analýzu férovosti přináší detailní informace, které by byly základě recenze férovosti jen těžko odhadnutelné. Ukázalo se například, že v otázkách k přijímacím testům na lékařskou fakultu se vyskytují diferencující položky, na které odpovídaly ženy výrazně lépe než muži. Byly to zejména úlohy týkající se dětských nemocí.

Pro odhady férovosti je možné využít i další statistické metody, například vizualizaci pomocí grafického zobrazení proporcí správných odpovědí, či analýzu kontingenčních tabulek (metoda Mantel-Haenszel). Všechny uvedené nástroje najde zájemce v aplikaci ShinyItemAnalysis.

Problematika férovosti svým rozsahem překračuje rozsah tohoto textu. Případným zájemcům doporučujeme publikace, kurzy a nástroje, které se tématem zabývají hlouběji (Vlčková 2014; Cígler 2020; Martinková et al. 2017a).



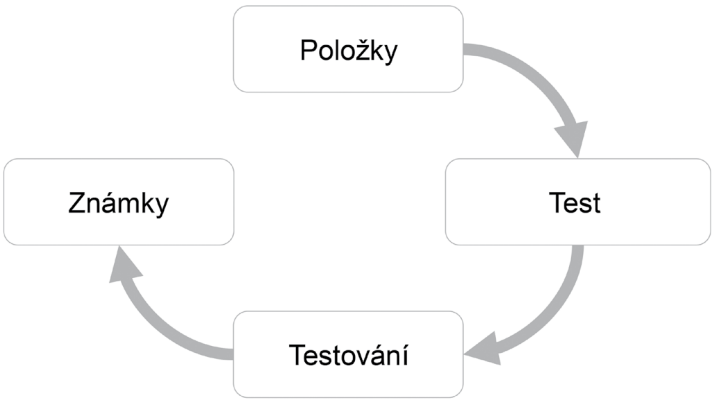
Obr. 6.8.1 Ilustrace nejjednoduššího případu neférového chování položky. Charakteristické křivky IRT pro dvě skupiny řešící stejnou, leč rozdílně fungující položku (viz příklad výše). Velikost plochy mezi křivkami odpovídá velikosti koeficientu DIF. Obě charakteristické křivky jsou stejně diskriminující, ale vykazují pro sledované skupiny odlišnou obtížnost. Případ, kdy neférová položka poskytuje v celém intervalu schopností výhodu jedné skupině studentů oproti druhé (jako zde), se označuje jako „uniform DIF“.



Obr. 6.8.2 Ilustrace nejednotného diferenciálního chování položky. Charakteristické křivky spočtené pro obě sledované podskupiny vykazují nejen různou obtížnost položky pro obě skupiny, ale i různou diskriminaci. Pro první skupinu (čárkovaná charakteristická křivka) je položka snazší na většině intervalu schopností, vyjma nejvyšších hodnot, kde se položka naopak stává snazší pro druhou skupinu (plná charakteristická křivka). Tento typ diferenciálního chování položky se označuje jako „non-uniform DIF“. Rozdílný tvar a hodnoty má pro obě skupiny i informační funkce této položky. Průběh křivek převzat z interaktivní tréninkové sekce webové aplikace ShinyItemAnalysis (Martinková et al. 2017a).

7 TESTOVÝ CYKLUS

Jako celá výuka je i testování cyklický proces. Při přípravě, provedení a vyhodnocení každého testu vytváříme (možná mimoděk) nejjednodušší základ testového cyklu.



Obr. 7.1 Schéma nejjednoduššího intuitivního testového cyklu

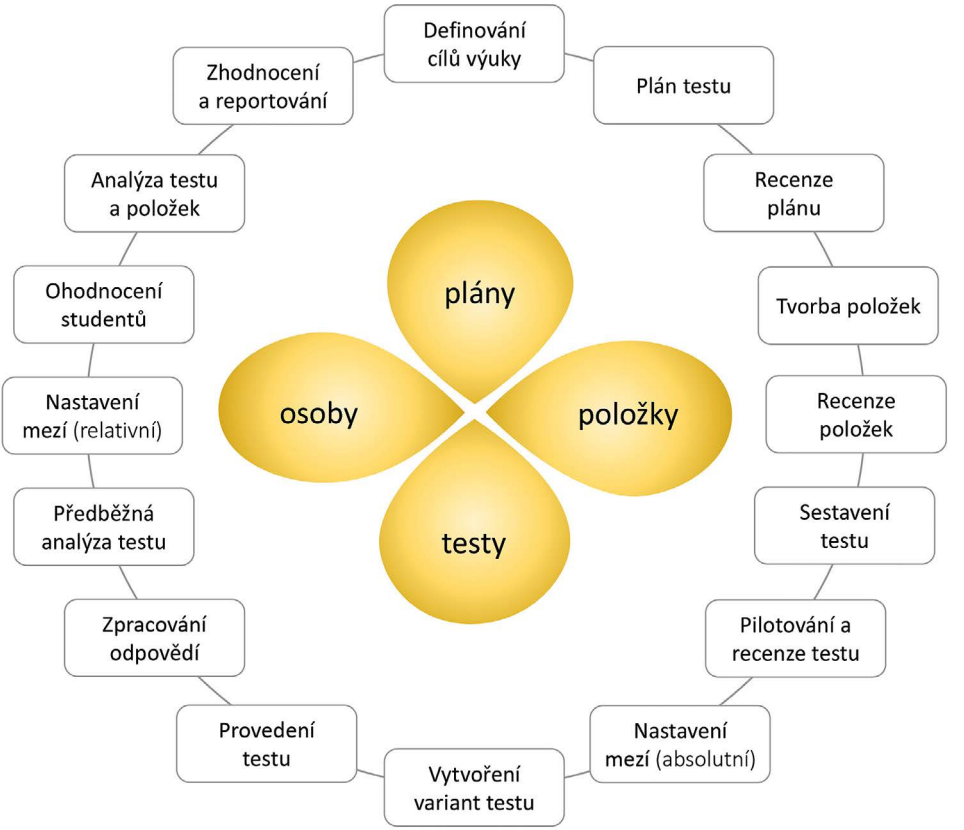
Jakmile začneme testy připravovat opakovaně a systematicky, začneme svoji zkušenost z předchozích běhů promítat do tvorby nových položek a testů a touto zpětnou vazbou vzniká první ucelený testový cyklus, na jehož konci jsme připraveni (lépe) pracovat na novém kole testu.

U testů velkého významu, které musí splňovat standardy validity a reliability, je třeba realizovat řadu kroků, které v intuitivní přípravě testu nejsou explicitně zdůrazněny. Typicky může testový cyklus u testu velkého významu vypadat následovně:

Osnova této knihy prakticky sleduje právě tento „velký“ testový cyklus. Pro většinu kroků existuje příslušná kapitola. Proto budeme v komentování jednotlivých kroků poměrně struční, neboť cílem je jen připomenout, o čem se v příslušné kapitole mluví.

Definování cílů výuky

Práce na testu by se měla odvíjet od ujasnění cílů. **Definováním cílů výuky** učitel vymezí rozsah učiva, které by měl student po absolvování kurzu umět. Učitel specifikuje klíčové kompetence, které je třeba otestovat.



Obr. 7.2 Schéma testového cyklu u testů velkého významu. Volně podle (Schuwirth a Vleuten 2011).

Plán testu

Návrh testu je dalším klíčovým bodem celého procesu. Je třeba stanovit, kolik úloh bude test obsahovat z každého tematického okruhu a jaké typy položek se použijí. Zvlášť významná tato fáze je, pokud se test připravuje ve více variantách, které mají být vzájemně srovnatelné. Cíle výuky se promítnou do výběru úloh a poměru zastoupení jednotlivých témat v připravovaném testu. Podle anglického pojmenování dříve užívaných modrých kopií stavebních plánů se tomuto *plánování testu* říká *blueprinting*, v češtině se nejčastěji používá termín *specifikační tabulka*.

Recenze plánu

Při přípravě důležitých hodnocení je třeba minimalizovat vliv individuálních preferencí jednotlivých pedagogů. Proto je plán testu potřeba nechat oponovat dalšími učiteli, aby se zastoupení témat a využití testových formátů zakládalo na konsenzu více vyučujících.

Tvorba položek

Možná nejnáročnější etapou testové přípravy je tvorba položek. Ve shodě s tématy určenými specifikační tabulkou a ve formátu, který jim specifikační tabulka předepisuje, navrhnou učitelé nové položky.

Recenze položek

Položky nesou zpravidla „rukopis“ autora. Proto je předložíme učitelům, kteří rovněž znají cílovou skupinu a probíranou látku. Při oponentuře otázek jsou položky předloženy k posouzení skupině odborníků (např. metodika přípravy testů programu Rogō doporučuje nejméně 5–9 osob, což je ovšem v našich podmínkách většinou nereálné), kteří podle připraveného formuláře procházejí testové úlohy a ověřují jednotlivé aspekty, které musí nová položka splňovat, a případně navrhnou nezbytné úpravy. Jednotlivé recenze potom musí projít zkušený autor položek, posoudit relevanci připomínek a upravit položky, pokud je třeba. Při supervizi recenzí může supervizor přidělovat recenzentům kredit za kvalitní recenze. Tím se recenzentům dostává zpětná vazba a současně vzniká informace o jejich výkonnosti a užitečnosti.

Sestavení testu

Autor testu vybírá z vytvořených položek tak, aby naplnil záměr blueprintu a současně dodržel ostatní (často nevyřčené) požadavky. Například aby udržel přiměřenou obtížnost testu, jeho časovou náročnost, aby počet výpočetních úloh nebyl vyšší než v jiných variantách, a podobně.

Pilotování a recenze testu

Má-li být test kvalitní, musí součástí jeho přípravy být i pilotní běh testu a recenze testu. Pro prověření chování položek i celého testu je vhodné test „pilotně“ vyzkoušet. Analýza výsledků *pilotního testu* může ukázat na (ne)schopnost položek rozlišovat studenty podle zvládnutí látky, ozřejmí jejich objektivní obtížnost atd. Pilotní testování je časově i organizačně náročné, proto se často jako pilotní testování používá až samotný první běh testu. Kromě pilotování necháme prověřit kvality testu ještě skupinou odborníků, kteří identifikují a odstraní poslední chyby, nejednoznačné či problematické formulace. I když se zdá, že díky všem předchozím kontrolám už v testu žádné problémy být nemohou, vždycky se tam najdou!

Nastavení mezí

Důležitým krokem je nastavení mezí pro průchod testem. Pokud je test vztažen ke kritériím, které musí účastník splnit, aby prošel, nazývá se tento krok *absolutní standardizace* a jeho čas je právě v tomto místě testového cyklu. Nastavení kritérií pro průchod testem předem dává účastníkům testu jistotu, že mez nebude nastavena účelově, aby některý konkrétní účastník ještě prošel. Pro učitele je objektivní nastavení meze prostředkem zajišťujícím, že testem projdou pouze dostatečně kompetentní studenti. Možností, jak najít „meze“ je víc, připomeňme např. zlaté standardy – Angoffovu a Ebelovu metodu.

Realizace testu

Jak jsme už uvedli, může mít písemný test podobu *papírovou*, nebo *počítačovou*. V obou případech je třeba zajistit vytvoření „testových verzí“, distribuci testů studentům a sběr jejich odpovědí. U testování, jehož výsledky mají významný dopad, musíme navíc zajistit rovné podmínky pro účastníky, například dozor během testu a další.

Zpracování odpovědí

Tento krok v testovém cyklu se týká především případu papírového testování, kdy sebrané odpovědní formuláře musí projít optickým čtením zaškrtnutých odpovědí.

Předběžná analýza testu

U testů velkého významu je žádoucí ještě před vyhodnocením výsledků test předběžně analyzovat. Z podezřelého chování položek můžeme poznat hrubé chyby typu špatně zapsaného klíče položky nebo chyby ve znění. Pro takové položky se ještě před vyhodnocením testu upraví klíč, aby položka fungovala, nebo se položka ze sčítání výsledků vyřadí (všichni účastníci za ni obdrží bod). Předejde se tím situaci, kdy by si po vyhlášení výsledků někdo stěžoval na chybnou úlohu a muselo by se dodatečně přepočítávat např. pořadí přijatých.

Klasifikace studentů

Oznámkování studentů je nejvýznamnějším výstupem testu. Při klasifikaci je možné porovnat počty bodů (celkové skóre) dosažené jednotlivými studenty a zjistit tak jejich relativní umístění. Pomocí expertního odhadu (např. Ebelovu nebo Angoffovou metodou) stanovíme hranici pro rozhodnutí „prošel“, nebo „neprošel“ (tzv. *absolutní standardizace*) a rozdělením intervalu úspěšnosti na potřebný počet dílů můžeme stanovit **klasifikaci studentů** v podobě klasifikačních stupňů – známek. K zajištění rovných podmínek účastníků přispívá anonymizace testů před úplným vyhodnocením (oznámkováním) testů.

Nastavení relativních mezí

Pokud je test zaměřen na porovnávání výkonu mezi studenty, pak nastavení mezí pro průchod tímto testem nejde udělat dříve než v tomto okamžiku, kdy jsou známy výsledky a jsou vyřešeny případné problémy odhalené rychlou analýzou. Účastníci testu jsou seřazeni podle dosaženého skóre v testu a dělící linie je nastavena buď podle některé metody relativní standardizace, nebo arbitrárně v případě, že cílem testu bylo vybrat vhodné adepty na omezený počet míst.

Ohodnocení studentů

Výsledky studentů jsou v tomto kroku převedeny na klasifikační hodnocení a poskytnuty studentům.

Analýza výsledků testu

Po testovém kole jsou k dispozici data, pomocí kterých je možné podrobněji zkoumat, jak se test ve skutečnosti choval. Zatímco v rychlé analýze šlo jen o identifikaci a eliminaci případných problémů, v testové analýze zkoumáme charakteristiky položek a testu. Díky tomu pak můžeme poskytnout zpětnou vazbu autorům a recenzentům, nakolik se při své práci „trefili do černého“. Test je měřící nástroj a jako každý nástroj má své vlastnosti, které je důležité znát. Optimální je mít možnost odhadnout kvalitu testu ještě *před* jeho ostrým nasazením, tj. již v rámci pilotního testování. Vlastnosti testu je poté potřeba ověřit na cílové skupině při ostrém nasazení. Při opakovaném použití testu je užitečné porovnávat výsledky v jednotlivých bězích testu.

Zhodnocení a reportování

Výsledky analýzy testu jsou reportovány jako zpětná vazba, jak autorům a recenzentům, tak i příslušným zodpovědným osobám v hierarchii instituce. Vzhledem k tomu, že reportování je u testů s významným dopadem běžnou agendou, obsahuje řada programů pro testovou analýzu i nástroje, které report nebo jeho části připraví. (Iteman, Xcalibre, Remark Office, ...).

8 POLOŽKOVÁ BANKA

Položková banka (nebo též „banka testových úloh“) je úložiště testů a testových úloh a jejich metadat.

Banky testových úloh úzce souvisejí s testovým cyklem popsaným výše, protože položková banka v širším pojetí může obsahovat nástroje pro vytváření položek, jejich recenzování a správu, dále pro plánování, vytváření a doručování testů, a nakonec i pro zhodnocení odpovědí, včetně analýzy testů a testových položek. Mohou tedy podporovat celý testový cyklus.

Spolu s testy a úlohami se uchovávají i metadata, včetně psychometrických charakteristik položek z předchozích běhů testů, jichž lze využít jako zpětné vazby pro zlepšování dalších běhů testu. Pokud systém udržuje i informaci o autorovi položky a recenzentech, může jim poskytnout zpětnou vazbu, jak se položka chovala v testu, a tím přispívat k jejich edukaci.

Položky mají většinou formát „s výběrem možností“, ale lze použít jakýkoli formát. Úlohy z banky slouží ke konstrukci testů distribuovaných jak klasicky, tak elektronicky.

David Vale popsal položkovou banku jako „organizovanou sbírku testových položek“ (Downing a Haladyna 2006b). Nejjednodušší položkovou bankou může být krabice od bot s lístky, na nichž jsou testové úlohy. Když byly k dispozici statistiky položek, psaly se často na zadní stranu karet, spolu s datem, kdy test proběhl. Pokud chtěl autor testů vytvořit nový test, prohledal ručně krabici s položkami, zkontroloval obsah úloh a jejich statistiky a vybral ty, které pak použil v testu (Weiss 2011).

Rozdíl, mezi položkovou bankou a prostou sadou testových položek je v tom, že položkové banky umožňují sledovat položku v celém jejich životním cyklu, v němž položka může být využita v řadě testů. Informace o chování úlohy v jednom cyklu se ukládají a využívají pro konstrukci kvalitnějších testů v dalších cyklech. Tato „znovupoužitelnost“ a zvyšování kvality díky vyhodnocování předchozích použití položek patří k základním rysům položkových bank.

> Na přelomu století byly předmětem výzkumu „znovupoužitelné výukové objekty“. V případě výukových materiálů se snaha uchovávat je v repozitářích i s metadaty pro další použití neprosadila. Popis pomocí metadat byl pro vytížené pedagogy příliš pracný

a jeho náročnost potlačila efekt úspory z opětovného využití. V případě položkových bank je situace jiná. Práce s metadaty je užitečná a většinou plně automatizovaná. Myšlenka využití „znovupoužitelných výukových objektů“ tak dochází po desetiletích svého naplnění, i když v jiném kontextu, než se původně plánovalo.

Položkové banky jsou nezbytnou součástí procesu kvalitního hodnocení. Kromě podpory tvorby testových otázek umí mnohem více: ukládat metadata o položkách, ukládat psychometrické vlastnosti položek v testech, sledovat jejich využití, řešit správu uživatelů, správu bezpečnosti a vynucovat správné pracovní postupy, které pomáhají dodržovat kvalitativní normy. Položkovou banku nepotřebujete, pokud děláte malý počet spíše formativních testů. Ale neobejdete se bez ní, pokud připravujete rozsáhlá a důležitá hodnocení studentského výkonu.

8.1 Položky a jejich metadata

Položkové banky neobsahují pouze text každé položky, ale také řadu informací týkajících se jejího původu, začlenění, využití a psychometrických charakteristik. Mezi tato metadata patří například:

- text položky,
- datum vytvoření,
- správná dopověď,
- formát položky,
- přiřazení k tématům,
- autor položky,
- recenzenti položky,
- vyjádření recenzentů (pro Angoffovu metodu),
- stav recenze (k recenzi, hotová, odmítnutá, k přepracování, ...),
- stav položky (např. nová, zrecenzovaná, aktivní, archivovaná, nahrazená novou verzí, ...),
- charakteristiky podle klasické testové teorie,
- charakteristiky podle teorie odpovědi na položku,
- zapojení položky v testových plánech,
- vztahy k ostatním položkám.

Při konstrukci položkové banky je třeba přihlédnout k potřebám konkrétního oboru a přizpůsobit mu například již členění položek. Položka totiž může být v databázi uložena jako celek, nebo po jednotlivých částech – jako kmen (medailonek), samotná otázka a nabídnuté odpovědi. Pak lze snadno generovat odvozené varianty položek (např. s odlišnými daty pro výpočet), ale je obtížnější udržet pořádek v informacích, v jaké podobě byla položka použita v kterém testu. Zvolit způsob ukládání je třeba provést na začátku, závisí na něm řada dalších vlastností položkové banky.

8.2 Typy vztahů mezi položkami v testech

U položek se ukládá řada metadat. Mezi jinými se ukládají i vztahy mezi položkami, které popisují, zda a za jakých okolností mohou dvě úlohy být spolu v jednom testu.

Přátelé	Položky se musí vyskytovat společně, neboť se opírají o společný základ, nebo používají společný podpůrný materiál.
Blízcí přátelé	Položky se musí vyskytovat společně. Pokud se objeví jedna, musí být přítomna i druhá.
Snobové	Položky typu „Blízcí přátelé“, které mohou být použity jen v určitém pořadí, aby byly srozumitelné.
Závislí	Položky, které se mohou vyskytnout pouze s podporou „Podporovatel“.
Podporovatelé	Položky bez samostatné interakce, které poskytují kontext pro následující „Závislé“ položky. Může jít například o text, na který se bude několik úloh ptát.
Antagonisté	Položky, které nesmějí být v testu blízko, neboť je jedna pro druhou nápovědou.
Nepřátelé	Položky nesmějí být ve stejném testu, protože se ptají na totéž.
Klon (potomek)	Položka byla odvozena změnou (opravou, vylepšením) rodičovské položky.

Tab. 8.2.1 Typy vztahů mezi položkami v testech

8.3 Funkcionality položkových bank

Od položkové banky obvykle očekáváme, že zajistí a bude podporovat:

- autentizaci přístupů,
- vytváření položek,
- ukládání položek,
- třídění položek,
- recenzi položek (včetně řízení postupu recenze),
- tvorbu testových plánů (blueprint),
- správu položek,
- správu testů,
- správu výukových cílů pro kategorizaci položek,
- tisk nebo elektronické provedení testu,
- načtení výsledkových formulářů,
- automatické psychometrické analýzy,
- logování aktivit,
- záznamy o exportech položek a testů,
- uchovávání dat psychometrických dat z předchozích použití položek.

Zásadní vlastností, která odlišuje položkovou banku od úložiště položek, je právě to, že položková banka uchovává psychometrické charakteristiky z předchozích kol testování. Umožňuje tak jejich využití pro tvorbu lepších položek a konstrukci lepších testů.

Jak si čtenář jistě povšiml, ve výčtu vlastností není explicitně uvedeno automatizované generování testů. Ukazuje se totiž, že návrh testu vyžaduje, aby bral autor v úvahu velké množství často nevyjádřených parametrů položek a jejich zastoupení v testu vyvažoval. Takže i v případě, že systém umí vygenerovat návrh sady testových položek, stejně se předpokládá, že tento návrh projde korekturou „autora“ testu, který je zodpovědný za jeho vyváženost. Položková banka k tomu nabízí autorovi řadu nástrojů, například možnost sledovat, jak úlohy pokrývají testovanou oblast (blueprinting), kolik je v testu výpočetních úloh, a podobně.

Komplexní položkové banky mají zabudované další užitečné mechanismy, které například hlídají změny závislých parametrů položek. Po přepracování položky tak dokážou změnit proměnou „stav recenze“ z hodnoty „k přepracování“ na hodnotu „k recenzi“ apod.

Poznámka: Vliv rozhodnutí o potřebnosti forenzních analýz testu na strukturu položkové banky. Samotná položková banka nepotřebuje ukládat informaci o identitě testovaného. Pokud to autoři položkové banky považují v daném kontextu za potřebné, mohou ukládat do položkové banky informace o ročníku, pohlaví a dalších atributech, ale pro ukládání konkrétní identity většinou není důvod. Situace se změní, když se ukáže, že je potřeba statistickými nástroji prověřovat, zda při testování nedochází k nedovolenému jednání. V tom případě je nezbytné s identitou testovaných pracovat. Jde o koncepční otázku, protože se pak musíme zabývat nejen *statistikami testových úloh*, ale i *statistikami testovaných*. Při návrhu položkové banky je vhodné na tuto okolnost pamatovat a počítat s ní předem.

8.4 Výhody položkových bank

Položková banka by měla být v základu každého seriózního systému pro testování. Její použití totiž přináší řadu výhod (Weiss 2011):

1. Umožňuje opakovanou tvorbu testů s predikovatelnými vlastnostmi.
2. Poskytuje možnost objektivně zjistit specifika jednotlivých autorů položek. Může se ukázat, že někteří autoři položek připravují systematicky lehčí nebo naopak obtížnější položky. Někteří upřednostňují jednu tematickou oblast, nebo jeden typ úloh (např. výpočetní). Potřebujete-li např. doplnit testovou sadu o specifickou položku, víte, na koho se obrátit.
3. Pravidelná práce s autory vám umožňuje je školit a zvyšovat jejich dovednosti při tvorbě úloh.
4. Banka nutí ke standardizovanému postupu při přípravě testových položek (obsahová recenze, jazyková korektura, typografická kontrola, stanovení obtížnosti, vyhodnocení po testu...), což je zárukou systematického zvyšování kvality.
5. Udržuje se pořádek v různých verzích položek. Když je potřeba úlohu opravit nebo opravit, vzniká podle rozsahu změny buď její nová verze, nebo jen upravená varianta (např. pokud jde jen o opravu překlepů nebo úpravu typografie).
6. Všechny položky jsou přiřazeny ke konkrétním tématům a systém v nich umožňuje vyhledávat podle řady kritérií. Tím je při tvorbě testů zajištěno lepší pokrytí testované látky, usnadní se dodržení plánu testu, zamezí se opakování a nedochází k problémům s položkami s neznámým nebo nepotřebným zaměřením.
7. Položkové banky umožňují k položkám přiřadit výsledky psychometrické analýzy proběhlých testů. Je tak možné vytřídit málo kvalitní položky, nebo sledovat, zda položka nebyla mezi dvěma testy vynesena.
8. V položkové bance jsou nastavována oprávnění jednotlivých uživatelů na základě rolí. Současně jsou všechny aktivity logovány, zejména změny položek, hromadné exporty a přístupy k finalizovaným testům. To vše přispívá ke zvýšení bezpečnosti testu.
9. Položková banka by měla umožnit snadnou identifikaci duplicitních položek a položek, které mají nějaký typ vzájemné vazby (nepřátelé, přátelé, blízcí přátelé, ...).
10. Položková banka zvyšuje efektivitu a kvalitu testování tím, že prohlíží na položky jako na *znovupoužitelné objekty* a podporuje celý cyklus vývoje testů.

8.5 Příklady položkových bank

Položková banka je v podstatě jednoduchá databáze, může tedy být uložena v databázovém systému, nebo dokonce v prostředí tabulkového procesoru.

Příklad položkové banky **vytvořené v Excelu** byl prezentován na konferenci Association for Educational Assessment – Europe 2016 (Verschoor a Jongkamp 2016). Řešení bylo relativně jednoduché, a přitom zcela funkční a rozhodně můžeme potenciálním zájemcům tento krok doporučit. I kdyby to bylo jen dočasné řešení – pomůže vám vyjasnit, jaké máte potřeby a co požadujete od případného budoucího komplexnějšího řešení.

Vlastní položkové banky si vyvinula většina velkých společností, které se zbývají testováním. V Česku mají bezesporu nějakou formu položkové banky například SCIO nebo Cermat. Na některých vysokých školách mají části položkové banky integrované přímo v informačním systému školy.

Položková banka 1. LF UK

Samostatnou položkovou banku si vyvinuta i 1. LF UK. Jedná se komplexní položkovou banku pro formát položek MTF. Podobně jako v případě testovacího systému Rogō se jedná o webovou aplikaci, která běží na všech hlavních prohlížečích. Banka, jejíž vývoj se datuje od roku 2014, podporuje všechny kroky testového cyklu od blueprintingu přes tvorbu položek, jejich recenzování a správu verzí. Banka umožňuje tvorbu testů, jejich recenzování, tiskovou přípravu, import výsledků, položkovou analýzu a reportování výsledků testu. V bance je asi 10 tisíc položek včetně metadat o výsledcích použití v předchozích kolech. Vzhledem k využití při zkouškách velkého významu je banka silně zabezpečena a každý přístup do ní je logován. Vlastnosti právě této položkové banky jsou základem obecného popisu položkových bank uvedeného výše. A zrcadlově – tato banka má všechny vlastnosti požadované v obecném modelu.

Řada zájemců o testování hledá ekonomicky přijatelnou cestu, jak pořídit některou komerční položkovou banku. Nabízených možností je celá řada, ale model jejich licencování zpravidla vychází z prostředí, kde je na tyto účely alokováno podstatně více prostředků, což tato řešení činí pro tuzemské zájemce prakticky nedostupnými.

Zkušenost s TAO of testing

Na pomezí komerčních produktů stojí systém „TAO of testing“, který nás zaujal dostupností bezplatné verze. Protože by to mohlo zlákat i další zájemce, považujeme za užitečné naši zkušenost sdílet.

TAO je jednak v čínské filozofii výraz pro základní princip vesmíru, ale též francouzská zkratka pro Testing Assisté par Ordinateur (počítačově podporované testování). Platforma TAO je vyvíjena na Lucemburské univerzitě jako open source projekt. Poskytuje účastníkům procesu počítačového testování komplexní sadu funkcí, které podporují vytváření, správu a administraci elektronických testů. Pokrývá zejména:

- vývoj a správu položek,
- správu testovaných,

- vytváření a správu testů,
- správu autorů a recenzentů,
- doručení testů,
- správu výsledků.

TAO je otevřený a modulární systém, který vychází z předpokladu, že žádné řešení nemůže vyhovovat všem, takže se očekává, že si jej uživatelé přizpůsobí pro své potřeby. Jde o webovou aplikaci, která běží na serveru a nevyžaduje cokoliv instalovat na počítači uživatele. Podporuje překlady do národních jazyků. Autorům nabízí pro intuitivní tvorbu položek WYSIWYG editor, včetně integrace multimédií.

Systém je sice open source, ale „zadarmo" je jen zdánlivě. Máte dvě možnosti, jak jej využívat. Buď využijete placenou cloudovou instalaci poskytovatele, která stojí jako ostatní položkové banky na trhu, nebo si TAO nainstalujete na svoje servery sami. Dokumentace je ovšem nedostatečná a instalace a aktualizace jsou špatně popsané. Instalační skripty obsahují chyby a v manuálech jsou odkazy na neexistující skripty a další zdroje. Výrobce sice nabízí podporu, ale ta je draze zpoplatněna.

V systému chybí podpora pro adresářové služby (LDAP), což znemožňuje využít v konkrétní instituci její stávající ověřování identit (a správu uživatelských jmen a hesel). To je velmi nepraktické při nasazení ve velkých institucích, které by tak kvůli testování musely udržovat adresářů několik. Systém TAO rovněž neobsahuje nástroje pro analýzu testů a položek (pouze export do formátů QTI 2.2. nebo CSV), takže pro testové a položkové analýzy je potřeba používat software jiných výrobců.

Tuto položkovou banku jsme implementovali a experimentálně provozovali (1. LF UK 2017), ale problémy s dokumentací a aktualizacemi i nutností řešit uživatelská jména a hesla studentů separátně mimo stávající adresářové služby (LDAP) vedly k ukončení tohoto experimentu.

8.6 Rozsáhlé banky testových úloh

V některých případech přerostla běžná práce s položkovou bankou obvyklý rozměr. Zmiňme dva takové zajímavé případy. Nejzajímavější z těchto spoluprací je britská *Medical Schools Council Assessment Alliance* (MSA-AA), jež sdružuje všech 31 lékařských škol ve Velké Británii (Medical Schools Council nedatováno).

Medical Schools Council Assessment Alliance

Provozuje pro ně společnou položkovou banku. Společným cílem všech účastníků je zlepšit hodnocení výsledků výuky na lékařských fakultách.

Aliance navazuje na činnost Sdružení lékařských fakult – *Universities Medical Assessment Partnership* (UMAP), které bylo založeno v roce 2003 za účelem spolupráce při tvorbě a sdílení testových položek. Sdružení se postupně rozrůstalo o další školy a po roce 2009 se přeměnilo na MSC-AA.

Společná položková banka obsahuje především položky ve formátu s jednou nejlepší odpovědí (SBA), ale přibývají i úlohy pro stanice OSCE a „multiple-mini interview“. Banka je přístupná všem zúčastněným školám. Otázky vznikají ve spolupráci a procházejí rozsáhlým testováním kvality a standardizací. Všechny lékařské vysoké školy v UK se dohodly, že zahrnou do závěrečných zkoušek dohodnutý podíl sdílených otázek, čímž zajistí vzájemnou psychometricky validní porovnatelnost „státních“ zkoušek.

Item Management System

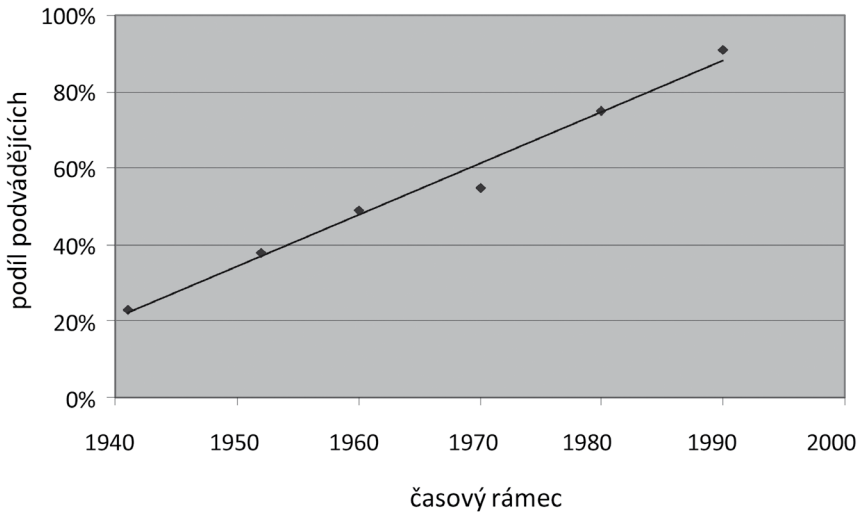
V německy mluvících zemích se na mnoha vysokých školách používá položková banka *Item Management System* (IMS). Vznikla v roce 2006 jako výsledek spolupráce lékařských fakult univerzity v Heidelbergu, v Berlíně a v Mnichově. Skupina se postupně rozrostla na 77 institucí zejména v Německu a ve Švýcarsku a je zastřešována konsorciem *Umbrella Consortium for Assessment Networks* (UCAN).

Item Management System je databáze položek, která umí spolupracovat s aplikacemi pro počítačové testování, pro papírové či mobilní testování, pro analýzu výsledků a tak dále. V systému bylo k březnu 2021 uloženo 700 000 položek (bez rozlišení formátu), z nichž 125 000 bylo sdíleno (Institute for Communication and Assessment Research nedatováno, Hochlehnert et al. 2012).

Umbrella Consortium for Assessment Networks (UCAN), které se o položkovou banku stará, je deklarováno jako nezisková organizace, ale navenek funguje jako komerční subjekt. Licence se počítá podle počtu testovaných. Ceny jsou nastaveny pro západoevropské poměry.

9 BEZPEČNOST TESTOVÁNÍ

Jakkoliv máme tendenci věřit tomu, že vše se vyvíjí směrem k lepšímu, z publikovaných údajů se zdá, že tendence podvádět při zkouškách dlouhodobě spíš roste (Bernardi et al. 2008, Cizek a Wollack 2017).



Obr. 9.1 Vývoj pravděpodobnosti podvádění studentů u zkoušky v průběhu půl století (Sims 2010)

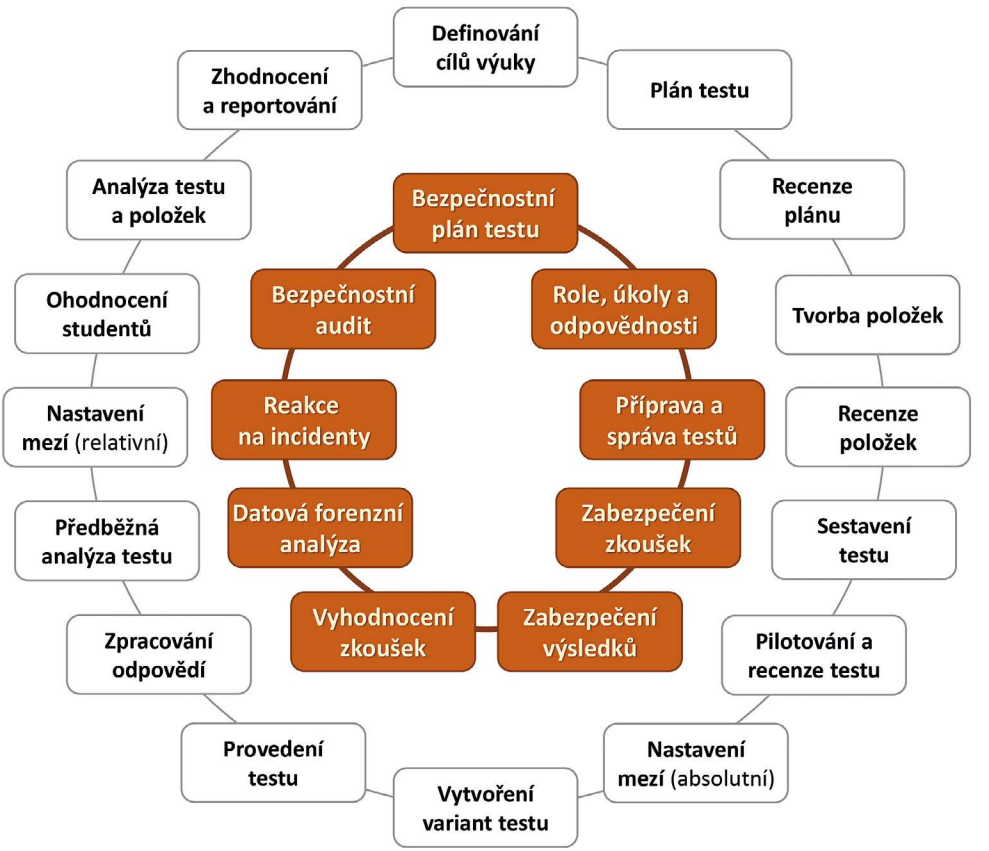
Podle provedených výzkumů se míra podvádění v testech v průběhu minulého století dramaticky zvýšila (Bernardi et al. 2008). Například mezi lety 1963 a 1993 se míra vážného podvádění při písemných testech zvýšila z 39 % na 64 % (McCabe et al. 2001).

Testy jsou vstupní branou k mnoha vzdělávacím a profesním cílům. Není proto divu, že motivace k podvádění je vysoká. Při neformálním průzkumu provedeném v roce 2007 mezi 30 000 americkými vysokoškoláky jich 60,8 % přiznalo, že během studia podvádělo (Online Education Database 2010). Tentýž průzkum ukázal, že 16,5 % z nich to ani nepociťuje jako etický problém. V jiných studiích až 85 % studentů uvádí, že se alespoň jednou za své studium dopustili při testech podvodu (Klerk et al. 2019). Současně se zvyšuje sociální tolerance k podvádění, zvláště pokud se uskutečňuje prostřednictvím internetu (Vrbová 2013).

Překvapivé zjištění přinesl průzkum provedený na Fordhamově univerzitě: poukázal na významný rozdíl mezi studijními průměry podvádějících studentů a jejich poctivých protějšků. Podvodníci patří do skupiny se statisticky významně lepšími studijními výsledky než ti, kdo nepodvádějí. Je na místě klást si otázku, jakou roli při formování těchto postojů hraje motivace zaměřená výhradně na úspěch.

Současně s tím 91 % studentů považovalo ignorování podvádění ze strany vyučujících za silně neetické (Vrbová 2013).

Za těchto okolností je zřejmé, že u testů velkého významu je potřeba věnovat pozornost jejich zabezpečení. V širším kontextu úvah o bezpečnosti testování musíme brát na zřetel sumu hodnot, které by mohly být při narušení bezpečnosti ohroženy. Pokud by podvádějící byli úspěšní, mohli by testem projít uchazeči, které by test měl naopak vyloučit. Zmařena by tak byla nejen práce učitelů, kteří testy připravovali, ale ohrožena by byla i pověst a důvěryhodnost celé instituce, která testy provozuje. Jedním z předpokladů validity sumativního hodnocení je jeho věrohodnost a objektivita. Význam zajištění bezpečnosti hodnocení přitom roste s důležitostí zkoušky.



Obr. 9.2 Schéma bezpečnostního cyklu rozhodného testování. Volně podle (Klerk et al. 2019).

Je zvláštní, že o tomto důležitém tématu pojednává méně literatury, než bychom očekávali. Z části to může být tím, že potřebné know-how bylo roztříštěné a v držení testujících subjektů. Zčásti pak tím, že zveřejňování postupů pro odhalování podvádění je v rozporu s jejich záměrem. Nicméně v posledních desetiletích se přece jen začínají objevovat i texty pokrývající celou tuto oblast (Foster 2013; Cizek a Wollack 2017).

Většina hodnocení na vysokých školách má periodicitu spojenou s rytmem akademického roku. V důsledku toho by se i péče o bezpečnost měla pravidelně opakovat. Instituce nastaví pravidla (např. přijímacího řízení) a účastníci se v rámci těchto pravidel pokusí dosáhnout co nejlepšího výsledku. Instituce pak svá pravidla koriguje, aby optimalizovala výběr, při zachování nestrannosti a objektivity. Bezpečnost testování z tohoto pohledu není ustálený stav, ale cyklický proces.

9.1 Bezpečnost testování z pohledu řízení rizik

Při posuzování významu (a racionality vynaložených nákladů) jednotlivých aspektů bezpečnosti testování můžeme použít metodiku známou z oblasti řízení rizik. V rámci instituce a v kontextu připravovaného testování bychom se měli zamyslet nad:

- identifikací aktiv,
- ohodnocením aktiv,
- identifikací hrozeb,
- odhadem pravděpodobnosti výskytu jednotlivých hrozeb,
- odhadem zranitelnosti aktiva hrozbou,
- odhadem celkového rizika vyplývajícího z jednotlivých ohrožení aktiv.

Identifikace aktiv

Aktivy se myslí vše, co pro organizaci představuje hodnotu. Ačkoliv to nemusí být z úzkého pohledu na testování zřejmé, mezi hlavní aktiva ve školství patří zejména důvěryhodnost celé instituce. Existují případy, kdy hrubé chyby v zabezpečení procesu testování (při přijímacím řízení) vedly až k fatálním následkům v podobě odebrání akreditace.

Ohodnocení aktiv

Při ohodnocení aktiv můžeme odhadnout například cenu položkové banky, respektive jejího obsahu. V položkové bance 1. LF UK je asi 10 tisíc položek. Náklady na jejich pořízení byly asi 1500 Kč za položku. Odtud dostáváme cenu za celý obsah položkové banky 15 milionů Kč. Odhadnout můžeme i cenu vybudované důvěryhodnosti instituce. Jde o dlouhodobě budované hodnoty s velkými náklady. Pokud by došlo např. ke skandálu kolem přijímacího řízení, pak to může znamenat, že byly znehodnoceny náklady na PR za dobu třeba 5 let. Kdyby ztráta důvěry v regulérnost řízení vedla až k odebrání akreditace, ztratí fakulta příjmy za výuku studentů, tedy jeden z největších zdrojů příjmu, a to na několik let. U větší fakulty tak ztráta dosáhne řádově stovek milionů Kč. Mezi chráněná aktiva ale patří např. i osobní údaje účastníků testů, které jsou chráněny podle GDPR pod hrozbou drakonických pokut.

Identifikace hrozeb

Hrozbami rozumíme takové scénáře, v nichž mohou být ohrožena aktiva organizace. Při testování jde zejména o vnější a vnitřní hrozby. Mezi hrozbami zaujímají zvláštní místo pokusy o úmyslné ovlivňování výsledků nedovolenými postupy. Existuje řada typů neetického a podvodného jednání, které mohou ohrozit vypovídací schopnost testu:

> **Vynesení položek**, nebo též neoprávněné získání předběžných znalostí, může nastat, pokud účastníci předchozích běhů vynesou znění otázek. Buď si obsah zkoušky zapamatují, nebo zkopírují zadání mobilem, nebo ho opíšou. Cílem je dosáhnout v testu výhody proti ostatním testovaným, díky znalosti konkrétních testových otázek. Neoprávněný přístup k testovacím otázkám přináší podvádějícím nespravedlivou výhodu oproti poctivým účastníkům testu.
> **Nedovolená spolupráce.** Dva nebo více testovaných se mohou pokoušet spolupracovat na vypracování testu. Například opisovat odpovědi, nebo sdílet odpovědi během testu prostřednictvím textových zpráv a podobně.
> **Záměna identity.** K znehodnocení vypovídací schopnosti testu může dojít, pokud test absolvuje někdo jiný než skutečný uchazeč. Tomuto „testovacímu proxy" lze zabránit dodržením vysokých standardů pro ověření identity. Zvláště při distančních testech, kdy jsou možnosti pro ověřování identity omezené, je třeba věnovat tomuto problému velkou pozornost.
> **Nedovolená pomoc.** Výsledek testu může být zkreslen i „koluzí" (utajenou dohodou), pokud se testovanému dostane pomoci od personálu, který testy organizuje či vyhodnocuje. Koluze znamená, že testový dohled, nebo správce testu poskytl zkoušenému neoprávněnou pomoc nebo nějakým způsobem manipuloval s testovými daty nebo testovací relací. Příkladem tajné dohody může být situace, kdy dohlížitel umožní zkoušenému odchýlit se od schválených testovacích postupů, získat přístup k neautorizovaným zdrojům nebo mu umožní překročit schválený čas na vyplnění testu. Utajená dohoda se může také vztahovat na manipulaci se záznamy o zkoušce, jako je změna odpovědí zkoušeného ze špatné na správnou, nebo přidání chybějících odpovědí.
> **Nedovolené pomůcky a zdroje.** Podle průzkumu mezi patnáctiletými žáky v Čechách byl v roce 2013 stále nejrozšířenějším způsobem podvádění používání „taháků", teprve na dalších místech byly technické prostředky jako mobilní telefony a podobně (Vrbová 2013).

Hrozby pro bezpečnost se u jednotlivých druhů testování liší. Papírové zkoušky mohou být náchylnější na kopírování odpovědí než zkoušky na počítači (zvláště v případě adaptivního testování), zatímco počítačové testování může být náchylnější k použití neautorizovaných zdrojů. Zásady a postupy zabezpečení by měly být přizpůsobeny tak, aby vyhovovaly danému druhu testů.

Pro každý typ ohrožení je třeba připravit:

- odhad pravděpodobnosti a potenciálních důsledků každého z možných případů,
- preventivní opatření k omezení hrozeb,
- následné postupy pro minimalizaci dopadu mimořádných událostí.

Ačkoliv je oceňování rizik pracné a nemusí být na první pohled pochopitelné, proč se jím zabývat, jeho smysl je zásadní – pomoci zajistit na ochranu důležitých hodnot prostředky, které jsou v rozumné relaci ke chráněným hodnotám.

Je zřejmé, že prevence a eliminace bezpečnostních hrozeb si vyžaduje systematický přístup. Proto při sumativním testování velkého významu bývá vytvořen bezpečnostní plán testu, který specifikuje kdo, kdy a čím se má zabývat, aby bylo dosaženo potřebné úrovně bezpečnosti. Pojďme si takový bezpečnostní plán projít krok za krokem.

9.2 Bezpečnostní plán testu

Bezpečnostní plán testu popisuje hodnoty, které je třeba chránit, známá rizika a postupy, jež mají rizika omezit.

Plán zabezpečení testu je komplexní soubor zásad, postupů a dokumentů, které nastiňují a řídí akce související s bezpečností testu. Od vypracování plánu zkoušky po rekapitulaci výsledků při bezpečnostním auditu se „bezpečnost“ týká téměř každého kroku. Využití skóre dosažených v testu pro posouzení výkonu kandidátů předpokládá důvěru v integritu a objektivitu testu. Bez důvěry by byla ohrožena i důvěryhodnost.

Co je potřeba udělat, aby testová skóre byla důvěryhodná a interpretovatelná? Minimálně to vyžaduje mít spolehlivý testovací bezpečnostní plán a řídit se jím.

Většina zásad a postupů v rámci bezpečnostního plánu testu je založena na zdravém rozumu. Například je nezbytné mít komunikační kanál pro jasné a jednoznačné zasílání zpráv uchazečům. Jak jinak lze od uchazečů očekávat, že se budou řídit pravidly, pokud nebudou tato pravidla spolu s odpovídajícími důsledky vysvětlena? Přijaté postupy musejí dávat smysl, dobře zapadat do daného testovacího programu, musí být vynutitelné a musí být právně obhajitelné. Navržené postupy by se měly sladit s hrozbami, které jsou specifické pro vaše programy, a měly by být přizpůsobeny konkrétním potřebám. Zatímco v jednom případě může být hlavním problémem vynesení položek mezi termíny zkoušek, v jiném uspořádání může být největší hrozbou podvod s identitou uchazeče. Hrozby se v jednotlivých programech liší a bezpečnostní plány by měly ochranu před těmito hrozbami řešit.

Role, úkoly, odpovědnosti

Příprava důležitých testů je kolektivní práce. I z hlediska bezpečnosti by byla věrohodnost obtížně zajistitelná, pokud by se všechny pravomoci koncentrovaly v rukou jediného člověka.

S omezením těchto rizik pomáhá přístup založený na rolích. Všichni pracovníci spolupracující na testování by měli mít specifikované role a pracovat jen v rozsahu těchto rolí. Někdo může mít roli „autor položky“, někdo roli „recenzent položek“, další „autor testu“ nebo „správce testování“. Bezpečnostní limity rolí pomohou zajistit, že např. ten, kdo má na starosti správu seznamu testovaných studentů, možná nikdy neuvidí žádné testové položky.

Mezi odpovědnosti, které musí bezpečnostní tým zajistit, patří mimo jiné i ochrana interních informací před vynesením. V rámci tzv. „měkké bezpečnosti“ se o tuto důvěrnost staráme

výběrem odpovědných pracovníků, jejichž morální integrita napovídá, že se o utajovaných informacích nebudou zbytečně šířit, či nepodlehnou pokušení tyto informace poskytnout někomu za úplatu. V konceptu „tvrdé bezpečnosti“ (například když nemáme o zapojených pracovnících dostatek informací) se používá pro ochranu utajovaných informací testovací bezpečnostní dohoda, někdy označovaná jako dohoda o mlčenlivosti (*non disclosure agreement*, NDA). Jde zpravidla o jednostrannou, právně závaznou smlouvu mezi institucí vyvíjející test (nebo vlastníkem obsahu) a další stranou řešící dílčí úkol. Dohoda o mlčenlivosti obvykle stanoví, jaké informace nebo materiály jsou považovány za důvěrné a/nebo chráněné, jaká je lhůta pro zachování důvěrnosti a jaké jsou důsledky porušení dohody.

K opakujícím se úkonům patří aktualizace právních zásad a postupů a školení zaměstnanců o testové bezpečnosti. V době, kdy je většina testů produkována a uchovávána v elektronickém prostředí, je úkolem testového týmu zabezpečit testová data na lokálních nebo cloudových serverech. Přístup k těmto serverům musí být omezen jen na p(r)ověřené pracovníky, monitorován a logován.

Je dobrou praxí vyžadovat, aby každý, kdo má přístup k testovému obsahu nebo jiným chráněným informacím, byl vyškolen a podepsal dohodu o mlčenlivosti. Patří sem odborníci, kteří se podílejí na vývoji testů, pracovníci, kteří monitorují provádění zkoušky, zaměstnanci, kteří zpracovávají testovací materiály a výsledky, učitelé, kteří přijímají nebo ukládají testovací materiály atd. Smlouvy o mlčenlivosti by měly být každoročně aktualizovány a uchovávány v evidenci po dobu stanovenou v bezpečnostním plánu zkoušky (obvykle nejméně tři roky).

Příprava a správa testů

Existuje řada bezpečnostních opatření, která by měla být provedena před samotným testem. Sem patří nejen bezpečná příprava obsahu testu, ale i monitorování webů a sociálních médií. Nebezpečí vynesení položek se umocňuje s technologiemi sdílení obsahu. Existují specializované stránky, které od jednotlivců sbírají jimi zachycené položky k certifikacím a zkouškám, kumulují je podle kategorií a za úplatu pak nabízejí zájemcům. Tyto stránky je možné najít pod heslem „brain dump“. Bezpečnostní příprava proto předpokládá, že tým připravující test bude sledovat sociální sítě, zkoušet cílené dotazy do webových vyhledávačů při hledání uniklých položek či sledovat blogy komentující danou zkoušku či certifikaci, aby mohl uniklé položky včas identifikovat.

Při podezření na nelegální praktiky je možné použít techniku označovanou jako „tajné nakupování“. Tento typ ověřování bezpečnosti zkoušky předpokládá, že domluvený spolupracovník testového týmu se jako student zaregistruje k provedení zkoušky a po jejím provedení podá zprávu o bezpečnosti testu z pohledu zkoušeného. Tato forma monitorování bezpečnosti je sice nákladná, ale v případě pochybností může poskytnout velmi cenné a jinak nedostupné údaje.

Vzhledem k významu je třeba věnovat pozornost i distribuci citlivých materiálů a přístupu k nim. V testovém plánu by proto měly být popsány postupy, jak jsou chráněné materiály distribuovány, shromažďovány a archivovány a kdo k nim má přístup. Nahlížení a zásahy do citlivých materiálů (např. do znění ostrých testů) musí být buď zaznamenáno technickými prostředky (logování, kamerové záznamy), nebo provedeno komisionálně (nejméně ve dvou

lidech) a o úkonu by měl by být proveden zápis. Nekontrolovaným rizikem jsou individuální přístupy a zásahy, o kterých není zpětně žádný doklad.

V neposlední řadě je třeba se zabývat otázkou školení. Každý, kdo se účastní testovacího cyklu, by měl být vyškolen v oblasti bezpečnosti testu. Školení se může zabývat celou řadou témat, včetně, ale nejen, správného zacházení s testovacími materiály, zřízení nebo udržování bezpečného testovacího prostředí, kritických aspektů dohody o zachování důvěrnosti, práv a povinností zkoušejících, či přijatelných postupů testového dozoru. Školení může také zahrnovat scénáře typu „co dělat, když". Školení by mělo být přizpůsobeno tak, aby bylo v souladu s rolemi různých členů týmu, včetně odborníků na předměty, dohledů, administrátorů a koordinátorů zkoušek, pracovníků pro vývoj obsahu, psychometriků a managementu. Školení by měli absolvovat i pracovníci třetích stran, kteří na testování spolupracují. Zajistit bezpečnost testování vyžaduje spolupráci celého týmu.

Zásady testovacího dne

Dalším důležitým aspektem bezpečnosti jsou tzv. *zásady testovacího dne*. Je testovací prostředí bezpečné? Jsou pracovníci dohledu dostatečně proškoleni o bezpečnosti testů? Jaké jsou požadavky na přihlášení? Jak se účastníci identifikují? Kolik forem identifikace je potřeba? Je průběh testu natočen kamerovým systémem? Existuje předem stanovený zasedací pořádek testovaných v místnosti? Existuje bezpečné místo pro ukládání osobních věcí, jako jsou mobilní telefony a studijní materiály? Jsou povoleny kalkulačky? Je dodáván odpovědní formulář? Pokud ano, je formulář individualizovaný? Sbírají se formuláře na konci testu? Používají se chrániče obrazovky u počítačových monitorů? Jsou pracovní stanice odděleny? Jsou během testu povoleny přestávky nebo odchody na toalety?

Komunikace s uchazeči začíná v dostatečném předstihu před datem zkoušky a pokračuje až do okamžiku, kdy jsou oznámeny výsledky. Pravidla musí být jasně stanovena a šířena zájemcům a zúčastněným stranám. Kromě toho musí být jasně stanoveny a šířeny důsledky porušení pravidel. Před testováním lze od zkoušejících požadovat, aby potvrdili, že četli, porozuměli a souhlasili s dodržováním vyžadovaných pravidel.

Při pravidelném testování velkého významu se neobejdeme bez nějaké formy komplexního testového systému (položkové banky). To ovšem přináší nový druh rizik, protože cenné informace (znění položek, ale i znění připravených ostrých testů) jsou zde koncentrovány ve finální podobě na jednom místě po dlouhou dobu, což zvyšuje riziko jejich nežádoucí expozice. K důležitým bezpečnostním opatřením týkajícím se položkových bank patří technické zajištění trvalého logování rizikových událostí, zejména spojených s exporty testů, nebo zobrazením většího množství testových položek, či přímo celých testů.

Zabezpečení výsledků

Další složkou zabezpečení jsou postupy při skladování a distribuci citlivých materiálů (např. zadání testů) a uchovávaní testových výsledků. Tento postup určuje, jak jsou chráněné materiály distribuovány, shromažďovány a archivovány. Ukládána jsou také jména a funkce osob odpovědných za provádění těchto postupů. Data a podpisy od každé osoby zapojené do testování a dohledu jsou shromažďovány a archivovány jako součást historie testů. Obecně platí, že při práci s citlivými údaji je třeba přístupy buď logovat, nebo dělat pod kontrolou dvojích očí a vést o provedených úkonech protokoly.

Rychlá analýza testových dat

Pro odhalování náznaků nesrovnalostí v právě proběhlých (ale ještě neobodovaných) testech je mimořádně cenným nástrojem *rychlá analýza testových dat*. Umožňuje například ještě před klasifikací studentů odhalit případné nejasnosti ve znění položek, chyby v klíči určujícím správné odpovědi a podobně. Podezřelé položky např. s velmi vysokou nebo nízkou obtížností, nebo s velmi nízkou diskriminační schopností jsou podrobeny obsahové kontrole a v případě chyb, dvojznačností nebo nepřesností je taková položka je vyloučena z bodování, nebo je upraven klíč pro její bodování. Informaci o problematických položkách dostávají i autoři a recenzenti, aby je opravili před dalším použitím (Martinková et al. 2017b). Podobně může rychlá analýza zachytit i některé nestandardní vzorce chování poukazující na případné bezpečnostní problémy.

Vyhodnocení průběhu testového kola

Pro zajištění věrohodnosti testů by měl v testující organizaci existovat *postup pro hlášení incidentů a nesrovnalostí* ve správě a zabezpečení testu. Účastníci testu, členové pedagogického dohledu a další testovací pracovníci by měli znát mechanismus pro hlášení incidentů, anomálií nebo možných porušení pravidel. Nabídnutá forma by měla sahat od anonymního upozornění až po formální zprávu.

Reakce na incidenty

Bezpečnostní plán testu by měl stanovit, jak budou incidenty registrovány, zpracovávány a vyšetřovány. Má být zřejmé, za jakých okolností bude zneplatněno dosažené skóre a kdy se přikročí k případným sankcím.

Bezpečnostní audit testu

Při bezpečnostním auditu testu bezpečnostní tým rekapituluje opatření, která byla preventivně učiněna, jejich efektivitu, ohrožení, která byla zaznamenána, jak byla vyřešena a jaké úpravy bezpečnostních pravidel je třeba provést před dalšími testy.

9.3 Bezpečnostní analýza testů

Při narušení akademické integrity nemusí skóre testů odrážet schopnosti a znalosti testovaných. Forenzní analýza testů (*educational data forensics*, EDF) je statistická analýza výsledků testů s cílem detekovat odchylky, které potenciálně naznačují neoprávněný zásah, zvýhodnění, nebo přímo testovací podvod. Pokud by docházelo k porušování akademické integrity na úrovni správců testu nebo administrátorů položkové banky, je forenzní analýza prakticky jediným nástrojem, jak tuto činnost systematicky odhalovat.

Analýza nám má odpovědět na
Otázky zaměřené na jedince:

- Je na tomto vyšetřovaném něco neobvyklého?
- Odpověděl na každou položku „C“?
- Odpovídal příliš rychle?
- Strávil 10 minut u každé z prvních 5 položek a zbytek přeskočil?
- Získal vysoké skóre v podezřele krátké době?
- Změnil nápadně mnoho špatných odpovědí na dobré?

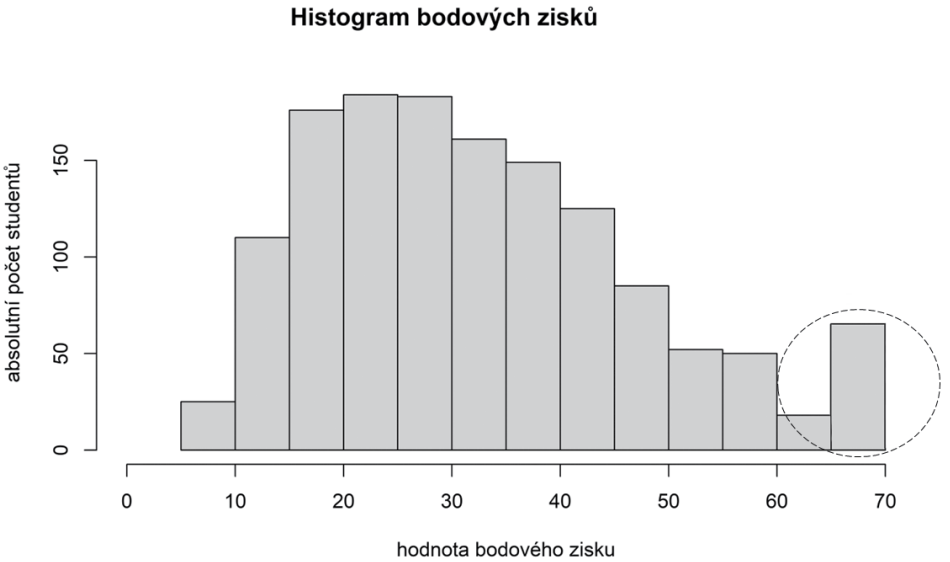
Otázky zaměřené na vztahy mezi jedinci:
- Jsou odpovědi některých účastníků nápadně podobné?
- Seděli tito účastníci poblíž sebe? Ve stejné učebně?
- Ukazuje se při porovnání tohoto vyšetřovaného s ostatními něco neobvyklého?
- Existují v jeho okolí jedinci, kteří mají téměř stejné odpovědi?

Otázky na úrovni skupiny:
- Vedou si některé školy nebo učitelé neobvykle dobře?
- Mají některá testovací centra nezvykle vysokou míru úspěšnosti a současně krátkou dobu testu?
- Jsou podobně zodpovězené testy společné pro určitou skupinu účastníků testu?
- Co je společným znakem této skupiny?
- Odpovídá nějaká skupina uchazečů výrazně lépe na otázky z jednoho profilového předmětu?
- Odpovídá nějaká skupina uchazečů výrazně lépe na otázky, které jsou nové, nebo naopak staré?
- Nebo nově zrecenzované? Zrecenzované jedním recenzentem?
- Jsou významné rozdíly mezi učebnami?
- Jsou významné rozdíly mezi uchazeči z různých kol testu?

9.3.1 Statistické indikace možného podvodného jednání

Existuje mnoho různých forenzních datových metod, které lze použít k detekci podvádění (Cizek a Wollack 2017). Statistické metody pro detekci podezřelých nesrovnalostí mohou zahrnovat:

- Hodnocení podobnosti odpovědí mezi dvojicemi zkoumaných. Nejjednodušší metody používají popisnou statistiku ke shrnutí počtu (nebo podílu) společně správných odpovědí nebo společných chyb. Například *responses in common index* (RIC) je počet otázek, na které mají dva zkoumaní stejnou odpověď. Složitější metody pracují s odhadem pravděpodobnosti, zda podobnost společných odpovědí může být ještě náhodná.
- Analýza změněných (smazaných) odpovědí sleduje počet změněných odpovědí studentů v odpovědních arších a testovacích programech. Nepravděpodobně velký počet změněných odpovědí ve třídě může naznačovat neoprávněnou manipulaci (např. hromadné opisování v při absenci dohledu). Počet změn ze špatné odpovědi na dobrou je mimořádně silným indikátorem podvodného jednání (Maynes 2013, Ranger et al. 2020).
- Analýza předpokládané vs. aktuální úspěšnosti: Statistická analýza výsledků testů z předchozího roku může předpovědět budoucí výkon. Neočekávaně úspěšné souhrnné výsledky testů mohou indikovat podvádění, zvláště když se velké zisky nezopakují v dalším roce, nebo vynesení testů, pokud se vysoká úspěšnost potvrdí i v dalších letech. Efekt zlepšování výsledků pocházejících z lepší výuky je postupnější a dlouhodobý.
- Analýza odpovědí studentů: Za podezřelé je třeba považovat, pokud studenti neodpoví na velké množství snadných otázek a současně mají správně zodpovězené nepravděpodobné množství obtížných otázek. Podobně mohou testující hledat i další statisticky významné podobnosti napříč testy.



Obr. 9.3.1 Histogram bodových zisků v testu ilustrující příklad analýzy na úrovni skupiny. V kroužku je neproporčně velká skupina mimořádně úspěšných respondentů, kteří dosáhli téměř plného počtu bodů. Dvouvrcholové rozdělení skóre vždy indikuje nehomogenní skupinu. V tomto případě by se mohlo jednat účastníky, pro které byl test příliš snadný (ale pak bychom očekávali, že rozdělení bude více „normální"). Tento průběh by proto spíše mohl odpovídat situaci, kdy omezená skupina respondentů měla předem k dispozici znění testu. Takovému „dvouvrcholovému" testu je vždy třeba věnovat zvýšenou pozornost a zkoumat, jestli další indicie případné podezření nepodporují.

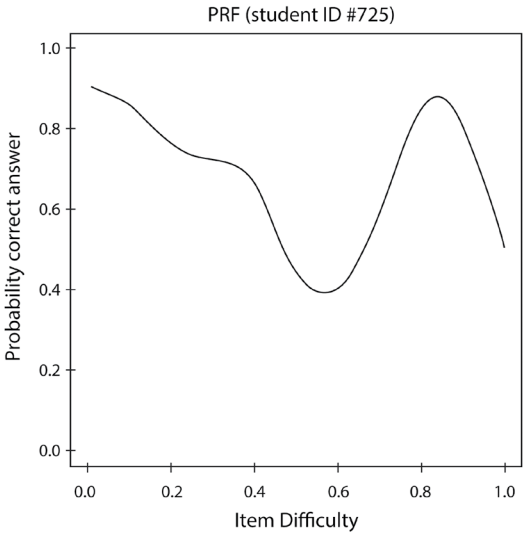
- Porovnání skóre mezi subjekty: Podezřelé je, pokud dojde k významné odlišnosti ve výsledcích u subjektů, jejichž výsledky spolu jinak vysoce korelují. Např. studenti v rámci jedné testové místnosti dosáhnou nepravděpodobně vysokého skóre v jednom předmětu.
- Neshoda mezi výsledky testů a předchozími studijními výsledky: Pokud v testech získají vysoké skóre studenti, kteří mají současně horší předchozí studijní výsledky, může to ukazovat na podvádění. Inovativní je v tomto směru přístup využívající pro detekci těchto anomálií strojového učení (Kamalov et al. 2021).

9.3.2 Nástroje pro forenzní analýzu testů

Hledáme způsob, jak z testových dat identifikovat nepravděpodobné stavy, které mohou ukazovat na případné podvádění. Uživatelsky přátelských softwarových nástrojů pro datovou forenzní analýzu není příliš mnoho.

9.3.2.1 PerFit

Jedna ze strategií je taková, že můžeme pro každého studenta vytvořit graf relativní úspěšnosti zodpovězení položek seřazených podle obtížnosti. Z logiky věci lze očekávat, že graf by měl být monotónně klesající funkcí s rostoucí obtížností položek. Významné odchylky jsou dobře rozpoznatelné. Pro tuto analýzu můžeme použít např. balíček PerFit v R (Tendeiro et al. 2016).



Obr. 9.3.2 Ilustrace využití programu PerFit pro identifikaci nepravděpodobných výsledků v testu. Pravděpodobnost správné odpovědi by měla s rostoucí obtížností klesat. Pokud tomu tak není, jako v tomto případě, je to indikace něčeho nestandardního, čemu je třeba věnovat pozornost.

Jde o uplatnění analýzy „person-fit“, která s určitou (asi 25%) senzitivitou a určitou (asi 90%) specificitou ukazuje nestandardní výstup testu pro daného studenta. Nemusí jít přímo o podvádění (kopírování nebo znalost otázek předem), může jít i o náhodné tipování třeba jen v určité části testu apod. Přestože senzitivita a specificita tohoto zkoumání nejsou samospasitelné, může jít o cenný způsob vytěžení dat, která už beztak existují.

Balíček nepotřebuje žádná externí data. Pracuje s maticí otázek a studentů, kde je jen hodnota 1 (správně) nebo 0 (nesprávně) coby dichotomické skóre položek. Nástroj sám vypočítá obtížnost položky a pravděpodobnost správného zodpovězení pro daného studenta. Výsledné grafy jsou založené na hrubých datech z daného testu, nic víc není potřeba.

Postup je dobře použitelný pro případy, kdy mají všichni stejný test, případně kdy se dají data na stejný test přepočítat (např. pokud měli všichni stejné úlohy, pouze v jiném pořadí a s proházenými možnostmi). Cesta od matice ke grafu je přímočará, stačí 2–3 řádky kódu a získáte graf pro daného studenta.

9.3.2.2 SIFT

SIFT (*Software for Investigation Fraud in Testing*) je nástroj využívající pokročilých statistických metod pro vyšetřování podvodů při testování. Poskytuje jej bezplatně (za registraci) jeden z předních dodavatelů komerčních testových systémů – společnost *Assessment Systems Corporation* (ASC). K programu je k dispozici uživatelský manuál a ukázková data, nikoli však podpora, kterou je možné si dokoupit. SIFT vypočítává řadu indexů ukazujících na různé druhy podvodů (opisování, pomoc učitele, uniklé položky a další) a může agregovat výsledky seskupením podle proměnných, jakými jsou učebna, nebo poloha testovaného v rámci této učebny apod. Podporuje všechny tři oblasti analýz – zaměřené na jedince, na relace mezi

nimi i na skupiny. SIFT poskytuje objektivně změřené statistiky pro rozhodování, ale jejich interpretace v dané situaci je na uživateli (Thompson a Sift 2016).

9.3.2.3 CopyDetect

CopyDetect (Zopluoglu, 2016) je balíček v open-source R statistickém programovacím jazyce (R Core Team, 2013), který v rámci modelu IRT i mimo něj počítá několik indexů podvádění. Je mezi nimi index Omega, zavedený Wollackem (1997), K indexy (Linden a Sotaridona 2006) a S indexy (Sotaridona a Meijer 2003). CopyDetect zpracovává najednou vždy jen jeden pár zkoumaných. Je tedy na uživateli, aby si dopsal rutinu pro zpracování větších dávek dat. U balíčků R je třeba vzít v potaz, že jde o open-source software, takže je třeba přistupovat k němu s jistou mírou opatrnosti.

Díky statistickým metodám můžeme vyslovit podezření na nepovolenou spolupráci při vyplňování testu, ale závěry bychom měli dělat opatrně. Statistické postupy by neměly být jediným důkazem opisování, zvláště pokud se používají pro obecné screeningové účely. Je sice zřejmé, že čím vyšší je shoda mezi odpověďmi, tím je pravděpodobnější, že došlo k testovacímu podvodu, ale ani vysoká míra shody není nezvratným důkazem, že je opravdu způsobená podváděním. Vždy existuje šance, že shoda v testech je (byť velmi nepravděpodobným) výsledkem poctivého vyplnění testů. Pokud naopak někdo opíše méně než 10 % položek, nejsou to statistické metody schopné odlišit od náhodných jevů.

9.4 Příklady bezpečnostních incidentů

Zdokumentované a zveřejněné příklady bezpečnostních incidentů jsou vzácné, protože ohrožují pověst instituce, jejíž procesy byly incidentem zasaženy. Instituce mají tendenci informace nezveřejňovat, a pokud tak už učiní, pak ve zcela nekonkrétní podobě, která je neužitečná těm, kdo hledají poučení. O to cennější jsou případy, kdy na veřejnost pronikl dostatek informací, aby bylo možné si učinit představu, jak k narušení integrity procesu hodnocení došlo.

Zkoušky fyzioterapeutů na Filipínách

Americká Federace státních komisí pro fyzikální terapii (FSBPT) řešila v roce 2007 problém. Při zkouškách fyzioterapeutů v detašovaném centru na Filipínách měla zřejmě část účastníků testů k dispozici otázky, které zachytili účastníci předchozích testů. Tyto otázky pak byly dalším testovaným hromadně poskytnuty, patrně samotným testovacím centrem v Manile. Testovaných byl velký počet a test byl zpoplatněný. Přinutit všechny k zopakování testu by bylo obtížné – tíže důkazního břemene by se tím přenášela i na poctivé, navíc by hrozily žaloby na ušlý zisk za zpoždění licencí. Na druhou stranu rezignovat a potvrdit podezřelé výsledky testů by znamenalo ohrozit integritu celého testování a dobré jméno FSBPT.

Za této situace se federace obrátila na společnost Caveon, která se specializuje na problematiku bezpečnosti testování a poskytuje v této oblasti širokou škálu služeb. Této společnosti, respektive její složce Caveon Data Forensics, byla poskytnuta kompletní testová data ze všech testovacích míst za poslední dva roky k forenzní analýze. K identifikaci odlišně vyplněných testů použila společnost tři nezávislé statistické ukazatele. Nejprve byl porovnán výkon v kompromitovaných testovacích otázkách (o kterých bylo známo, že unikly v testovacích přípravných centrech na Filipínách) s výkonem v nekompromitovaných testovacích

položkách. Za druhé, byla zkoumána podobnost vzorců odpovědí mezi kandidáty, přičemž vyšší stupně podobnosti naznačovaly možnost předchozí znalosti obsahu testu. Třetí analýza počítala pravděpodobnost, že se konkrétní účastník testu zúčastnil kurzu, na kterém byly distribuovány stažené položky. Kombinací spočítaných indexů bylo možné dosáhnout detekce podvodného jednání s rizikem chyby menším než 1:1 000 000. Z prověřovaných 23 500 testů tak bylo vytipováno dvacet, které měly všechny tři sledované ukazatele odchylné od normálu. Na základě toho byly zmíněné testy prohlášeny za neplatné a ostatní uznány jako platné (FSBPT 2012). Povšimněte si, jak opatrný postoj federace FSBPT a společnost Caveon zaujaly. Omezily se jen na anulování malého podílu podezřelých testů, u nichž byla jistota podvodu téměř 100 %. Podezřelý výsledek ještě nedokazuje podvod. Kumulace podezření však umožňuje vyslovit velmi relevantní závěry.

Bezpečnostní incidenty při přijímacích řízeních na českých vysokých školách

V českém vysokoškolském prostředí se v posledních desetiletích řešeno několik bezpečnostních incidentů.

Případ první

V roce 1999 byla zpochybněna integrita přijímacího řízení na Právnické fakultě UK. Protože již několik let před tím prosakovaly informace, že přijímací řízení na tuto fakultu je neférové, vyzvali novináři veřejnost ke spolupráci. V den konání náhradního kola přijímacího řízení přinesl v anonymitě zůstávající občan vzorově vyplněné přijímací testy do podatelny redakce novin Právo. Podle neověřených informací byla cena, za kterou bylo možné vypracovanou verzi koupit, sto tisíc korun. Nevypracovanou bylo možno pořídit za padesát tisíc Kč.

Fakulta v reakci přiznala, že vyřešené testy unikly na veřejnost, ale odmítla odpovědnost. Univerzita to označila za „organizovaný útok gangsterské mafie proti univerzitě“. Pověst instituce tak ohrozil nejen samotný únik testů, ale i nedbalé vyšetřování a související podezření, že na této fakultě existoval celý úplatkářský systém (Čulík 1999).

Případ druhý

O čtyři roky později, v roce 2003, došlo na stejné fakultě k problému s kvalitou testových otázek. Nepřijatí studenti nechali analyzovat podle nich nespravedlivý test a ukázalo se, že nejméně v devíti (ale spíše ve třiceti) položkách byly věcné chyby. Chyby se vyskytovaly zejména v logických testech (testy všeobecné studijní připravenosti) a v otázkách všeobecného přehledu. Aby nedošlo k poškození žádného z uchazečů o studium, musely být výsledky přepočítány. Místo 650 studentů bylo nuceně do prvního ročníku přijato o 260 studentů víc.

Po těchto zkušenostech vedení fakulty razantně změnilo způsob přijímacího řízení. Příprava otázek ze všeobecného přehledu zůstala v režii fakulty, testy a testové otázky z logiky a z všeobecných studijních předpokladů byly svěřeny agentuře Scio. Test byl bezprostředně před zkouškou distribuován přímo do jednotlivých měst, kde se zkoušky konaly – nebyl množen ani skladován na samotné fakultě (Univerzita Karlova 2005).

Případ třetí

V roce 2018 vyšlo najevo, že Lékařská fakulta Ostravské univerzity (LF OU) umožňovala již od svého vzniku (2011) obcházení výsledků přijímacího řízení a přijímání studentů, kteří ve skutečnosti neprošli testem. V roce 2018 byl do prvního ročníku například přijat student,

který v testu z profilových předmětů získal jen 43 z 90 možných bodů, ačkoliv hranice pro přijetí byla 46. Od roku 2011 do roku 2018 bylo takto „mimo pořadí" přijato každý rok kolem pěti uchazečů. Pro přijímání těchto neúspěšných uchazečů byl (zne)užit proces odvolání proti výsledku přijímacího řízení, jehož vyřizování bylo netransparentní.

Případ čtvrtý

Na 1. LF UK došlo k pokusu o podvod při přijímacím řízení v roce 2016. Při písemném testu z fyziky si pozorný učitelský dozor v testovací místnosti všiml, že uchazeč odevzdává vyplněný odpovědní formulář jiné verze testu, než měl řešit. Všechna číselná označení testové verze přitom překryl malůvkami, takže nebylo možno jednoduše zkontrolovat, kterou verzi testu původně dostal. Problém byl v tom, že ve stresové situaci před začátkem testování nepřiřadili testovaným místa dozorující učitelé, ale nechali uchazeče, aby si sedli podle svého. Jedinec bez potřebných vědomostí, ale dobře obeznámený s procedurou, byl domluven s jiným, dobře připraveným. Podvádějící účastník pak opsal celý odpovědní formulář od vedle sedícího kolegy. Nebýt pozorného dozoru, který zaregistroval neshodu verzí při odevzdávání testu, nemuselo by se na podvod vůbec přijít. Podivuhodné přitom bylo, že podvod byl vymyšlen a připraven s hlubokou znalostí procesů, podle nichž tehdy testování probíhalo. Přitom takto důkladnou znalost není možné získat na základě jednorázové individuální zkušenosti. Existuje proto podezření, že metodu připravila některá ze společností, které studentům nabízejí přípravu na přijímací zkoušky. Fakulta na odhalený pokus promptně zareagovala a změnila nejen průběh testového dne, ale individualizovala i testový sešit a odpovědní arch a přidala tuto zkušenost do školení pro dohlížející učitele.

Případ masivního zpochybnění výsledků přijímacího řízení v USA

V roce 2019 propukl ve Spojených státech skandál, když se ukázalo, že firma zabývající se oficiálně poradenstvím při přijímaní na vysoké školy ve skutečnosti od roku 2011 organizovala podvody, přičemž za úplatky v hodnotě 25 mil. dolarů pomohla asi 750 studentům při přijetí na celkem 11 elitních vysokých škol. Organizátor podvodů (William Singer) uplácel psychology, aby vystavili lékařské dobrozdání o zdravotním hendikepu uchazeče, které by mu přineslo více času pro vyplnění testového formuláře (potvrzení přišlo na 4000–5000 USD). Nejméně dvě testová centra evidentně spolupracovala s organizátory podvodů. Přinejmenším ve 20 případech došlo k podvodu pomocí záměny identit, kdy testovaného zastoupil vysoce kompetentní náhradník. Druhou metodou ovlivňování výsledků přijímacího řízení bylo opatřování falešných dokladů o provozování vrcholového sportu, ke kterému se při přijímání studentů na vysoké školy v USA přihlíží. V případu bylo obviněno 53 osob. O skandálu byl natočen dokumentární film *Operation Varsity Blues: The College Admissions Scandal* (2020). Skandál poukázal nejen na mezery v zajištění spravedlivého přijímaní uchazečů, ale i na zvláštní roli prestižních univerzit, jejichž absolvováním získá student nejen vzdělání, ale i styky a společenský status nezbytný pro průnik do nejvyšších pater finančně lukrativních oborů, jako jsou právo a finance (Winter 2019; Baker 2019).

Shrnutí

Ukazuje se, že největší riziko představují ti, kdo mají buď moc, nebo informace. Koncentrovaná znalost přijímacího řízení (například ve firmách připravujících studenty na přijímací řízení) svádí k zneužití těchto vědomostí k obcházení systému.

9.5 Prevence podvodného jednání

Pro strukturování úvah o faktorech ovlivňujících pravděpodobnost podvádění můžeme použít „trojúhelník podvodu“ (*fraud triangle*), model často používaný k posouzení pravděpodobnosti etického selhání v různých oblastech lidského konání.

Model vznikl na základě hypotézy, že důvěryhodné osoby se dopouštějí nečestného jednání, pokud se domnívají, že jsou v bezvýchodné situaci, a současně mají příležitost vyřešit tuto situaci porušením pravidel a dokážou si své jednání nějak obhájit před sebou samými (Cressey 1953).

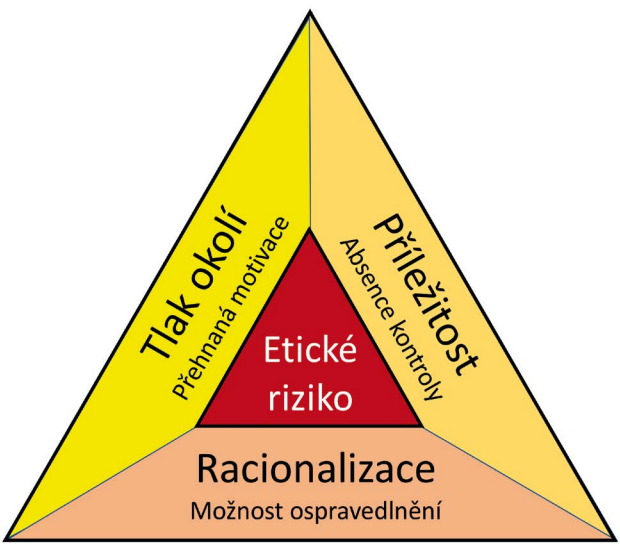
Identifikované faktory ovlivňující nepoctivé jednání jsou:

- tlak,
- příležitost,
- racionalizace.

Podívejme se nyní tyto oblasti podrobněji, abychom pochopili důvody, proč studenti podvádějí (Simmons 2018). Pokud porozumíme **motivaci k podvodnému jednání**, můžeme se pokusit ji snížit (Foltýnek 2021).

Tlak

Tlak je vliv okolí, nebo zprostředkovaně vlastní psychiky, na dosažení nereálných cílů. Jedinec má pocit bezvýchodné situace, což ho „nutí“ sáhnout po nekorektním řešení.



Obr. 9.5.1 „Trojúhelník podvodu“ je vizualizací faktorů, které společně mohou vést podvodnému jednání. Jsou to: **vnější tlak**, případně ambice přesahující schopnosti, **příležitost**, nebo obecně nedostatek kontroly a **racionalizace**, tedy možnost obhajoby nečestného jednání před sebou samotným.

Asi nejčastěji uváděnou motivací k podvádění je snaha dosáhnout lepšího hodnocení, než by odpovídalo skutečně nabytým znalostem a dovednostem (Carlton a Krou 2021). Může to být vyvoláno tím, že známky jsou z *nástroje* hodnocení povýšeny na *cíl* učení. Často vzniká tlak, aby student měl „dobré známky" bez ohledu na to, co se doopravdy naučí. Takový tlak mohou vytvářet např. rodiče, spolužáci nebo stipendijní řád. V nižších stupních vzdělávání dokonce studijní průměr může mít zásadní dopad na další osud studenta – může na něm záviset třeba přijetí na střední nebo vysokou školu. V okamžiku, kdy je cílem studenta získat dobrou známku, stává se podvádění logicky jednou z možných cest, jak tohoto cíle dosáhnout.

Skutečným cílem vysokoškolského vzdělávání je získání dovedností a kompetencí pro nějaké povolání, práci či roli. Hodnocení je nástroj, který „jen" měří, do jaké míry se tohoto cíle daří dosáhnout. K získání znalostí a dovedností podvádění stěží pomůže. Podvodnému jednání tak můžeme bránit snížením tlaku na známky jako takové, a naopak jasným definováním cílů výuky. Je třeba, aby studenti porozuměli, proč se mají konkrétní znalosti a dovednosti naučit a k čemu jim budou dobré v praxi. Cíle výuky bychom jim měli srozumitelnou formou sdělovat a měli bychom je motivovat k jejich dosažení – nikoli k dosažení dobrého hodnocení. Studentům musí být zřejmé, že se učí pro sebe, nikoli pro uzavření předmětu.

Zdá se, že i přehnaná vnitřní motivace může vytvářet podobný tlak jako ambiciózní rodinné zázemí. Klíčový je zřejmě nesoulad mezi skutečnými výsledky a očekáváními, která mají buď sami studenti, nebo která jsou na ně kladena okolím. Tendence k podvádění se tak překvapivě zvyšuje u nejlepších (nejvíce motivovaných) (McCabe et al. 2012).

Tlak může vytvářet i nedostatek času, nebo pocit nedostatku času, na zvládnutí látky. Proto je tak zásadní komunikovat se studenty, dbát na to, aby chápali výukové cíle a aby dokázali odhadnout čas potřebný pro přípravu na zkoušku. O vlivu pocitu „nedostatku času" svědčí i práce, které prokázaly zvýšenou tendenci k podvádění u studentů, kteří jsou více zapojeni do mimoškolních aktivit (Ma et al. 2013). Diskutuje se ovšem, zda v pozadí těchto případů není spíše tendence napodobovat „úspěšné" vzory a usnadnit si tak racionalizaci podvodného jednání (Vowell a Chen 2004).

Je také vhodné pomoci studentům porozumět tlakům okolí, které je ovlivňují, a racionálně zhodnotit jejich význam. Stanovení adekvátních cílů a vytvoření odpovídajících hodnotových žebříčků pomůže studentům k nacházení správné motivace.

Příležitost

Snižování příležitostí k podvodnému jednání u testu či zkoušky může podpořit pečlivý pedagogický dohled. Ze studií vyplývá, že sklon k podvádění se značně sníží, pokud jsou si studenti vědomi, že zkouška je dozorovaná (Dyer et al. 2020). Ochotu podvádět snižuje i nastavení případné sankce tak, aby další studenty od podvodného jednání odradilo.

Příležitosti k podvádění snižuje i kvalitní organizace testu. Pokud učitel určí zasedací pořádek u zkoušky, je riziko opisování nižší, než když si studenti sami mohou vybrat, vedle koho si sednou. Stejně tak se dá riziko manipulace s výsledky testu snížit tím, že test je až do vyhodnocení anonymní. Teprve po přidělení bodů se testové formuláře opět propojí s identifikací osob. Promyšlenost a dobré uspořádání testového procesu může prostor k nežádoucím aktivitám výrazně omezit.

Racionalizace

Podvádějící jedinec se snaží najít racionální zdůvodnění pro své chování. Pokud si může v duchu říci, že škola s ním taky nejedná poctivě, je pro něj snazší si odůvodnit svoje neetické chování. Mezi často uváděné důvody zvyšující sklon k podvádění patří zkoušení z témat, jež jsou z pohledu studentů zbytečná a okrajová (Burnett et al. 2016). Studenti to vnímají jako „nepoctivé" jednání ze strany školy a cítí se oprávněni podvádět taky.

„Podvádění je přeci běžné:" Někteří studenti uvádějí, že u zkoušek nemají zábrany podvádět, neboť to „dělá každý" (Kennedy 2019). Nemají pocit, že by dělali něco špatného a nevnímají společenskou nebezpečnost podvádění.

Učitel by měl dát jasně najevo, že podvádění u zkoušek považuje za nepřijatelné. Součástí vzdělávání by měla být i *metakognice* – pochopení, jak smýšlím a proč. Součástí metakognice je i schopnost odhadnout vlastní možnosti. Metakognice pomůže stanovit cíle, posílit motivaci k učení, upevnit akademickou integritu a morální zásady. Je tedy potřeba se studenty průběžně a v rámci mnoha různých předmětů mluvit o tom, jak a proč studují a čeho chtějí dosáhnout.

Další faktory

Pomoc slabšímu: Někteří studenti dávají u zkoušek a testů opisovat jiným, nebo jim jinak nedovoleně pomáhají. Tím se sami stávají účastníky podvodného jednání.

Mimo zkoušky je pomoc slabšímu společensky oceňovaná. Ani u zkoušek není tento druh nedovoleného jednání společností vnímán jednoznačně jako vysloveně nežádoucí. Problémem je, že zkoušky a testy jsou ve většině případů individuální. Většina vysokoškolských studentů se přitom připravuje pro profese, které jsou týmové.

Nedovolená pomoc jinému při zkoušce přestane dávat smysl v okamžiku, kdy se namísto čistě individuálního hodnocení začnou znalosti a dovednosti hodnotit v rámci týmové spolupráce. Taková forma hodnocení navíc může pomoci studenty lépe připravit pro praxi. Problémem ovšem zůstává, že vzdělávací systém požaduje, abychom nakonec i takové hodnocení „rozebrali" na hodnocení jednotlivců. Dokonce je nutné, abychom hodnocení jednotlivce očistili od vlivu ostatních členů týmu. Přesto by se hodnocení týmové práce mělo stát pravidelnou součástí jak formativního testování, tak i praktického zkoušení.

Podvádění je výhodné: Někteří studenti se k podvádění rozhodnou, protože to považují za výhodnější než investovat do přípravy na zkoušku. Jiní přicházejí ke zkoušce s vědomím nebo obavou, že nejsou dostatečně připraveni, a podvádění se jeví jako únosná strategie k „vyřešení" situace (University of North Texas nedatováno).

V obou případech motivaci k podvádění posiluje rozdělení vzdělávání na fázi, kdy se student učí, a na fázi, kdy získává zpětnou vazbu a je hodnocen.

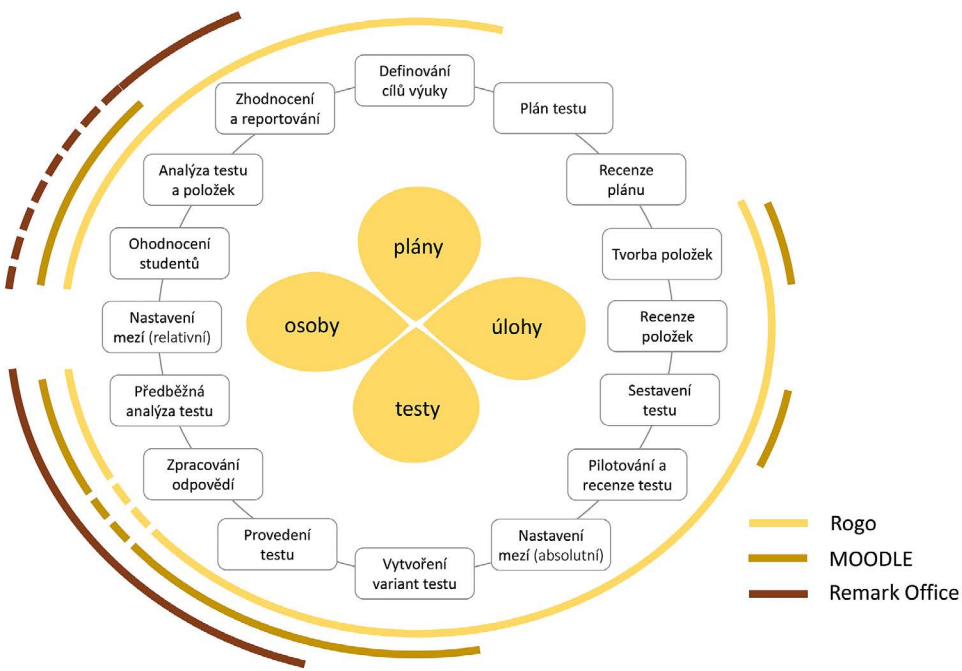
Studenti nemají tendenci podvádět, pokud se cítí dobře připraveni. Stejně tak motivace k podvádění klesá, pokud zkoušený do přípravy investoval hodně času a úsilí. Jako prevence podvodného jednání se tak jeví větší důraz na aktivní učení a osvojování dovedností během výuky, současně s častým poskytováním zpětné vazby. Zkouška by neměla být izolovaným aktem na konci kurzu. Výhodnější je, pokud v průběhu celé výuky student prochází velkým počtem

dílčích formativních testování. Intenzivní zpětná vazba pomáhá v motivaci k učení. Před závěrečnou sumativní zkouškou navíc student má dobrou představu, co očekávat, a dokáže realisticky odhadnout, jakou má naději splnit podmínky. Snižuje se tím testová úzkost, odstraní se někdy i zbytečné obavy z neúspěchu a tendence k podvádění klesá.

10 NÁSTROJE PRO TESTOVÁNÍ A ANALÝZY

10.1 Software pro testování

Nástrojů pro podporu formativního i sumativního testování jsou desítky. Průběžně vznikají nové, některé naopak ztrácejí na významu a jsou opouštěny. Stávající jsou pak součástí intenzivního vývoje, zejména směrem k mobilním platformám. V této kapitole, zaměřené na testové nástroje, se nebudeme pokoušet o ucelený přehled dostupných produktů, neboť statický formát této publikace nedovoluje zachytit dynamický vývoj, ale omezíme se na zprostředkování našich zkušeností s těmi nástroji, které se nám osvědčily, nebo naopak neosvědčily.



Obr. 10.1.1 Pokrytí testového cyklu testovými nástroji Rogō, Moodle a Remark Office

10.1.1 Rogō

Mezi nástroji pro elektronické testování zaujímá význačné místo program **Rogō**. Mimořádné postavení tohoto systému spočívá v tom, že je to **kvalitní**, **bezpečný** a snadno **použitelný** testovací nástroj, který je přitom **volně šiřitelný**.

Výhody (kvalita)

- Každá úloha má vlastní záznam o změnách a použití v testech.
- Přizvaní externí spolupracovníci mají snadný a bezpečný přístup k recenzím.
- Při distančním testování je možné nastavit ukončení testu po vyčerpání časové dotace.
- Položky použité v testu se uzamknou.
- U právě probíhajícího testu nelze přidávat ani odebírat položky.
- Možnost nastavení časového limitu testu.
- Odpovědi studentů lze analyzovat a jednoduše zhodnotit vlastnosti položek.
- Hodnocení lze dle výkonu kohorty upravit po skončení testu pomocí Hofsteeho metody.
- Pro nastavení hodnocení testu jsou k dispozici rovněž modifikovaná Angoffova a Ebellova metoda.

Bezpečnost

- K připojení se používá výhradně protokol HTTPS. Data jsou šifrována pomocí 256bitového SSL.
- Bezproblémové a bezpečné sdílení materiálů v rámci týmů zaměstnanců pro společnou práci na hodnocení.
- Přizpůsobitelné kontroly a váhy pro zajištění spravedlivého hodnocení pro všechny uživatele.
- Pro přístup studentů a učitelů umí využívat zavedené místní autentizační systémy.
- Při výpadku internetového připojení se ztratí jen odpověď na aktuálně otevřené stránce testu, nikoliv celý test.
- Přístup k testu se dá omezit na vybrané IP adresy.

Použitelnost

- Může běžet na serverech Windows i Linux.
- Kompatibilita s LDAP – pro uživatele není potřeba vytvářet další přihlašovací údaje.
- Podpora jazykových mutací.
- Systémy nápovědy na míru – samostatná nápověda pro uživatele podle rolí.
- Rogō je webová aplikace, která běží na všech hlavních prohlížečích – Chrome, Edge, Firefox, Safari, Internet Explorer.
- Adeptům se speciálními potřebami lze upravit jak vzhled testu, tak i časovou dotaci.

Licence

Z pohledu licence je webová aplikace pro on-line testování Rogō svobodně šiřitelný *open source* program, uvolněný pod GPL verze 3.0. Je tedy možné kód měnit, rozšiřovat jej a přispívat tak k projektu. V praxi to funguje tak, že požadavky na úpravy kódu, ať už se týkají nahlášení problémů, nebo jde o návrhy na nové funkcionality, se zapisují do fronty požadavků a komunita je postupně řeší.

Historie

Testovací systém byl od roku 2003 vyvíjen na lékařské fakultě University of Nottingham pod názvem „TouchStone“ („prubířský kámen“). Systém Rogō vznikl na lékařské fakultě a je velmi dobře přizpůsoben zvláště pro výuku medicíny. Mimo jiné dovoluje použít v úlohách interaktivní obrázky, na nichž student myší vyznačí hledaný objekt. Systém pak vyhodnotí, zda hledanou strukturu označil s požadovanou přesností.

Po úspěchu na domovské fakultě byl rozšířen na celou univerzitu, převeden na software s otevřeným zdrojovým kódem a při té příležitosti i přejmenován, aby nedocházelo k záměnám s jinými systémy. *Rogō* v latině znamená „ptám se“ (Wilson 2012). Komunita zabývající se vývojem Rogō získala finanční podporu společnosti JISC, která umožnila další rozvoj systému, mimo jiné možnost překladu do národních jazyků. V ČR je systém Rogō nainstalován (na serverech 1. LF UK) na adrese https://www.rogo.cz/ a kromě 1. LF UK slouží i dalším fakultám UK. 1. LF UK připravila český překlad prostředí a pracuje kontinuálně i na lokalizaci nápovědy. Díky podpoře LDAP jsou studenti do systému importování automaticky ze SIS (Studijní informační systém UK) a mohou se autentifikovat svým CAS účtem (Centrální autentizační systém UK).

Specifické vlastnosti

Na rozdíl od jiných programů Rogō pokrývá a podporuje mnoho kroků cyklu tvorby testů, od spolupráce při přípravě testových úloh, přes jejich oponování z hlediska obtížnosti a relevance, tvorbu plánu testu, standardizaci, až po vyhodnocení kvality otázek. Takto komplexní řešení přináší podstatné výhody. Např. k oponentuře testových úloh je vhodné přizvat řadu vlastních i externích odborníků, což bývá obvykle časově a organizačně náročné. Pokud přitom návrhy položek kolují mezi větším počtem lidí, je velmi obtížné zajistit jejich utajení. V Rogō jsou naopak oponenti vyzváni k připojení do systému, takže testové úlohy systém vůbec neopustí. Komentáře a připomínky se opět vkládají přímo do Rogō a autoři úloh na ně mohou ihned reagovat. Poté, co test proběhl, je možné zobrazit popisné charakteristiky, histogram celkových skóre všech studentů, nebo obtížnosti položek. Rogō automaticky vypočítává diskriminační indexy pro každou položku testu, což umožňuje identifikovat špatně sestavené úlohy a vyloučit je z dalšího používání. **Z hlediska uplatňování moderních postupů v testování je systém zcela unikátní a jeho zavedení podporuje rozšíření správné testové praxe do terénu.**

Systém umožňuje distribuovat jak papírové, tak i on-line testy, a to pro sebehodnocení i pro zabezpečené sumativní testování (Baylem et al. 2011).

Rogō obsahuje nástroje umožňující automaticky importovat do systému studenty a předměty, které mají zapsány. Studenti se pak díky podpoře adresářové služby LDAP, kterou používá celá Univerzita Karlova, mohou rovnou přihlásit svým CAS účtem a v Rogō jsou přiřazeni ke všem svým předmětům.

Systém umožňuje pedagogům vytvářet řadu typů testů a průzkumů:

- formativní hodnocení,
- sumativní hodnocení,
- testy pokroku,
- průzkumy (dotazníky),
- e-OBSE (klinické zkoušení),

- offline testy,
- vzájemné hodnocení (studentské).

Každý z těchto typů testů a průzkumů může používat řadu forem položek:
- vymezení plochy,
- dichotomické úlohy,
- úlohy s výběrem odpovědi,
- extended Matching,
- multiple true false,
- doplňování textu,
- označení bodů v obrazu,
- Likertova škála,
- test shody scénářů,
- textová pole.

Výhody:

Tab. 10.1.1 Výhody systému Rogō

● Nízké náklady na pořízení. ● Podpora celého procesu testování. ● Podpora týmové spolupráce. ● Vysoká úroveň zabezpečení. ● Velký výběr typů testových úloh včetně multimediálních.

Tab. 10.1.2 Nevýhody systému Rogō

● Menší komunita udržující a rozvíjející systém. ● Potřeba lokální podpory a administrace. ● Ne zcela intuitivní ovládání. ● Časová náročnost při osvojování programu.

10.1.2 Moodle

Systém pro řízení výuky Moodle je celosvětově rozšířené online výukové prostředí. Tuto open-source platformu využívá více než 250 milionů studenů. Moodle je dnes asi nejrozšířenější systém pro řízení výuky na vysokých školách. Vznikl v roce 2002 a průběžně se aktualizuje. Svým otevřeným kódem, zabezpečením a ochranou osobních údajů představuje pro mnoho vysokých škol a univerzit lákavé řešení. Na vývoji Moodle spolupracuje rozsáhlá a aktivní komunita. Služby, které přesahují možnosti jednotlivých správců, nabízejí specializované firmy s kvalifikací *Moodle partner*. Jedná se zejména o služby, jako je hosting, přizpůsobení, podpora, školení nebo i komplexní správa celých projektů v Moodle.

Výhody:
- cena,
- rozšíření,

- komunita vývojářů,
- flexibilita díky mnoha modulům,
- možnosti integrace,
- mobilní i PC rozhraní,
- podpora LDAP.

Nevýhody:
- nepřehlednost,
- chaotické uživatelské rozhraní,
- příliš mnoho modulů,
- absence vedení a jednotné koncepce,
- serverová náročnost při souběžné práci více uživatelů,
- upgrady mohou znehodnotit předchozí práci,
- z pohledu moderních mobilních aplikací jde o „moloch“.

Zásadní předností Moodle je jeho rozšíření a flexibilita. Na druhou stranu je potřeba mít na paměti, že při velkém počtu současných přístupů (typická situace pro testování) se systém může zahltit (Korviny et al. 2009). Řešení je ve vhodném dimenzování infrastruktury, např. rozložení zátěže na více serverů (Korviny a Foltyn 2012).

Výhody pro testování

LMS Moodle není vysloveně zaměřen na testování, ale přesto poskytuje některé zajímavé možnosti. Pokud jej používáte při proktorovaném testování, je možné využít zabezpečený prohlížeč Safe Exam Browser. Pro adaptivní testování je k dispozici modul Ada Quiz. Adaptivní kvíz vede studenta itinerářem otázek, přizpůsobeným právě jeho znalostem.
- více typů úloh,
- míchání úloh,
- časový limit testu,
- automatická klasifikace,
- lokalizace,
- dobré zabezpečení.

Nevýhody pro testování

Z pohledu položkového bankovnictví je škoda, že Moodle neshromažďuje u úloh informace o jejich fungování v předchozích použitích. To by umožňovalo používat Moodle více v roli položkové banky. Zvláštní je, že statistiky v analýzách testů mají svébytné názvosloví, takže učitel občas musí experimentálně zjišťovat, co kterým označením autoři Moodle mysleli.
- malá podpora týmové práce,
- nutnost vyškolení učitelů,
- mnoho jiných funkcionalit mimo testování,
- neintuitivní rozhraní,
- pracnost přípravy testu.

Nicméně i přes tyto výhrady je Moodle hojně používaný nástroj, výtečně využitelný i pro testování. Testové úlohy lze připravit i mimo prostředí Moodle, nebo je možné je převzít z položkové banky. Moodle podporuje standardy interoperability QTI a řadu importních formátů.

10.1.3 Remark Office

Testování velkých skupin studentů se často provádí pomocí „tužky a papíru“. Výhodou je nezávislost na technických prostředcích a velmi dobrá prokazatelnost v případě sporu. Nicméně vyhodnocování odevzdaných odpovědních formulářů může být úzkým hrdlem této technologie. Zvláště v případě velkého počtu testovaných a velkého významu zkoušky je třeba zajistit, aby byl proces rychlý, pokud možno bezchybný, prokazatelný a reprodukovatelný. Vyhodnocování pomocí průsvitky těchto parametrů nedosahuje. Používají se proto programy pro optické rozpoznávání znaků (*optical mark recognition*, OMR).

Jedním z takových programů je Remark Office OMR. Tento program slouží k rozpoznávání naskenovaných odpovědních formulářů, dotazníků či testů a jejich převedení do elektronické podoby.

Program dovoluje načítat data přímo ze skeneru nebo z uložených souborů. Sebraná data mohou být exportována do různých datových formátů nebo přímo zpracována. Program dále umožňuje tvorbu reportů nad sebranými daty. Seznam dostupných statistik a druhů reportů lze nalézt na webu výrobce (www.gravic.com).

Předem musí být načten prázdný formulář – šablona, na kterém jsou označena místa pro optické čtení. Šablona definuje typ proměnné, název a popis proměnné apod. Software rozpozná začerněné kolečko, prázdné kolečko, a dokonce i začerněné, ale později přeškrtnuté kolečko. Poradí si s nevyplněnými odpověďmi, umazanými a poškozenými formuláři, vícenásobnými značkami a dalšími anomáliemi. Remark čte čárové kódy a umí rozpoznávat text (ORC). U rukou psaného textu dokáže pole uložit jako grafický soubor k dalšímu zpracování.

Pokud automatika narazí na případ, kdy si neví rady, poskytne obsluze zvětšený pohled na příslušné místo formuláře a vyžádá si pomoc. Je to uživatelsky velmi přátelské a nerozpoznaných případů je minimum. S mírnou nadsázkou firma razí heslo: Když to dokážete přečíst vy, dokážeme to přečíst taky.

Software je velmi užitečný nejen pro vyhodnocování testů a anket, ale např. pro akademické volby a jiné příležitosti, kdy je třeba zpracovat velké množství papírových formulářů.

Trvalá licence stojí cca 30 tis. Kč a roční podpora řádově 20 % z této ceny. Vzhledem k funkčnosti se cena nezdá být přehnaná. Nepraktické se může jevit fixování licence na konkrétní počítač.

10.1.4 Socrative

Socrative (https://www.socrative.com) je nástroj pro online testy, kvízy a průzkumy. Původně sloužil především pro frontální výuku, v níž nahrazoval dříve používaná „hlasovátka“. Socrative umožňuje vytvářet kvízy s výběrovými úlohami (pravda/nepravda, s jedinou nejlepší odpovědí i multiple true-false) i úlohy s krátkou tvořenou odpovědí. Studenti odpovídají pomocí mobilních telefonů, učitel řídí kvíz z počítače nebo také z mobilního telefonu.

Předností Socrative je kvalitně zpracované rozhraní pro učitele, které na jedné straně umožňuje poměrně velkou variabilitu testů a kvízů a jejich snadné přizpůsobení tomu, co je v danou chvíli potřeba, na druhou stranu je ale jednoduché a přehledné. Učiteli se v něm dobře orientuje a práce se Socrative neodvádí nežádoucím způsobem pozornost od výkladu nebo komunikace se studenty.

Po skončení kvízu je možné úlohy automaticky obodovat, ohodnocení úloh s krátkou tvořenou odpovědí může učitel upravit nebo takové úlohy bodovat zcela ručně. K dispozici jsou pak vizualizace, jak studenti odpovídali, které se opět snadno využijí při další práci se studenty. Je také možné stáhnout několik forem reportů, od přehledu odpovědí pro každého studenta až po tabulky s výsledky pro učitele.

Kromě celých kvízů je možné studentům pokládat i jednotlivé otázky. Zajímavý nástroj je k dispozici pro otázky s tvořenou odpovědí, kde učitel může nejprve nechat studenty, aby napsali své odpovědi, a pak je v druhém kroku může nechat hlasovat o tom, která odpověď je nejlepší. Mezi další funkce Socrative patří zpětnovazebný dotazník na konci lekce.

Základní verze Socrative je zdarma. Kvízy je možné používat opakovaně, vytvářet jejich nové verze a ukládat je v různých složkách. Placená verze umožňuje vytvořit více „místností“, což je užitečné, pokud jeden učitel využívá tento nástroj pro více různých kurzů a má v něm větší množství kvízů.

10.1.5 Kahoot!

Kahoot! je aplikace pro vytváření kvízů, která se používá především v nižších stupních vzdělávání. Poskytuje hravé, stimulující prostředí, které umožňuje vytvářet atraktivní soutěže. Žáci nebo studenti opět odpovídají pomocí mobilního telefonu. Kromě kvízů pro jednotlivce podporuje Kahoot! i soutěže mezi týmy. Vytvořené kvízy lze sdílet a na internetu se dá najít mnoho různých soutěží, testů a kvízů v Kahoot! z nejrůznějších oblastí.

10.1.6 Mentimeter

Jiným nástrojem, ve kterém učitel pokládá posluchačům otázky a nechává je odpovídat pomocí jejich mobilního telefonu nebo notebooku, je Mentimeter. Na rozdíl od předchozích není vhodný pro testování, ani v něm není možné pohodlně sledovat, kterou odpověď napsal který student. Mentimeter se hodí především pro pokládání postojových otázek a pro stimulaci diskuse na určité téma. Obsahuje řadu možností, jak vizualizovat odpovědi. Nejčastěji se využívá *world cloud*. Posluchači jsou vyzváni, aby odpovídali jednotlivými slovy nebo slovními spojeními (lze nastavit, kolik odpovědí může jeden student odeslat). Výsledek se zobrazí jako „oblak“, ve kterém je slovo nebo slovní spojení tím větší, čím častěji se v odpovědích vyskytuje. Využívají se také průzkumy pomocí různých typů stupnic, jako jsou Likertovy škály, jejichž výsledky se opět přehledně vizualizují.

Ve verzi dostupné zdarma je omezený počet otázek, které je možné použít v rámci jedné prezentace. Lze ale vytvářet větší počet prezentací, již vytvořené průzkumy je možné používat opakovaně, výsledky průzkumů lze stáhnout a dále s nimi pracovat.

10.1.7 Interoperabilita testových nástrojů

Uvážíme-li situaci, kdy tvorba testových úloh může probíhat v jednom prostředí (např. v položkové bance) a administrace testu v prostředí jiném, je důležité zajistit přenositelnost testových úloh mezi platformami. Jednoduché formáty výměny podporují často přenos jen mezi několika určitými programy, nebo podporují jen několik málo formátů testových úloh. Na druhou stranu jsou tyto formáty přehledné a pochopitelné (např. Aiken). Na opačném konci pomyslné stupnice komplexnosti stojí všeobecně přijímané standardy interoperability výukových systémů, z nichž nejpoužívanější je QTI.

Standard QTI

Question and Test Interoperability (QTI) je otevřený standard pro výměnu testových úloh, který vytvořilo IMS Global Learning Consortium. K vytvoření standardu výměny otázek QTI vedla snaha zabránit znehodnocení práce, která byla vložena do přípravy úloh, při změně technologie testování. QTI je založen na formátu XML a definuje formáty a protokoly interoperability testů, od papírových, přes digitální, adaptivní až k proktorovaným. Vývojáři potom tyto standardy integrují do svých řešení (Boussakuk et al. 2021).

Pro integraci testovacích nástrojů s výukovým prostředím vyvíjí konsorcium IMS i standardy výměny dat s e-learningovými nástroji LTI (*learning tools interoperability*). Standardy LTI umožňují přenos dat, např. známek z testovacího programu do výukového prostředí. LTI doplňuje QTI tím, že poskytuje způsob, jak integrovat testovací systém s výukovou platformou, jako je LMS, nebo studijním informačním systémem (IMS Global nedatováno).

10.2 Software pro analýzy testů

Analýza testu a její reportování patří k důležitým krokům v procesu testování. Při položkové a testové analýze můžeme zjistit, jak se chová náš test jako celek a jaké jsou vlastnosti jednotlivých položek. Díky této zpětné vazbě můžeme test pro další kola korigovat a vylepšovat. Komerčních nástrojů pro analýzu testů a položek je mnoho desítek (Wikipedia 2001). Volně dostupných nástrojů je výrazně méně (Nelson 2017). Některé moduly analytických nástrojů mohou být obsaženy přímo v programech pro testování (Rogō), či systémech pro správu výuky (Moodle), ale pro opravdu důkladnou analýzu je třeba použít specializované nástroje, či statistické prostředí s příslušnými knihovnami.

Specializované komerční programy bývají uživatelsky přátelské, velmi sofistikované a poměrně nákladné. Volně dostupná nekomerční řešení mají zase většinou vysoký práh obtížnosti.

Protože předpokládáme, že naši čtenáři se rekrutují spíše z akademických kruhů, kde je vyžadována vysoká kvalita analýzy a je snáze dostupná mentální kapacita než finanční zdroje, začneme výtečným analytickým nástrojem – statistickým software R. Ti, kdo nemají problém se zadáváním příkazů z příkazové řádky, mohou použít některý z mnoha balíčků v „R knihovně“ CRAN zaměřený na oblast „Psychometric Models and Methods“. Pro ty, kdo mají přece jen raději více uživatelsky přátelská řešení, je tu webová aplikace ShinyItemAnalysis odvozená od stejnojmenného balíčku v knihovně R.

ShinyItemAnalysis

Na webu volně dostupná aplikace ShinyItemAnalysis od Patrície Martinkové a jejích kolegů byla původně vytvořena pro analýzu přijímacích testů na vysoké školy. Nově nabízí i širokou škálu dalších analýz v oblasti didaktických a psychologických měření (Martinková et al. 2021). Umožňuje provádět testové a položkové analýzy včetně grafických výstupů (distraktorovou analýzu, dvoubarevné DD grafy, ...). Pomocí přednahraných dat si můžete vyzkoušet analýzy na ukázkových datech. Nabízených metod je hodně a v relaci k tomu je nápověda trochu stručná. Nicméně, když víte, co potřebujete, tak to zas takový problém není.

Ne zcela bezprahové je nahrávání dat do systému. Pokud máte chybu ve formátu dat, nedostanete žádnou zprávu o příčině selhání. Můžete tak strávit dost času, než problém odhalíte, ale nenechte se odradit, podruhé to již půjde lépe. Jinak je to totiž opravdu jedinečný nástroj. Najdete jej na adrese http://www.shinyitemanalysis.org/

jMetrik

jMetrik je bezplatný psychometrický software s otevřeným zdrojovým kódem. Byl vyvinut J. Patrickem Meyerem na University of Virginia. Nabízené psychometrické metody zahrnují klasickou analýzu položek, odhad spolehlivosti, škálování testů, diferenciální fungování položek, teorii odpovědi na položku, Raschovy modely a další. K programu je obsáhlá nápověda. Nicméně podle některých autorů je jMetrik poněkud těžkopádný (Nelson 2017). Nově je dostupný samostatný modul IRT illustrator, který umožňuje vykreslovat různé funkce teorie odpovědi na položku (IRT). jMetrik i IRT illustrator jsou čistě Javová aplikace, fungující na všech operačních systémech, které mají aktuální verzi javy. Více informací a software samotný najdou čtenáři na adrese https://itemanalysis.com/

Z komerčních analytických nástrojů si dovolíme zmínit tři – Lertap, Iteman a Xcalibre, které jsme měli možnost vyzkoušet.

Lertap

V Austrálii působící odborník na psychometrii Larry Nelson vytvořil řadu programů pro analýzy testů. Poslední z této řady LERTAP5 je komplexní softwarový balík pro analýzy testů, využívající Microsoft Excel. Počítá analýzu výsledků testů, položek, včetně grafických výstupů. Nabízí nástroje pro detekci podvádění. I je Lertap5 spíše orientován na metody klasické teorie testů (CTT), nabízí také základní Raschovy analýzy pro dichotomické testové položky (Nelson 2017). Výpočty nejsou nejrychlejší, což je způsobeno prostředím Excelu, v němž probíhají. Volně dostupná verze obsahuje všechny funkce, ale nezpracuje více než 250 datových záznamů. Trvalá licence (vázaná na počítač) stojí 78 dolarů, čím se tento produkt řadí na pomezí mezi komerčními a nekomerčními nástroji.

Iteman

Program Iteman je zajímavý komerční software pro analýzu položek a testů pomocí klasické teorie testů (CTT). Unikátní je tím, že vytváří obsáhlé a profesionálně zpracované zprávy ve formátu Microsoft Word o kvalitě testových položek, o testu jako celku a o jeho psychometrických vlastnostech, a to včetně vložené grafiky a tabulek. Popis jedné ze starších verzí přináší Byčkovský a Marková (2003). Iteman je nyní (ve verzi 4) k dispozici buď jako

cloudová verze, nebo aplikace pro Windows. Cloudová verze umožňuje používat software kdekoli. Verze pro Windows zase umožňuje uchovávat všechna (potenciálně citlivá) data na jednom počítači. Nekomerční licence (pro akademickou sféru) programu Iteman stojí 1295 USD ročně. Ukázková verze má omezení na 100 studentů a 100 položek. Více informací, popis aktuální verze a licencování najde čtenář na stránkách výrobce http://www.assess.com/.

Xcalibre

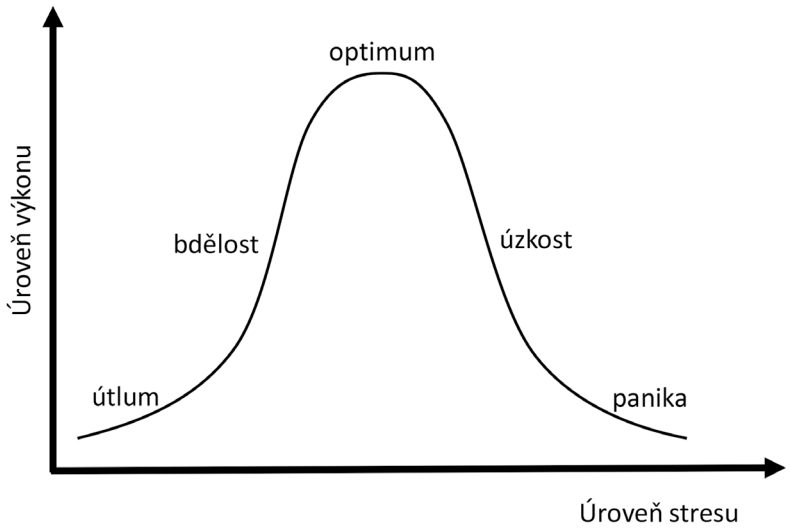
Xcalibre 4, od stejného výrobce, představuje výkonný nástroj pro analýzu testů založený na teorii odpovědi na položku (IRT). Program má velmi uživatelsky přátelské rozhraní. Poskytuje profesionální zprávy shrnující výsledky analýzy, včetně vložených tabulek a grafů. Umožňuje analyzovat velké soubory dat, provádět srovnávací studie nebo vylepšovat položky pomocí distraktorové analýzy. Ovládání je typu „point-and-click“, nepotřebujete psát žádný kód. Nekomerční verze programu Xcalibre stojí ročně 1495 USD.

11 DODATKY

11.1 Testová úzkost

Výsledek testu mohou ovlivňovat i emoce testovaného. Nejvíce se v této souvislosti hovoří o úzkosti (*testová úzkost, stres ze zkoušení*) (Embse et al. 2018).

Většina lidí prožívá před zkouškou určitou míru stresu nebo úzkosti. Některé jedince to dokonce může motivovat k lepším výkonům. Pokud je však míra stresu taková, že negativně ovlivní výkon testovaného při zkoušce, hovoříme o testové úzkosti.



Obr. 11.1.1 Yerkes-Dodsonův zákon, popisující vztah mezi stimulací a výkonem (Yerkes a Dodson 1908). Zdá se přitom, že pro výkon složitějších činností je optimální nižší míra stimulace, než u činností jednodušších (Nakonečný 1992).

Při zkoumání souvislosti výkonu a stresu obecně se ukázalo, že různé úkoly vyžadují pro optimální výkon různé úrovně vzrušení. Například obtížné nebo intelektuálně náročné úkoly vyžadují nižší úroveň vzrušení (k usnadnění soustředění), zatímco úkoly vyžadující vytrvalost

lze lépe provádět s vyššími úrovněmi vzrušení (ke zvýšení motivace). Yerkes-Dodsonův zákon, který popisuje vztah *vzrušení* (motivace, stresu) a *výkonu* naznačuje, že výkon jedince roste jen do určitého stupně excitace. Další zvyšování úrovně motivace a vzrušení již výkon naopak snižuje, přičemž hraniční míra užitečné excitace je individuální. Mechanismus patrně souvisí s působením stresových hormonů. K navození stresu vedou situace, které jsou nové, nepředvídatelné, mimo kontrolu jedince, nebo nesou riziko negativního sociálního hodnocení (vyloučení) (Lupien et al. 2007). Změny ve výkonu se ovšem neprojevují stejně u všech jeho typů. Stres například zlepšuje zapamatování faktických údajů, ale současně se zhoršují kreativní funkce. Zvyšuje se rychlost, ale snižuje se přesnost. Zátěž vhodné intenzity může zlepšovat výkon, nebo alespoň některé jeho složky, ve fázi učení. Pokud však přijmeme fakt, že testy mají zkoušet především porozumění a dovednosti, nikoli jen vybavování izolovaných faktů, vyžadujeme v průběhu testu i zapojení vyšších („tvořivějších“) kognitivních funkcí. Zdá se tedy, že i mírná zátěž zhoršuje výkon během zkoušení. Jakkoli můžeme diskutovat o přiměřenosti zátěže v průběhu výuky, obecně se přijímá názor, že v okamžiku zkoušení a testování bychom měli omezit stresující faktory na minimum (Hembree 1988; Andrews a Wilding 2004; Madsen 1982; Kader 2016).

Testová úzkost je jedním z faktorů, které **snižují reliabilitu testu**. Jde o jednu z forem tzv. *akademické úzkosti*. Spouštějí ji jednak kontextově specifické podněty (např. instruktáž před testem), jednak reakce specifické pro akademický předmět (např. úzkost z matematiky). Odhaduje se, že testová úzkost postihuje asi 15 až 22 % studentů (Embse et al. 2018).

Míra testové úzkosti zpravidla závisí na typu a významu zkoušky. Největší bývá u zkoušek velkého významu. Ovlivňuje ji řadě faktorů. Během více než padesáti let, během nichž se testová úzkost podrobněji zkoumá, vznikla řada teoretických modelů tohoto jevu. Většina vychází ze dvou nejstarších konceptů. *Interferenční model* testové úzkosti předpokládá, že horší výkon u testu lze vysvětlit faktory (např. emocemi nebo obavami), které narušují vybavování informací a práci s nimi. Naproti tomu *deficitní model* testové úzkosti předpokládá, že testová úzkost je důsledkem nedostatečných znalostí a schopností, včetně např. schopnosti efektivně studovat, vnímání vlastních schopností (*self-efficacy*), motivace či zvládání strategií, jak vyplnit test. Ani jeden z těchto dvou modelů nedokáže zcela vysvětlit variabilitu a dynamiku testové úzkosti, vznikají proto další teoretické koncepty. Novější přístupy zahrnují i vlivy vnějšího prostředí a také sociální vlivy, například prostředí, v němž vzdělávání probíhá, a vztahy mezi studenty a učiteli i studenty navzájem.

Do značné míry lze zobecnit, že testová úzkost je mírnější u studentů, kteří:

- mají lepší studijní výsledky,
- měli lepší výsledky u přijímacích zkoušek,
- mají lepší kognitivní a verbální schopnosti,
- mají větší sebedůvěru,
- očekávají, že zkouška bude snazší, nebo ji za snadnou považují.

Zajímavá je **souvislost s motivací**. Vnitřní motivace studenta ke studiu testovou úzkost zmenšuje. Testová úzkost se naopak zvětšuje vnější motivací, zejména pokud jde o motivaci negativní (např. zdůrazňování případných dopadů neúspěchu).

Podobně testová úzkost souvisí se schopnostmi vypořádat se s problémy. Je menší u osob, které při překonávání překážek využívají strategií zaměřených na odstranění nebo překonání stresoru. Naproti tomu osoby, které volí vyhýbavé strategie, vykazují větší míru testové úzkosti.

Míra testové úzkosti koreluje i s některými demografickými prediktory. Větší testovou úzkost mívají ženy, i když při srovnání různých studií se zdá, že závislost na pohlaví se postupně oslabuje (Hembree 1988, Embse et al. 2018). Významně více však bývají testovou úzkostí zasaženy osoby, které samy sebe vnímají jako příslušníky některé minority.

Prevence a zmírnění dopadů testové úzkosti

Testová úzkost snižuje reliabilitu testu. Zhoršuje výkon některých testovaných a neumožní jim zcela využít znalostí a dovedností, které mají, při vyplňování testu. U různých testovaných se přitom projevuje různou měrou, takže se stává nezanedbatelným zdrojem variability výsledného skóre v testu. Je proto žádoucí testové úzkosti předcházet, popřípadě její dopady minimalizovat.

Strategie pro prevenci a zvládání testové úzkosti lze rozdělit na opatření na straně učitele (organizátora testu) a opatření na straně testovaného.

Opatření doporučovaná studentům zasaženým testovou úzkostí jsou většinou založena na dostatečné přípravě na test, psychohygieně, relaxačních technikách, zvyšování sebevědomí, překonání nereálných obav apod. Některé vzdělávací instituce organizují programy, v nichž se snaží u studentů zasažených testovou úzkostí intervenovat a učí je testovou úzkost překonat (Weems et al. 2010).

Opatření na straně učitele nezahrnují jen samotný test nebo přípravu na něj. Záleží na celkovém nastavení výuky. Zásadní je, aby test zkoušel to, co se učí, tj. aby jeho obsah nebyl pro studenty překvapivý (Aydin 2007). Test tedy musí být odpovídajícím způsobem naplánovaný a validní. Důležité je také, aby studenti rozuměli tomu, jak bude test hodnocen, a měli v hodnocení důvěru.

Účinně proti testové úzkosti působí podpora metakognitivního přístupu k učení se (Mealey a Host 1992). Studenti by měli pochopit, proč se učí, jaké jsou cíle učení, jak a proč výuka probíhá, jaké jsou součásti vzdělávacího procesu, jaký je význam testu atd. Významná je také sociální podpora a vytvoření sociálních vazeb. Testová úzkost je větší u studentů, kteří se učí izolovaně od ostatních. Zařazování skupinové práce do výuky testovou úzkost snižuje.

Mezi relativně jednoduchá opatření, kterými může učitel testovou úzkost zmenšovat, patří např.:

- Studenty předem seznámíme s tématy a rozsahem testu.
- Umožníme studentům vyzkoušet si předem testové prostředí, zvláště pokud test probíhá v elektronické formě.
- Studenty předem seznámíme s formátem úloh a způsobem odpovídání na ně. Pokud se odpovídá pomocí formuláře, vysvětlíme, jak přesně se s formulářem pracuje.
- Umožníme studentům absolvovat „zkoušku na nečisto“ (*mock test*). Zkouška nanečisto může být i velmi krátká, s jen několika úlohami, měla by ale obsahovat všechny prvky

skutečného testu (např. přípravu pracovního místa, ověření identity, stejný způsob zadání a odpovídání na úlohy).
- Probereme předem se studenty témata, která jsou klíčová a v testu se budou objevovat. Tím se sníží tzv. „obsahová nejistota".
- Pomůžeme studentům rozvrhnout si čas potřebný na přípravu na zkoušku.
- Před testy většího významu by měla předcházet řada dílčích formativních hodnocení, která studenta k sumativní zkoušce „vedou", ukazují mu, nakolik dosahuje očekávaných znalostí a dovedností, jaká jsou jeho slabá a silná místa, a současně jej postupně připravují na obsah a rozsah sumativní zkoušky.
- Při samotné zkoušce se snažíme minimalizovat faktory, které by mohly studenty rozptylovat a rušit. V době zkoušky se snažíme připravit předvídatelné a vlídné prostředí.

Proti snahám o zmírňování testové úzkosti může stát skutečnost, že některá povolání jsou spojena i s prací pod tlakem a není možné do nich nechat vstupovat „skleníkové bytosti". Oprávněným požadavkem pak může být, abychom ověřili, že student dokáže efektivně a přesně pracovat i ve stresu. V takovém případě by ale studenti měli být tlaku vystavováni především v průběhu výuky a formativních hodnocení, nikoli ve standardizované závěrečné sumativní zkoušce. Práce v zátěžových situacích také může být součástí praktického zkoušení. Ve většině písemných zkoušek a testů je však testová úzkost nežádoucí, neboť snížením reliability testu nakonec ohrožuje i jeho validitu a hodnotitelnost.

11.2 Náklady testování

Příprava kvalitních testových úloh, provedení a vyhodnocení testů jsou odborně náročné, pracné a nákladné úkony. Při velkém počtu zkoušených studentů lze očekávat, že cena jednoho testu klesne („úspory z rozsahu"), protože fixní náklady se rozdělí na víc částí.

Pro malé počty hodnocených je rozhodně méně pracné zkoušet ústně než připravovat, provádět a hodnotit testy. K rozhodnutí zkoušet malý počet studentů pomocí testů obvykle vedou závažné důvody. Hodnocení studentů pomocí testů může být metodou volby, pokud potřebujeme, aby výsledky byly prokazatelně objektivní a reprodukovatelné, např. když hrozí, že se proti nim budou testovaní odvolávat. Z ekonomického hlediska se testování vyplatí při velkém počtu zkoušených, neboť vysoké pořizovací náklady budou vyváženy nízkými provozními náklady.

Náklady na položku

Vytvoření kvalitních testových úloh vyžaduje tým expertů v příslušném oboru, proškolených navíc v metodice tvorby testů. Další výdaje přináší recenzování otázek a pilotní testování s dostatečně velkou skupinou studentů.

Náklady na přípravu kalibrovaných položek pro důležité testy se odhadují na 1000 USD za úlohu; obecně se dá říci, že náklady na položku neklesají pod 300 USD (Downing a Haladyna 2006a).

Celkové náklady na vývoj jedné kvalitní položky do adaptivního přijímacího testu spočítal např. Rudner (2010). Ukázal, že kvalitní kalibrovaná položka stojí (v USA)

1500–2000 USD. Porovnáme-li údaj s odhadem nutného počtu položek v položkové bance pro běžné adaptivní testování od Breithauptové (asi 2000 položek v bance), dojdeme k astronomické částce 3–4 miliony USD za obsah položkové banky (Breihaupt et al. 2010, Gierl et al. 2012c).

Náklady na položky se mohou snížit, můžeme-li znovu použít již hotové úlohy. Mohou to být položky, které jsme připravili a kalibrovali v minulých kolech testování, pokud jsme si jisti, že nebyly předchozím použitím exponovány.

Náklady na položku na 1. LF UK

Na 1. LF UK jsme spočítali náklady na novou položku v akademickém roce 2020/21. Bylo připraveno 230 položek, při výsledné ceně 1 500,- Kč za položku. Do ceny se největší měrou promítají náklady na práci autorů a recenzentů a v menší míře náklady na provoz a pořízení položkové banky. Je zřejmé, že v porovnání se zahraničním máme výhodu nízké ceny kvalifikované pracovní síly.

Náklady na testovaného

Zajímavé je podívat se na náklady testování vztažené na jednoho otestovaného studenta. Je zde patrný velký rozptyl údajů vzhledem k různým podmínkám a nárokům.

Centrum pro výzkum vzdělávacích standardů a studentského testování (CRESST) ve zprávě z roku 1996 počítalo náklady hodnocení pomocí testů se započtením platů učitelů. Jeho odhady se pohybovaly od 848 do 1 792 USD na studenta (Picus et al. 1996).

Jak významné mnohou být úspory z rozsahu, ukazuje příklad známých zkoušek ACT a SAT, které jsou ve Spojených státech součástí přijímacího řízení na vysoké školy. Tyto zkoušky v roce 2020 absolvovalo 1,65 a 1,1 milionu testovaných. Ceny za tyto testy, které pokrývají veškeré náklady na vývoj, bodování a administraci testů, se pohybují od 20 do 70 USD.

Ačkoli jsou odhady nákladů a přínosů obvykle závislé na kontextu a rozsahu, je zřejmé, že přínosy počítačově podporovaného testování podstatně převažují nad jeho náklady (Phelps 2002).

Náklady na testovaného na 1. LF UK

Opět můžeme porovnat tyto náklady s náklady při přijímacím řízení na 1. LF UK v akademickém roce 2020/21. Omezíme-li se na magisterské obory, do kterých se hlásilo cca 4000 uchazečů, dostáváme náklady na jednoho testovaného přibližně 250 Kč. Z toho největší část nákladů tvoří tzv. „náklady testového dne“, které představují více než třetinu celkové sumy. Další zhruba třetina jsou tiskové náklady a zbylá část se dělí na tvorbu úloh a náklady položkové banky. I zde je patrné, že jsme na zlomku (cca třetině) porovnávaných nákladů zkoušek typu SAT a ACT, a to navzdory nepoměru v počtu testovaných. Menší poměr nákladů než v případě tvorby otázek, kde byl poměr více než 1:20, přisuzujeme rozdílu v počtu testovaných a nestlačitelnosti nákladů tisku, které nejsou tak geograficky citlivé jako náklady na pracovní sílu.

11.3 Zkratky v textech o testování

A-level Advanced Level (plným názvem General Certificate of Education Advanced Level) je označení certifikátu a zkoušek, které jsou ve Velké Británii součástí státní maturity. A-level jsou na mnoha univerzitách přijímány jako jeden z významných ukazatelů vhodnosti uchazečů o přijetí pro studium na vysoké škole. Pro výběr studentů na lékařské fakulty bývala požadována nejvyšší hodnocení ze tří předmětů (A), které musely obsahovat chemii a alespoň jednu další přírodní vědu nebo matematiku.

ACT American College Testing je jeden ze dvou nejpoužívanějších přijímacích testů na vysoké školy v USA a v Kanadě. Většina škol dává studentům na výběr, který z testů absolvují, a stanovují jen počty bodů nutné pro přijetí v tom kterém testu. ACT se skládá se ze čtyř částí: angličtiny (45 min.), matematiky (60 min.), čtení (35 min.) a vědeckého myšlení (35 min.).

AERA American Educational Research Association – Americká společnost pro výzkum ve vzdělávání.

AIG Automatic Item Generation – automatická tvorba testových položek.

AMEE Association for Medical Education in Europe – původně evropské, nyní celosvětové sdružení pro vzdělávání lékařů.

APA American Psychological Association – Americká psychologická společnost.

BMAT BioMedical Admissions Test – test používaný ve Velké Británii pro přijetí na tradiční lékařské fakulty s rozdělením preklinické a klinické výuky a důrazem na výuku vědy v prvních letech studia. BMAT je následníkem Medical and Veterinary Admissions Test (MVAT). O přínosu tohoto testu, respektive jeho jednotlivých částí, se vedou diskuse.

BTU Banka testových úloh, též položková banka, je databáze testových úloh, umožňující s úlohou uložit i informace o jejím vytvoření, využití a psychometrických vlastnostech.

CAA Computer Assisted Assessment – počítačově podporované hodnocení.

CAT Computer Adaptive Testing – adaptivní počítačové testování.

CBA Competency based assessment – hodnocení založené na kompetencích. Porovnává jedince s požadovanými standardy.

CBA/CBT Computer-Based Assessment/Testing – hodnocení (testování) prostřednictvím počítače nebo jiného podobného zařízení, např. tabletu, mobilního telefonu, aj. (protiklad k PPT).

CBD Case-Based Discussion označuje strukturovanou diskusi o klinických případech řízenou lékařem, testující klinické uvažování.

CFT Computerized Fixed-form Tests – klasický test v počítačové formě.

Class rank Pořadí výkonu studenta střední školy ve srovnání s ostatními studenty ve třídě. Viz též *high school class rank* (HSCR).

CRT Criterion-Referenced Test – standardizovaný test porovnávající výkon studenta s předem stanovenými standardy vyžadovanými pro úspěšné absolvování testu (srovnej s NRT).

CTT Classical Test Theory – klasická teorie testů je nástroj analýzy testů. Je jednodušší a snáze se používá než dokonalejší nástroj analýzy testů IRT (teorie odpovědi na položku).

DIF Differential Item Functioning – index rozdílného fungování položky indikuje odlišné chování testové úlohy pro skupiny se stejnou úrovní znalosti (schopnosti, studijního výkonu), ale odlišným etnickým, nebo genderovým složením.

DOPS Direct Observation of Procedural Skills – přímé sledování procedurálních dovedností.

ECD Evidence-Centered assessment Design – na důkazech založený přístup ke konstrukci hodnocení výsledků výuky.

EDF Educational Data Forensics – bezpečnostní analýza testů.

EMQ EMQ nebo též EMI (Extended Matching Question/Item) jsou „rozšířené přiřazovací otázky". Jde o výběrové úlohy s jedinou nejlepší odpovědí, v nichž testovaný vybírá z většího počtu (typicky kolem dvaceti) možností. Stejná sada možností se zpravidla používá pro několik úloh, které jsou v testu za sebou.

ETS Educational Testing Service je nezisková vzdělávací organizace zaměřená na testování a hodnocení. ETS vyvíjí různé standardizované testy pro střední a vysoké školství v USA a také spravuje mezinárodní testy včetně jazykových zkoušek TOEFL.

FYGPA First Year Grade Point Average – studijní průměry v prvním ročníku vysoké školy jsou používány pro odhad akademické výkonosti studenta na vysoké škole.

GAMSAT Graduate Australian Medical School Admissions Test je test pro výběr uchazečů o magisterské studium medicíny, stomatologie a veterinárního lékařství vyvinutý v roce 1995 Australskou radou pro výzkum vzdělávání (ACER). Používá se pro výběr studentů na magisterský stupeň studia zdravotnických vysokých škol v Austrálii a od roku 1999 i na některých školách ve Velké Británii a v Irsku.

GAT General Aptitude Test – test všeobecné studijní připravenosti (VSP) je souhrnný název pro schopnostní testy, které testují rozumové schopnosti uchazečů (na rozdíl od testů znalostních). V testu VSP bývají zpravidla otázky zaměřené na prostorové vztahy, logické souvislosti a představivost.

GCSE General Certificate of Secondary Education – certifikát o středoškolském vzdělání v příslušném oboru používaný v Anglii, srovnatelný s maturitním vysvědčením. Vydává se 14–16letým studentům po splnění příslušné zkoušky.

GDPR General Data Protection Regulation – je obecné nařízení o ochraně osobních údajů zakotvené v legislativě EU pro ochranu osobních dat občanů.

GPA Grade Point Average – studijní průměry. Studijní průměry ze střední školy bývají jako dobrý prediktor úspěšnosti studia zařazovány mezi výběrová kritéria pro přijetí na některé fakulty.

HEI Higher Education Institution – vysokoškolské instituce (vysoké školy)

HSCR High School Class Rank je měřítko studijní výkonnosti konkrétního studenta v poměru k výkonu ostatních ve třídě. Jiné označení pro tento parametr je Clas rank. Vypočte se jako pořadí studenta stanovené na základě GPA a vydělené počtem studentů ve třídě. Výsledkem je percentil nejlepších studentů, mezi něž student patří. Zejména v zahraničí tento údaj poskytují některé střední školy. Velké veřejné školy poskytují tento údaj častěji než malé soukromé školy. HSCR se spolu s GPA často používá pro ohodnocení studenta při přijímání na vysoké školy.

HSGPA High-School Grade Point Average – označení pro studijní průměr na střední škole (USA). Viz též GPA a uGPA.

ICC Item Characteristic Curve – charakteristická křivka položky vyjadřuje vztah mezi měřeným latentním rysem testovaného (znalostí) a pravděpodobností správné odpovědi na položku. Užívá se v teorii odpovědi na položku. Je to jiné pojmenování pro ICF.

ICF Item Characteristic Function – charakteristická funkce položky vyjadřuje vztah mezi měřeným latentním rysem testovaného (znalostí) a pravděpodobností správné odpovědi na položku. Užívá se v teorii odpovědi na položku.

IDEAL IDEAL Consortium (International Database for Enhanced Assessments and Learning) – dobrovolné sdružení 23 vysokých škol z celého světa sdílejících testové úlohy z oboru medicíny v angličtině.

IMS Item Management System – položková banka v širším slova smyslu – systém pro tvorbu, uchování, sdílení a doručování položek.

IRT Item Response Theory – teorie odpovědi na položku – moderní nástroj analýzy testů umožňující odhadnout vlastnosti položky pro různé úrovně znalosti.

IRM Item Response Models – označuje skupinu matematických modelů, které se snaží vysvětlit vztah mezi latentními znaky a jejich projevy (tj. pozorovanými výsledky, odpověďmi nebo výkonem) pomocí teorie odezvy na položku (IRT).

JISC Joint Information Systems Committee – společný výbor pro informační systémy je nevládní veřejná instituce ve Velké Británii, jejímž úkolem je podporovat vysokoškolské vzdělávání zaváděním informačních a komunikačních technologií.

LMS Learning Management System – systémy pro organizování výuky, např. Moodle, BlackBoard, Adobe Connect atd.

LTI Learning Tools Interoperability – standard pro spolupráci e-learningových nástrojů a prostředí.

MCAT Medical College Admissin Test – standardizovaný „počítačový" přijímací test na lékařské fakulty v USA, zavedený Asociací amerických lékařských fakult v USA.

MCAT(R) V letech 1991–1992 byl MCAT znovu revidován a restrukturalizován a jeho nová podoba nese výše uvedené označení.

MCQ Multiple Choice Question – otázka se mnohočetným výběrem odpovědi, nejobecnější kategorie, která zahrnuje všechny formy testových úloh s výběrem odpovědi.

MEFANET MEdical FAculties NETwork – dobrovolné sdružení českých a slovenských lékařských a zdravotnických fakult spolupracujících na elektronické podpoře výuky.

MEQ Modifikovaný esej (Modified Assay Question) je formát otevřených úloh, v nichž se očekává tvořená odpověď delší než v SAQ, ale podstatně kratší než esej. Testovaný postupně odpovídá na dílčí otázky, mezi nimiž získává různé doplňující informace. Pomocí tohoto typu úloh lze posoudit analytické uvažování, interpretaci dat a kritické rozhodování.

MeSH Medical Subject Headings – řízený slovník deskriptorů pro indexování v medicíně a biologii.

MRQs Multiple Response Questions značí otázky s mnohočetným výběrem odpovědí, ve kterých je správně (a je třeba vyznačit) více nabízených odpovědí – synonymum pro MTF.

MSC-AA Medical Schools Council Assessment Alliance je organizace lékařských vysokých škol ve Velké Británii spolupracujících při hodnocení výsledků pregraduální výuky. Jejím předchůdcem byl UMAP.

MSF Multisource Feedback – (též 360degree feedback) vícezdrojové hodnocení (360° zpětná vazba) je metoda hodnocení, při níž je jedinci poskytována zpětná vazba pomyslným kruhem respondentů, kteří s ním přicházejí do styku, a porovnávána s jeho sebehodnocením. Přínosem tohoto hodnocení je, že podává informaci o tom, jak si jedinec v očích ostatních počíná.

MTF Multiple True/False question. Otázka s výběrem z několika odpovědí, z nichž několik může být správných. Testovaný u každé nabízené odpovědi zvažuje, zda je správná, či nesprávná. V praxi se často zaměňuje s daleko obecnějším termínem MCQ.

NBME National Board of Medical Examiners je nezávislá, nezisková organizace, která se zabývá posuzováním kvality vzdělání zdravotnických pracovníků. NBME vyvíjí a spravuje USMLE (národní licenční zkoušky pro výkon lékařství).

NCME National Council on Measurement in Education – Národní rada pro měření ve vzdělávání je americká profesionální organizace pro jednotlivce zapojené do posuzování, hodnocení, testování a dalších aspektů měření výsledků vzdělávání. Vydává čtvrtletník The Journal of Educational Measurement (JEM).

NDA Non-disclosure agreement – dohoda o mlčenlivosti.

NRT Norm-Referenced Testing je standardizovaná zkouška, v níž výkon jedince je hodnocen v porovnání s výkony relevantní populace (srovnej s CRT).

OMR Optical Mark Recognition – optické rozpoznávání značek používané pro strojové vyhodnocování odpovědních formulářů

OSCE/OSPE Objective Structured Clinical/Practical Examination – objektivně strukturované klinické/praktické zkoušení je způsob objektivního hodnocení výsledků výuky klinických/praktických dovedností. Zkoušení je většinou organizované jako 5–10minutová zastavení na stanovištích, kde zkoušený řeší příslušný úkol.

P&P Paper-and-Pencil – pomocí pera a papíru.

PPT Paper-and-Pencil Testing – testování v papírové verzi.

QTI Question and Test Interoperability specification – mezinárodní standard pro interoperabilitu testových systémů.

RIR Item Rest Correlation – Index RIR (též psáno RIR) je korelační koeficient mezi úspěšností v dané testové položce a celkovým počtem bodů v testu při vyloučení dané položky. Koeficient RIR nabývá hodnot od -1 do 1 a používá se k hodnocení diskriminační schopnosti položky. Dobře diskriminující položka by měla dosahovat hodnoty RIR nejméně 0,3. Výrazně menší nebo záporné hodnoty indikují, že položka není rozlišující, nebo diskriminuje opačně než test.

RIT Item Test Correlation – Index RIT (psáno též RIT) označuje korelační koeficient mezi úspěšností v dané testové úloze a celkovým počtem bodů v testu. Koeficient RIT se používá podobně jako index RIR.

RIC Responses in Common index je počet položek, na které dali dva zkoumaní stejnou odpověď.

RTE Response Time Effort – doba odezvy, čas do vyplnění odpovědi.

SAQ Otázka s krátkou odpovědí (Short-Answer Question) je otevřená úloha, na niž má odpovídající vytvořit velmi krátkou odpověď (jednoslovnou, slovní spojení). Odpovědí, které se vyhodnotí jako správné, může být několik.

SAT Standardized Admissions Tests tvoří spolu s ACT dva nejrozšířenější testy pro zjišťování připravenosti středoškoláků na vysokou školu v USA. V aktuální podobě platné od roku 2005 trvá SAT tři a tři čtvrtě hodiny a skládá se ze tří částí: kritického čtení, matematiky a psaní. Za každou část je možné dosáhnout až 800 bodů. V testu je zahrnuta „experimentální" část, která se nepoužívá pro vyhodnocení studentových schopností, ale pro zhodnocení otázky samotné, pro případné budoucí použití v testech SAT.

SBA Single Best Answer Question. Otázka s výběrem z obvykle tří až pěti nabízených variant odpovědí. Testovaný volí právě jednu z nabídnutých odpovědí. Ostatní možnosti (distraktory) jsou buď nesprávné, nebo (častěji) jde o kvalitativně výrazně méně vhodné odpovědi na otázku.

uGPA Undergraduate grade point average – průměr známek na střední škole, často používaný pro predikci studijního úspěchu na vysokých školách. Viz též GPA a HSGPA.

UKCAT UK Clinical Aptitude Test – je test vytvořený v roce 2006 konsorciem britských lékařských a stomatologických fakult pro testování duševních schopností uchazečů. UKCAT je navržen tak, aby testoval schopnosti a postoje, nikoli akademický úspěch, který je dobře predikován pomocí A-leves, GCSE nebo GPA. Test je tedy zaměřen na schopnost kritického logického myšlení a schopnost vyvozovat závěry. Přínos tohoto testu je sporný a podněcuje v UK diskusi o vhodnosti psychologického testování uchazečů pro výběr studentů medicíny.

ULI Upper-Lower index je index pro hodnocení citlivosti neboli diskriminační schopnosti položky.

UMAP Universities Medical Assessment Partnership – dřívější dobrovolné sdružení lékařských fakult ve Velké Británii založené v roce 2003 za účelem spolupráce při tvorbě a sdílení testových otázek. Sdružení se roce 2009 přeměnilo na současnou MSC-AA.

UMAT Undergraduate Medicine and Health Sciences Admission Test se používá pro výběr středoškolských uchazečů o studium medicíny v Austrálii a na Novém Zélandu. Po absolvování bakalářského stupně jsou uchazeči o navazující „magisterské" studium vybíráni pomocí GAMSAT.

USMLE United States Medical Licensing Examination je oficiální zkouška pro absolventy lékařských fakult pro vstup do postgraduálních programů klinické medicíny v USA.

VLE Virtual Learning Environment – virtuální prostředí pro výuku, například Moodle.

VSP Test všeobecné studijní připravenosti je souhrnný název pro schopnostní testy, které testují rozumové schopnosti uchazečů. V testech VSP bývají zpravidla otázky zaměřené na prostorové vztahy, logické souvislosti a představivost. Viz též GAT.

12 DOSLOV

Testování je dnes samozřejmou součástí výuky na vysokých školách. Záměrem autorů bylo poskytnout první orientaci učitelům, které tato oblast zajímá. Cílem byla spíše popularizace postupů a metod než jejich detailní zkoumání.

Dovolíme si na tomto místě poděkovat všem, kdo nám pomáhali se v tématech zorientovat, kdo nás ponoukali a inspirovali.

Věříme, že každý zájem probouzí v lidech otázky, a budeme rádi, pokud budete hledat odpovědi, které tento text přesahují. Přejeme všem, kdo se touto cestou vydají, aby je to bavilo stejně jako nás.

A na závěr naše oblíbené memento:
Výsledek testu se má k hodnocení studenta jako laboratorní výsledek k diagnóze.

13 LITERATURA

1. LF UK 2017. *Tao of Testing: instalace na 1. LF UK*. Dostupné z: https://tao.lf1.cuni.cz/.

Abdellatif, H. a Al-Sharani, A. M. 2019. Effect of Blueprinting Methods on Test Difficulty, Discrimination, and Reliability Indices: Cross-sectional Study in an Integrated Learning Program. *Advances in Medical Education and Practice*. 10: 23–30. ISSN 1179-7258. DOI:10.2147/AMEP.S190827.

Adesope, O. O. et al. 2017. Rethinking the Use of Tests: A Meta-Analysis of Practice Testing. *Review of Educational Research*. 87 (3): 659–701. ISSN 0034-6543. DOI: 10.3102/0034654316689306.

Al-Faris, E. A. et al. 2010. A Practical Discussion to Avoid Common Pitfalls When Constructing Multiple Choice Questions Items. *J Family Community Med* . 17 (2): 96–102. DOI:. ISSN 1319-1683 (print), 2229–340X.

Albano, A. D. 2016. Equate: An R Package for Observed-Score Linking and Equating. *Journal of Statistical Software*. 74 (8). ISSN 1548-7660. DOI:10.18637/jss.v074.i08.

Alderson, J. et al. 1995. *Language Test Construction and Evaluation*. New York, NY, USA: Cambridge University Press. 310 s. ISBN 0-521-47255-5.

American Educational Research Association 2014. *Standards for Educational and Psychological Testing: AERA, APA & NCME*. Washington: American Educational Research Association. IX. 230 s. ISBN 9780935302356.

American Psychological Association 2020. Testwise. *APA Dictionary of Psychology*. Washington DC: American Psychological Association. Dostupné z: https://dictionary.apa.org/testwise.

Andrews, B. a Wilding, J. M. 2004. The Relation of Depression and Anxiety to Life-stress and Achievement in Students. *British Journal of Psychology.* 95: 509–521.

ASC. The Angoff Analysis Tool: A free Spreadsheet to Set Cutscores That Are Legally Defensible, Using the Modified-Angoff Method. *Assessment Systems Corporation (ASC)*. Dostupné z: https://assess.com/angoff-analysis-tool/.

Assessment Systems 2018. How the Angoff Analysis Tool Makes it Easy to Set Defensible Cutscores. *YouTube: Assessment Systems*. 22.3.2018. Dostupné z: https://www.youtube.com/watch?v=CQh6hJpDfl8.

Aydin, S. 2007. How Can Teachers Reduce Test Anxiety of L2 Learners. *Humanising Language Teaching.*

Ayers, W. 2001. *To Teach : The Journey of a Teacher.* 2. vydání. New York: Teachers College Press. 151 s. s. 116. ISBN 08-077-3985-5.

Aziz, S. 2005. *A Modified Ebel Standard Setting Method for a Medical School Clinical Skills Assessment.* Chicago: University of Illinois. 162 s.

Baker, V. 2019. Celebrity Parents and the Bizarre ‚Cheating' Scandal: US College Admissions Scandal. *BBC News*. Washington DC: BBC. 15. 3. 2019. Dostupné z: https://www.bbc.com/news/world-us-canada-47585336.

Baumgartnerová, G. a Kapustová, A. 2013. *Metodický materiál pro hodnotitele písemných prací z českého jazyka a literatury.* Praha: Centrum pro zjišťování výsledků vzdělávání. 37 s.

Baylem, N. J.et al. 2011. Would the MRCS Written Papers Benefit from Computerisation? The University of Nottingham Experience. *The Bulletin of the Royal College of Surgeons of England.* 93(1): 1–5. ISSN 14736357. DOI: 10.1308/147363511X546545.

Berger, R. 2018. Here's What's Wrong With Bloom's Taxonomy: A Deeper Learning Perspective. *Education Week*. 14. 3. 2018. Dostupné z: https://www.edweek.org/education/opinion-heres-whats-wrong-with-blooms-taxonomy-a-deeper-learning-perspective/2018/03.

Berk, R. A. 2002. *Humor as an Instructional Defibrillator: Evidence-based Techniques in Teaching and Assessment*. Sterling, Va.: Stylus. 268 s. ISBN 1579220630.

Bernardi, R. A. et al. 2008. Methods of Cheating and Deterrents to Classroom Cheating: An International Study. *Ethics & Behavior.* 18 (4): 373–391. ISSN 1050-8422. DOI:10.1080/10508420701713030

Boone, W.J. et al. 2017. Rasch Analysis: A Primer for School Psychology Researchers and Practitioners. *Cogent Education*. 4 (1). ISSN 2331-186X. DOI:10.1080/2331186X.2017.1416898.

Bourque, J. et al. 2020. Performance of the Ebel Standard-Setting Method in Spring 2019 Royal College of Physicians and Surgeons of Canada Internal Medicine Certification Examination Consisted of Multiple-choice Questions. *Journal of Educational Evaluation for Health Professions*. 20. 4. 2020. 17: 12. DOI:10.3352/jeehp.2020.17.12.

Boussakuk, M. et al. 2021. Designing and Developing e-Assessment Delivery System Under IMS QTI ver.2.2 Specification. *International Journal of Emerging Technologies in Learning (iJET)* 16 (01): 219–233. ISSN 1863-0383. DOI:10.3991/ijet.v16i01.16257.

Bowers, J. J. a Shindoll, R. R. 2014. A Comparison of the Angoff, Beuk, and Hofstee Methods for Setting a Passing Score. *ACT Research Report Series* 892.

Breithaupt, K. et al. 2010. Assembling an Inventory of Multistage Adaptive Testing Systems. In: *Elements of Adaptive Testing*. New York: Springer. s. 247–266. Dostupné také z: http://link.springer.com/chapter/10.1007%2F978-0-387-85461-8_13. DOI:10.1007/978-0-387-85461-8_13. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8.

Breslau, J. et al. 2003. Differential Item Functioning Between Ethnic Groups in the Epidemiological Assessment of Depression. *Journal of Nervous & Mental Disease*. 196 (4): 297–306. ISSN 0022-3018. DOI:10.1097/NMD.0b013e31816a490e.

Burnett, A. J.et al. 2016. Use of the Social Cognitive Theory to Frame University Students' Perceptions of Cheating. *Journal of Academic Ethics* 14 (1): 49–69. ISSN 1570-1727. DOI:10.1007/s10805-015-9252-4.

Burr, S. A. et al. 2017. Angoff Anchor Statements: Setting a Flawed Gold Standard? *MedEdPublish.* 6(3). ISSN 23127996. DOI:10.15694/mep.2017.000167.

Butterwick, D. J. et al. 2006. Development of Content-valid Technical Skill Assessment Instruments for Athletic Taping Skills. *Journal of Allied Health.* 35 (3): 149–157. ISSN (Print) 0090-7421, (Online) 1945-404X. PMID: 17036669.

Byčkovský, P. a Marková, M. 2003. Využití software ITEMAN k položkové analýze a analýze výsledků testů. In: *11. konference ČAPV – Sociální a kulturní souvislosti výchovy a vzdělávání: Sborník referátů*. Brno: Pedagogická fakulta, Masarykova univerzita. Dostupné z: http://www.ped.muni.cz/capv11/5sekce/5_CAPV_Byckovsky.pdf.

Byčkovský, P. a Zvára, K. 2007. *Konstrukce a analýza testů pro přijímací řízení*. Praha: Univerzita Karlova v Praze, Pedagogická fakulta. 79 s. ISBN 978-80-7290-331-3.

Cantor, J. A. 1989. A Validation of Ebel's Method for Performance Standard Setting through its Application with Comparison Approaches to a Selected Criterion-Referenced Test. *Educational and Psychological Measurement.* 49 (3): 709-721. ISSN (Print) 0013-1644; (Online) ISSN: 1552-3888.

Carlton, J. F. a Krou, M. 2021. Motivation is a Key Factor in whether Students Cheat. *The Conversation*. Dostupné z: https://theconversation.com/motivation-is-a-key-factor-in-whether-students-cheat-155274.

Case, R. 2013. The Unfortunate Consequences of Bloom's Taxonomy. *Social Education*. National Council for the Social Studies. 77 (4): 196–200. ISSN 0037-7724.

CERMAT 2018. Hodnotící zpráva Matematika+ 2018: Pokusné ověřování obsahu, formy, organizace a hodnocení výběrové zkoušky ze středoškolské matematiky. *CERMAT: Centrum pro zjišťování výsledků vzdělávání*. Praha. Dostupné z: https://data.cermat.cz/files/files/Matematika/MA-PLUS_hodnotici_zprava_2018.pdf.

Cígler, H. 2014. Jak začít s Teorií odpovědi na položku?: S pomocí knihy „Applying The Rasch Model: Fundamental Measurement in the Human Sciences". *Testfórum*. (3). Dostupné z: https://testforum.cz/article/download/TF2014-3-15/10487.

Cígler, H. 2020. *Přednáška 8: Férovost a zkreslení při testování*. Fakulta sociálních studií MU: Katedra psychologie. Brno: MUNI. 24. 11. 2020. Dostupné z: https://is.muni.cz/el/fss/podzim2020/PSYn4790/um/PSYn4790_2020_P08.pdf?lang=en.

Cizek, G. J. a Wollack, J. A., 2017. *Handbook of Quantitative Methods for Detecting Cheating on Tests*. New York and London: Routledge. ISBN 978-1-138-82180-4.

Clauser, B. E. et al. 2009. Judges' Use of Examinee Performance Data in an Angoff Standard-Setting Exercise for a Medical Licensing Examination: An Experimental Study. *Journal of Educational Measurement*. 46 (4): 390–407. ISSN 00220655. DOI:10.1111/j.1745-3984.2009.00089.x.

Cohen-Schotanus, J. a Vleuten, C. P. M. van der 2010. A Standard Setting Method With the Best Performing Students as Point of Reference: Practical and Affordable. *Medical Teacher*. 32 (2): 154–160. ISSN 0142-159X. DOI:10.3109/01421590903196979.

Council of Chief State School Officers 2021. *A Practitioner's Introduction to Equating: With Primers on Classical Test Theory and Item Response Theory*. Washington: Council of Chief State School Officers. Dostupné z: https://ccsso.org/resource-library/practitioners-introduction-equating.

Council on Higher Education 2013. *The Aims of Higher Education*. Pretoria: Case. ISBN 978-1-919856-84-1. Dostupné z: https://www.che.ac.za/sites/default/files/publications/kagisano9.pdf.

Cressey, D. R. 1953. *Other People's Money: A Study of the Social Psychology of Embezzlement.* Glencoe: Free Press. 191 s.

Crocker, L. M. a Algina, J. 1986. *Introduction to Classical and Modern Test Theory*. New York: Holt, Rinehart, and Winston. ISBN 0030616344.

Cruess, R. L. et al. 2016. Amending Miller's Pyramid to Include Professional Identity Formation. *Academic Medicine.* 91 (2): 180–185. ISSN 1040-2446. DOI:10.1097/ACM.0000000000000913.

CSPC ed. 2021. Extended Common Marking Scheme: Approved by CSPC on 7 June 2011 For Use from Academic Year 2011/12. *The University of Edinburgh*. Dostupné z: https://www.ed.ac.uk/timetabling-examinations/exams/regulations/common-marking-scheme.

Culbertson, J. C. 1995. An Essay Review: The Bell Curve: Class Structure and the Future of America. *Education Policy Analysis Archives*. 3 (2): 1–12. Dostupné z: http://epaa.asu.edu/ojs/article/view/645/767. ISSN 1068-2341.

Čeština pro cizince 2010. *Pokyny k organizaci zkoušky z českého jazyka pro trvalý pobyt v ČR*.

Čulík, J. 1999. *Jak se prodávají zkoušková zadání na právnické fakultě*. 14. 7. 1999 Dostupné z: http://www.ceskaskola.cz/1999/07/jan-culik-jak-se-prodavaji-zkouskova.html.

Davidson, C. N. 2011. *Now you see it: How the Brain Science of Attention Will Transform the Way We Live, Work and Learn : [object Object]*. Viking Adult. 342 s. ISBN 9780670022823.

Davier, M. V. 2019. *Training Optimus Prime, M.D.: Generating Medical Certification Items by Fine-Tuning OpenAI's gpt2 Transformer Model*. ArXiv: abs/1908.08594.

Dendir, S. a Maxwell, R. S. 2020. Cheating in Online Courses: Evidence From Online Proctoring. *Computers in Human Behavior Reports*. 2. ISSN 24519588. DOI:10.1016/j.chbr.2020.100033.

Denison, A. et al. 2016. Tablet Versus Paper Marking in Assessment: Feedback Matters. *Perspectives on Medical Education*. 5 (2): 108–113. ISSN 2212-2761. DOI:10.1007/s40037-016-0262-8.

Dennick, R. et al. 2009. Online eAssessment: AMEE Guide No. 39. *Medical Teacher*. 31 (3): 192–206. ISSN 0142-159X. DOI:10.1080/01421590902792406.

Diedenhofen, B. a Musch, J. 2017. PageFocus: Using Paradata to Detect and Prevent Ceating on Online Achievement Tests. *Behavior Research Methods*. 49: 1444–1459.

Dolejš, M. et al. 2012. *Testová příručka ke škále osobnostních rysů představujících riziko z hlediska užívání návykových látek: (SURPS – substance use risk profile scale)*. Praha: Klinika adiktologie, 1. lékařská fakulta Univerzity Karlovy v Praze a Všeobecná fakultní nemocnice v Praze ve vydavatelství Togga. 84 s. ISBN 978-80-87258-81-1.

Downing, S. M. a Haladyna, T. M. 2006a. Computerized Item Banking. In: *Handbook of test development*. Mahwah, N. J.: L. Erlbaum. s. 261–286. ISBN 9780805852646.

Downing, S. M. a Haladyna T. M. 2006b. *Handbook of Test Development.* Mahwah: Lawrence Erlbaum Associates. 778 s. ISBN 9780805852653.

Drasgow, F. et al. 2006. Technology and Testing. In: Brennan, R. L. *Educational Measurement*. 4. vydání. Praeger Publishers. 779 s. Washington, DC: American Council on Education. ISBN 0275981258, 9780275981259.

Driessen, E. et al. 2007. Portfolios in Medical Education: Why Do They Meet with Mixed Success? A Systematic Review. *Medical Education*. 41 (12): 1224–1233. ISSN 03080110. DOI:10.1111/j.1365-2923.2007.02944.x.

DuBois, P. H. 1970. *A History of Psychological Testing*. Michigan: Allyn and Bacon.

Durning, S. J. et al. 2016. Comparing Open-Book and Closed-Book Examinations. *Academic Medicine.* 91 (4): 583–599. ISSN 1040-2446. DOI:10.1097/ACM.0000000000000977.

Dyer, J. et al. 2020. *Academic Dishonesty and Testing: How Student Beliefs and Test Settings Impact Decisions to Cheat: How Student Beliefs and Test Settings Impact Decisions to Cheat*. 28. 4. 2020. Dostupné z: https://www.researchgate.net/publication/341296878_Academic_Dishonesty_and_Testing_How_Student_Beliefs_and_Test_Settings_Impact_Decisions_to_Cheat.

Educational testing service 2014. *EETS Standards for Quality and Fairness.* Dostupné také z https://www.ets.org/about/fairness.

Egarter, S. et al. 2020. Medical Assessment in the Age of Digitalisation. *BMC Medical Education*. 20(1). ISSN 1472-6920. DOI:10.1186/s12909-020-02014-7.

Embse, N. von der et al. 2018. Test Anxiety Effects, Predictors, and Correlates: A 30-year Meta-analytic Review. Journal of Affective Disorders. 227: 483–493. ISSN 01650327. DOI:10.1016/j.jad.2017.11.048.

Esemes.cz. 2014. Uzbekistán vypnul internet a SMS: Pro někoho možná radikálním krokem se rozhodli zakročit v Uzbekistánu proti korupci a podvodu. 8.8.2014. Dostupné z: https://www.esemes.cz/magazin/uzbekistan-vypnul-internet-a-sms/.

FairTest. 2007. The Limits of Standardized Tests for Diagnosing and Assisting Student Learning. *FairTest*. Jamaica Plain: National Center for Fair & Open Testing. Dostupné z: https://fairtest.org/limits-standardized-tests-diagnosing-and-assisting.

Fiřtová, L. 2021. Klonování úloh jako cesta k vyrovnání obtížnosti různých variant testu? In: *Konference Psychologická diagnostika*. Brno: MUNI FSS.

Foltýnek, T. 2021. Akademická integrita a jak ji utvářet: European Network for Academic Integrity. In: *Akademická integrita*. Slovenská akreditačná agentúra pre vysoké školstvo. Dostupné z: https://saavs.sk/wp-content/uploads/2021/06/Akademicka-integrita_Foltynnek.pdf.

Foster, D. 2013. Security Issues in Technology-Based Testing. In: *Handbook of Test Security*. London: Routledge. s. 39–83. ISBN 978-0415805643. DOI: 10.4324/9780203664803.ch3.

FSBPT 2012. *Forensic Analysis Conducted to Investigate Effect of Trafficking in Recalled Test Items Leads to Invalidation of 20 Candidate Test Scores*. The Federation of State Boards of Physical Therapy. DOI: https://www.fsbpt.org/forfaculty/yourquestions/index.asp#InvalidatedNPTEScores.

Gierl, M. J. a Haladyna, T. M. 2012. *Automatic Item Generation: Theory and Practice.* New York: Routledge. 256 s. ISBN 978-0-415-89750-1.

Gierl, M. J. a Lai, H. 2012. The Role of Item Models in Automatic Item Generation. *International Journal of Testing*. 12 (3): 273–298. ISSN 1530-5058. DOI:10.1080/15305058.2011.635830.

Gierl, M. J.et al. 2012. Using Automatic Item Generation to Create Multiple-choice Test Items. *Medical Education*. 46 (8): 757–765. ISSN 03080110. DOI:10.1111/j.1365-2923.2012.04289.x. ISSN 1365-2923. PMID: 22803753.

Hambelton, R. K. a Plake, B. S. 1995. Using an Extended Angoff Procedure to set Standards on Complex Performance Assessments. *Applied Measurement in Education.* 8 (1): 41–55. ISSN 0895-7347 (Print), 1532–4818 (Online). DOI: 10.1207/s15324818ame0801_4.

Hambleton, R. K. et al. 1991. *Fundamentals of Item Response Theory*. Newbury Park, Calif.: Sage Publications. ISBN 0803936478.

Han, K. T. 2009. IRTEQ: Windows Application That Implements IRT Scaling and Equating [computer program]. *Applied Psychological Measurement*. 33 (6): 491–493.

Hembree, R 1988. Correlates, Causes, Effects, and Treatment of Test Anxiety. *Review of Educational Research*. 58 (1): 47–77. DOI: 10.2307/1170348.

Herman, J. L. et al. 1992. *A Practical Guide to Alternative Assessment*. Association for Supervision and Curriculum Development. Alexandria, VA. ISBN-0-87120-197-6.

Herman, J. L. a Zuniga, S. A. Assessment: Portfolio Assessment. *Education Encyclopedia*. Dostupné z: https://education.stateuniversity.com/pages/1769/Assessment-PORTFOLIO-ASSESSMENT.html.

Hochlehnert, A. et al. 2012. Good Exams Made Easy: The Item Management System for Multiple Examination Formats. *BMC Medical Education*. 12 (1). ISSN 1472-6920. DOI:10.1186/1472-6920-12-63.

Homer, M. a Darling, J. C. 2016. Setting Standards in Knowledge Assessments: Comparing Ebel and Cohen via Rasch. *Medical Teacher*. 38 (12): 1267–1277. ISSN 0142-159X. DOI:10.1080/0142159X.2016.1230184.

Homer, M. et al. 2012. Psychometric Characteristics of Integrated Multi-specialty Examinations: Ebel Ratings and Unidimensionality. *Assessment & Evaluation in Higher Education*. 37 (7): 787–804. ISSN 0260-2938. DOI:10.1080/02602938.2011.573843.

Chvál, M. et al. 2015. *Hodnocení výsledků vzdělávání didaktickými testy*. Plzeň: Česká školní inspekce. s. 113. ISBN 978-80-905632-9-2.

IMS Global. Igniting Digital Assessment Innovation. *IMS Global: Learning Consortium*. Dostupné z: http://www.imsglobal.org/activity/qtiapip.

Institute for Communication and Assessment Research. *Umbrella Consortium for Assessment Networks*. Dostupné z: https://www.ucan-assess.org/.

Jacobs, L. C. a Chase, C. I. 1992. *Developing and Using Tests Effectively: a Guide for Faculty*. San Francisco: Jossey-Bass Publishers. ISBN 1-55542-481-3.

James, R. 2002. A Comparison of Norm-referencing and Criterion-referencing Methods for Determining Student Grades in Higher Education. In: *Assessing Learning in Australian Universities: Ideas, Strategies and Resources for Quality in Student Assessment.* Melbourne, Vic: Centre for the Study of Higher Education. ISBN 9780734029027.

Jelínek, M. a Květon, P. 2011. *Testování v psychologii: Teorie odpovědi na položku a počítačové adaptivní testování.* Praha: Grada. 160 s. ISBN 978-802-4735-153.

Jeřábek, O. a Bílek, M. 2010. *Teorie a praxe tvorby didaktických testů*. Olomouc: Univerzita Palackého v Olomouci. ISBN 978-80-244-2494-1.

Johanns, B. et al. 2017. A Systematic Review Comparing Open-book and Closed-book Examinations: Evaluating Efects on Development of Critical Thinking Skills. *Nurse Education in Practice*. 27: 89–94. ISSN 14715953. DOIi:10.1016/j.nepr.2017.08.018.

Jolly, B. 2010. Written Examinations . In: Sawanawik, T. *Understanding Medical Education: Theory and Practice*. Oxford : Wiley-Blackwell. 464 s. s. 208–230. DOI: 10.1002/9781444320282.ch15. ISBN 978-1-4051-9680-2.

Jørgensen M. et al. 2018. Contrasting Groups' Standard Setting for Consequences Analysis in Validity Studies: Reporting Considerations. *Advances in Simulation*. 3 (1). 1–7. ISSN 2059-0628. DOI:10.1186/s41077-018-0064-7.

Kader, A. A. 2016. Debilitating and Facilitating Test Anxiety and Student Motivation and Achievement in Principles of Microeconomics. *International Review of Economics Education*. 23: 40–46. ISSN 14773880. DOI:10.1016/j.iree.2016.07.002.

Kamalov, F. et al. 2021. Machine Learning Based Approach to Exam Cheating Detection. *PLoS ONE.* 16 (8).

Kean, J. et al. 2018. An Introduction to Item Response Theory and Rasch Analysis: Application Using the Eating Assessment Tool (EAT-10). *Brain Impairment*. 19 (1): 91–102. ISSN 1443-9646. DOI:10.1017/BrImp.2017.31.

Keith-Spiegel, P. et al. 1998. Why Professors Ignore Cheating: Opinions of a National Sample of Psychology Instructors. *Ethics & Behavior* 8 (3): 215–227. ISSN 1050-8422. DOI:10.1207/s15327019eb0803_3.

Kennedy, R. 2019. Why Students Cheat and How to Stop It. *ThoughtCo: World Largest Education Resource*. Dotdash. 16. 11. 2019. Dostupné z: https://www.thoughtco.com/cheating-basics-for-private-schools-2773348.

Klenowski, V. et al. 2006. Portfolios for Learning, Assessment and Professional Development in Higher Education. *Assessment & Evaluation in Higher Education*. 31 (3): 267–286. ISSN 0260-2938. DOI:10.1080/02602930500352816.

Klerk, S. de 2019. The Theory and Practice of Educational Data Forensics. In: *Theoretical and Practical Advances in Computer-based Educational Measurement*. *Methodology of Educational Measurement and Assessment*. Cham: Springer International Publishing. 6. 7. 2019. s. 381–399. ISBN 978-3-030-18479-7. DOI:10.1007/978-3-030-18480-3_20.

Kline, P. 1995. *The Handbook of Psychological Testing*. London: Routledge.

Kolen, M. J. et al. 2004. Test Equating, Scaling, and Linking: Methods and Practices. 2. vydání. New York: Springer. XXVI. 548 s. ISBN 0-387-40086-9.

Komenda, M. a Pokorná, A. 2011. *Benefity a úskalí elektronického testování.* Brno: Masarykova univerzita. Dostupné také z https://www.mefanet.cz/res/file/publikace/benefity-uskali-elektronickeho-testovani.pdf.

Korviny, P. a Foltyn, R. 2012. LMS Moodle v clusteru. In: *EUNIS-CZ. Open Source na vysokých školách: sborník příspěvků ke konferenci : Špindlerův Mlýn 23.–25.9.2012*. Západočeská univerzita. 71 s. ISBN 8026101499, 9788026101499.

Korviny, P. et al. 2009. *LMS Moodle na více serverech*. 443 s. s. 239–244. Proceedings of International Conferences: ICT Bridges, Sunflower 2009, Silesian Moodle Moot 2009. ISBN 978-80-248-2117-7. Dostupné také z : http://korviny.cz/clanky_pdf/smm2009-korviny_foltyn_kempny-clanek.pdf.

Kosh, A. E. et al. 2018. A Cost–Benefit Analysis of Automatic Item Generation. *Educational Measurement: Issues and Practice*. 38 (1): 48–53. ISSN 0731-1745. DOI:10.1111/emip.12237.

Krevič, N. 2019. Katalyzátor změn vyučování?: Inovace v hodnocení žáků. *Pro vzdělávání: Školské poradenské zařízení a zařízení pro další vzdělávání pedagogických pracovníků*. Praha: Národní ústav pro vzdělávání. Dostupné z: http://provzdelavani.nuv.cz/clanky/ze-zahranici/katalyzator-zmen-vyucovani-inovace-v-hodnoceni-za.

Kubiszyn, T. a Borich, G. 2000. *Educational Testing and Measurement.* Wiley. 530 s. ISBN 9780471364962.

Lafave, M. et al. 2008. Development of a Content-valid Standardized Orthopedic Assessment Tool (SOAT). *Advances in health sciences education : theory and practice.* 13 (4): 397–406. ISSN (Print) 1382-4996, (Online) 1573-1677. PMID: 17203268.

Linden, W. J. van der a Sotaridona, L. 2006. Detecting Answer Copying When the Regular Response Process Follows a Known Response Model. *Journal of Educational and Behavioral Statistics*. 31 (3): 283–304. ISSN 1076-9986. DOI:10.3102/10769986031003283.

Lupien, S. J. et al. 2007. The Effects of Stress and Stress Hormones on Human Cognition: Implications for the Field of Brain and Cognition. *Brain and Cognition*. 65 (3): 209–237. CiteSeerX 10.1.1.459.1378. DOI:10.1016/j.bandc.2007.02.007. PMID 17466428. S2CID 5778988.

Ma, Y.et al. 2013. Students' Academic Cheating in Chinese Universities: Prevalence, Influencing Factors, and Proposed Action. *Journal of Academic Ethics*. 11 (3): 169–184. ISSN 1570-1727. DOI:10.1007/s10805-013-9186-7.

Madsen, H. S. 1982. Determining the Debilitative Impact of Test Anxiety. *Language Learning*. 32 (1): 133–143. ISSN 00238333. DOI:10.1111/j.1467-1770.1982.tb00522.x.

Maierová, E. et al. 2015. Položková analýza testů studijních předpokladů jako součást zkvalitňování procesu přijímání na vysokou školu. In: *PHD EXISTENCE 2015: česko-slovenská psychologická konference (nejen) pro doktorandy a o doktorandech*. Olomouc: Univerzita Palackého v Olomouci, Filozofická fakulta. s. 75–84. ISBN 978-80-244-4694-3.

Malamed, C. 2020. Alternatives to Bloom's Taxonomy for Workplace Learning. *The eLearning Coach: Helping You Design Smarter Learning Experiences* . Dostupné z: https://theelearningcoach.com /elearning_design/alternatives-to-blooms-taxonomy/.

Martinková, P. a Drabinová, A. 2018. ShinyItemAnalysis for Teaching Psychometrics and to Enforce Routine Analysis of Educational Tests. *The R Journal.* 10 (2): 503–515. DOI: 10.32614/RJ-2018-074.

Martinková, P. et al. 2017a. ShinyItemAnalysis: Analýza přijímacích a jiných znalostních či psychologických testů. *Testfórum*. 6 (9): 16–35. DOI: 10.5817/TF2017-9-129.

Martinková, P. et al. 2017b. Semi-real-time Analyses of Item Characteristics for Medical School Admission Tests. *ACSIS*. 24. 9. 2017. s. 189–194. DOI:10.15439/2017F380.

Martinková, P. et al. 2021. Psychometrická analýza interaktivně a v R: Co je nového v ShinyItemAnalysis. In: *Konference Psychologická diagnostika*. Brno: MUNI, FSS.

Maynes, D. 2013. Educator Cheating and the Statistical Detection of Group-based Test Security Threats. In: *Handbook of Test Security.* s. 187–214. New York: Routledge, Psychology Press. ISBN 978-0-203-66480-3.

McCabe, D. L. 2001. Cheating in Academic Institutions: A Decade of Research. *Ethics & Behavior*. 11 (3): 219–232.

McCabe, D.L. et al. 2012. *Cheating in College*. The Johns Hopkins University Press. ISBN 9781421407166. Dostupné z: http://www.scopus.com/inward/citedby.url?scp=84906003037&partnerID=8YFLogxK.

McKinley, D. W. a Norcini, J. J. 2014. How to Set Standards on Performance-based Examinations: AMEE Guide No. 85. *Medical Teacher*. 36 (2): 97–110 [cit. 2021-11-29]. ISSN 0142-159X. DOI:10.3109/0142159X.2013.853119.

McLachlan, J. C. a Whiten, S. C. 2000. Marks, Scores and Grades: Scaling and Aggregating Student Assessment Outcomes. *Medical Education*. 34 (10): 788–797. ISSN 0308-0110. DOI: 10.1046/j.1365-2923.2000.00664.x.

Mealey, D. L. a Host, T. R. 1992. Coping with Test Anxiety. College Teaching. 40 (4): 147–150. ISSN 8756-7555. DOI:10.1080/87567555.1992.10532238.

Medical Schools Council. *MSC Assessment Alliance*. London: Medical Schools Council. Dostupné z: https://www.medschools.ac.uk/our-work/assessment/msc-assessment-alliance.

Mezera, A. *Školní měření a evaluace výsledků vzdělávání ve škole: Studijní materiál pro interní potřebu učitelů základních a středních škol*. Dostupné z: http://www.pppprаha7a8.cz/files/zaklady%20skolniho%20mereni.pdf.

Miller, G. E. 1990. The Assessment of Clinical Skills/Competence/Performance. *Academical Medicine*. 65 (9): 63–7. ISSN 1040-2446. PMID: 2400509.

Murphy, K. R. a Davidshofer, C. O. 2005. *Psychological testing: principles and applications*. Upper Saddle River, N.J., Pearson/Prentice Hall. ISBN 0-13-189172-3.

Nakonečný, M. 1992. *Motivace pracovního jednání a její řízení*. Praha: Management Press.

Nelson, L.R. 2017. Item Analysis Software for Classes: Measurement Classes with Lertap 5, jMetrik, SAS University, BILOG-MG, and Xcalibre. In: Curtin University. DOI:10.13140/RG.2.2.32532.71049. Dostupné z: https://www.researchgate.net/publication/317087828_Item_analysis_software_for_classes.

Nigam, A. et al. 2021. A Systematic Review on AI-based Proctoring Systems: Past, Present and Future. *Education and Information Technologies*. 26 (5): 6421–6445. ISSN 1360-2357. DOI:10.1007/s10639-021-10597-x.

Norcini, J.J. 2003. Setting Standards on Educational Tests. *Medical Education*. 37 (5): 464–469. ISSN 03080110. DOI:10.1046/j.1365-2923.2003.01495.x.

O´Neil, G. a Murphy F. 2010. *Assessment: Guide to Taxonomies of Learning*. Dublin: UCD Teaching and Learning. Dostupné z: https://www.ucd.ie/t4cms/ucdtla0034.pdf.

Online Education Database (OEDb) 2010. *8 Astonishing Stats on Academic Cheating*. Dostupné z: http://oedb.org/library/features/8-astonishing-stats-on-academic-cheating.

Patil, S. Y. et al. 2015. Blueprinting in Assessment: A Tool to Increase the Validity of Undergraduate Written Examinations in Pathology. *International Journal of Applied and Basic Medical Research*. 5(4): 76. ISSN 2229-516x. DOI:10.4103/2229-516X.162286.

Peñalver, E. A. 2015. Financial Translation. In: *Handbook of Research on Teaching Methods in Language Translation and Interpretation*. IGI Global:Advances in Educational Technologies and Instructional Design. s. 102–117. ISBN 9781466666153. DOI:10.4018/978-1-4666-6615-3.ch007.

Phelps, R. 2002. Estimating the Costs and Benefits of Educational Testing Programs. In: *The Education Consumers Consultants Network: Issues in Public Education: Research and Analysis from the Education Consumers Foundation*. Arlington: Education Consumers Foundation. Dostupné z: http://education-consumers.org/research/briefs_0202.htm.

Picus, L. O. et al. 1996. Estimating the Costs of Student Assessment in North Carolina and Kentucky: A State-Level Analysis. In: *CSE Report 408: National Center for Research on Evaluation, Standards, and Student Testing (CRESST)*. Los Angeles: University of California. Dostupné z: https://cresst.org/publications/cresst-publication-2780/.

Prakash, J. et al. 2020. Workplace Based Assessment: A Review of Available Tools and Their Relevance. *Industrial Psychiatry Journal*. 29 (2). ISSN 0972-6748. DOI:10.4103/ipj.ipj_225_20.

Ranger, J. et al. 2020. The Detection of Cheating on E-Exams in Higher Education-The Performance of Several Old and Some New Indicators. *Frontiers in Psychology*. 11: 568825. DOI.

Riffert, F. 2005. The Use and Misuse of Standardized Testing: A Whiteheadian Point of View. *Interchange*. Salzburg: University of Salzburg. 36 (1–2): 231–252. ISSN 0826-4805. DOI:10.1007/s10780-005-2360-0.

Rudner, L. M. 2010. Implementing the Graduate Management Admission Test Computerised Adaptive Test. In: *Elements of Adaptive Testing*. New York : Springer. s. 151–165. Dostupné také z: https://link.springer.com/chapter/10.1007/978-0-387-85461-8_8. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8.

Ryan, J. a Brockmann, F. 2009. *A Practitioner's Introduction to Equating: With Primers on Classical Test Theory and Item Response Theory*. Washington: Council of Chief State School Officers.

Sam, A. H. et al. 2020. High-stakes, Remote-access, Open-book Examinations. *Medical Education*. 54 (8): 767–768. ISSN 0308-0110. DOI:10.1111/medu.14247.

Seifert, K. 2011. Advantages and Disadvantages. *Educational Psychology*. OpenStax CNX:318-320. Dostupné z: https://www.opentextbooks.org.hk/ditatopic/6468.

Schindler, R. 2006. *Rukověť autora testových úloh*. Praha: Centrum pro zjišťování výsledků vzdělávání. ISBN 80-239-711-5.

Schmidt, F. L. a Hunter, J. E. 1977. Development of a General Solution to the Problem of Validity Generalization. *Journal of Applied Psychology*. 62: 529–540.

Schuwirth, L. W. T. a, Vleuten, C. P. M. van der 2011. General Overview of the Theories Used in Assessment: AMEE Guide No. 57. *Medical Teacher*. 33 (10): 783–797. ISSN 0142-159X. DOI:10.3109/0142159X.2011.611022.

Simmons, A. 2018. *Why Students Cheat—and What to Do About It: A teacher seeks answers from researchers and psychologists*. George Lucas Educational Foundation: Classroom Management. Dostupné z: https://www.edutopia.org/article/why-students-cheat-and-what-do-about-it.

Sims, R. L. 2010. The Relationship Between Academic Dishonesty and Unethical Business Practices. *Journal of Education for Business*. 68(4): 207–211. ISSN 0883-2323. DOI:10.1080/08832323.1993.10117614.

Soozandehfar, S. M. A. a Adeli, M. R. 2016. A Critical Appraisal of Bloom's Taxonomy. *American Research Journal of English and Literature*: An Academic Publishing House. 2016 (2): 1–9. ISSN 2378-9026. Dostupné z: https://www.arjonline.org/papers/arjel/v2-i1/14.pdf.

Sotaridona, L. S. a Meijer, R. R. 2003. Two New Statistics to Detect Answer Copying. *Journal of Educational Measurement*. 40 (1): 53–69. ISSN 0022-0655. DOI:10.1111/j.1745-3984.2003.tb01096.x.

Stemler, S. E. a Naples, A. 2021. Rasch Measurement v. Item Response Theory: Knowing When to Cross the Line. *Practical Assessment, Research, and Evaluation*. 26 (11). ISSN 1531-7714. DOI:10.7275/v2gd-4441.

Stitt-Bergh, M. a Hill, Y. What Is a Portfolio?: Using Portfolios in Program Assessment. *Assessment and Curriculum Support Center: Learning outcomes assessment for improvement*. Manoa, Hawaii: University of Hawai'i at Mānoa. Dostupné z: https://manoa.hawaii.edu/assessment/resources/using-portfolios-in-program-assessment/.

Streiner, D. L. 2003. Starting at the Beginning: An Introduction to Coefficient Alpha and Internal Consistency. *Journal of Personality Assessment*. 80 (1): 99–103. ISSN 0022-3891. DOI:10.1207/S15327752JPA8001_18.

Swerdlik, M. et al. 2012. *Psychological Testing and Assessment : An Introduction to Tests and Measurement.* McGraw-Hill Education. 752 s. ISBN 9780078035302.

Štuka, Č. et al. 2012. The Prediction and Probability for Successful Completion in Medical Sudy Based on Tests and Pre-admission Grades. *The New Educational Review*. 28 (2):138–152. Dostupné také z: http://www.educationalrev.us.edu.pl/vol/tner_2_2012.pdf. ISSN 1732-6729.

Štuka, Č. et al. 2013. *Testování při výuce medicíny: konstrukce a analýza testů na lékařských fakultách*. Praha: Karolinum. 155 s. ISBN 978-80-246-2369-6.

Tavakol, M. a Dennick, R. 2011a. *Post Examination Analysis of Objective Tests.* AMEE 2011. AMEE guide: sv. 54. ISBN 978-1-903934-91-3.

Tavakol, M. a Dennick, R. 2011b. Making Sense of Cronbach's Alpha. *International Journal of Medical Education*. 2: 53–55. ISSN 20426372. DOI:10.5116/ijme.4dfb.8dfd.

Tavakol, M. a Dennick, R. 2013. Psychometric Evaluation of a Knowledge Based Examination Using Rasch Analysis: An Illustrative Guide. *Medical Teacher*. 35 (1): e838–e848. ISSN 0142-159x. DOI :10.3109/0142159X.2012.737488.

Teachers Commons 2008. Criticisms of Bloom's Taxonomy: Educational Theorists Have Criticized Bloom's Taxonomy on a Few Grounds. *Teachers Commons: A Place for Teachers to Share*. 24.4.2008. Dostupné z: http://teachercommons.blogspot.com/2008/04/bloom-taxonomy-criticisms.html.

Tendeiro, J. N. et al. 2016. PerFit: An R Package for Person-Fit Analysis in IRT. *Journal of Statistical Software* 74 (5): 1–27. ISSN 1548-7660. DOI:10.18637/jss.v074.i05.

The Royal College of Pathologists 2019. Definitions: Workplace-based Assessment (WPBA). *Assessment Department*. Dostupné z: https://www.rcpath.org/trainees/assessment/workplace-based-assessment-wpba.html.

The University of Kansas. *Testwiseness and Guessing: What is Testwiseness and Guessing?*. Lawrence: The University of Kansas. Dostupné z: http://www.specialconnections.ku.edu/?q=assessment/quality_test_construction/teacher_tools/testwiseness_and_guessing.

Thompson, N. 2016. SIFT: A New Tool for Statistical Detection of Test Fraud: SIFT: Software for Investigating Test Fraud. *Assessment Systems Corporation (ASC)*. Dostupné z: https://assess.com/sift-new-tool-statistical-detection-test-fraud/.

Thompson, N. 2019. What Are the Possible Transformations for Scaled Scoring? *Assessment Systems Corporation (ASC): Psychometrics*. Dostupné z: https://assess.com/2019/07/13/what-are-the-possible-transformations-for-scaled-scoring/.

Tutkun, O. et al. 2012. Bloom‘s Revized Taxonomy and Critics on It. In: *The Online Journal of Counselling and Education*. s. 23-30. ISSN 2146-8192. Dostupné z: https://www.researchgate.net/publication/299850265_Bloom‘s_Revized_Taxonomy_and_Critics_on_It.

University of North Texas. Why Do Students Cheat? *UNT Teaching Commons: Center for Learning Experimentation, Application, and Research*. Dostupné z: https://teachingcommons.unt.edu/teaching-essentials/academic-integrity/why-do-students-cheat.

Univerzita Karlova 2005. *Výroční zpráva o činnosti za rok 2005*. Univerzita Karlova v Praze, Právnická fakulta. Praha. Dostupné z: https://www.prf.cuni.cz/dokumenty-download/1404044551/.

Verschoor, A. a Jongkamp, C. 2016. *Item Banking for Optimal Tests: AEA Europe pre-conference workshop*. Prague.

Violato, C. et al. 2002. Certification Examinations for Massage Therapists: A Psychometric Analysis. *Journal of Manipulative Physiological Therapeutics*. 25 (2): 111–115. Dostupné také z http://www.jmptonline.org/article/S0161-4754(02)70455-7/fulltext. ISSN 0161-4754. DOI: 10.1067/mmt.2002.121413.

Violato, C. et al. 2003. A Validity Study of Expert Judgement Procedures for Setting Cutoff Scores on High Stakes Credentialing Examinations Using Cluster Analysis. *Evaluation and the Health Professions*. 26 (1): 59–72. Dostupné také z http://www.internationalgme.org/Resources/Pubs/Validity%20Cutoff%20Scores%20-%20Violato.pdf. ISSN (Print) 0163–2787; (Online) 1552–3918. PMID: 22973420. DOI: 10.1177/0163278702250082.

Vlčková. K. 2014. *Férovost didaktických testů a jejich položek*. Diplomová práce. Praha: MFF UK.

Vleuten, C. P. M. van der 1996. The Assessment of Professional Competence: Developments, Research and Practical Implications. *Advances in Health Science Education.* 1 (1): 41–67. DOI: 10.1007/BF00596229.

Vleuten, C. van der. 2019a. OSCEs by Cees van der Vleuten. *Maastricht University*. Dostupné z: https://www.maastrichtuniversity.nl/news-events/newsletters/article/+5u+DZKHLUQtFBjwefD8Tg.

Vleuten, C. van der. 2019b. *Automatic Item Generation by Cees van der Vleuten*. Maastricht University. Dostupné z: https://www.maastrichtuniversity.nl/news-events/newsletters/article/NyJydZFCFpcpCYHi4Fadew.

Vleuten, C. P. M. van der a Schuwirth, L. W. T. 2005. Assessing Professional Competence: from Methods to Pogrammes. *Medical Education*. 39 (3): 309–317. DOI: 10.1111/j.1365-2929.2005.02094.x.

Vleuten, C. P. van der et al 1991. Pitfalls in the Pursuit of Objectivity: Issues of Reliability. *Med Educ.* 25 (2): 110–8 ISSN 0308-0110. Dostupné také z: https://www.ncbi.nlm.nih.gov/pubmed/2023552.

Vowell, P. R. a Chen, J. 2004. Predicting Academic Misconduct: A Comparative Test of Four Sociological Explanations. *Sociological Inquiry*. 74 (2): 226–249. ISSN 0038-0245. DOI: 10.1111/j.1475-682X.2004.00088.x.

Vrbová, J. 2013. „Co mi ve škole vadí víc, podvádění, či klamání?“ Postoje žáků k nečestnému chování ve škole v kontextu školního podvádění. *Studia paedagogica* 18 (2–3). ISSN 1803-7437. DOI:10.5817/SP2013-2-3-6.

Weems, C.F. et al. 2010. Test Anxiety Prevention and Intervention Programs in Schools: Program Development and Rationale. *School Mental Health.* 2 (2): 62–71. ISSN 1866-2625. DOI:10.1007/s12310-010-9032-7.

Weiss, D. J. 2011. Item Banking, Test Development, and Test Delivery. In: *The APA Handbook on Testing and Assessment in Psychology*. Washington DC: American Psychological Association. ISBN 978-1-4338-1227-9.

Wikipedia. 2001. Psychometric Software. *Wikipedia: the Free Encyclopedia*. San Francisco (CA): Wikimedia Foundation. Dostupné z: https://en.wikipedia.org/wiki/Psychometric_software.

Wikipedia. 2021. Grading Systems by Country. *Wikipedia: the Free Encyclopedia*. San Francisco (CA): Wikimedia Foundation. Dostupné z: https://en.wikipedia.org/wiki/Grading_systems_by_country.

Wilson, S. 2012. *Rogō: an Open Source Solution for High-stakes Assessment*. OSS Watch Team Blog: Open Source Software Advisory Service. Dostupné z: http://osswatch.jiscinvolve.org/wp/2012/09/13/Rogo-an-open-source-solution-for-high-stakes-assessment/.

Winter, T. 2019. College Cheating Ringleader Says He Helped More Than 750 Families with Admissions Scheme. *NBC NEWS.* 13. 3. 2019. Dostupné z: https://www.nbcnews.com/news/us-news/college-cheating-mastermind-says-he-helped-nearly-800-families-admissions-n982666.

Wollack, J. A. 1997. A Nominal Response Model Approach for Detecting Answer Copying. *Applied Psychological Measurement.* 21 (4): 307–320 [cit. 2021-10-6]. ISSN 0146-6216. DOI:10.1177/01466216970214002.

Yang, B. W. et al. 2019. Using Testing as a Learning Tool. *American Journal of Pharmaceutical Education*. 83 (9):.7324. DOI: 10.5688/ajpe7324. PMID: 31871352. PMCID: PMC6920642.

Yerkes R. M. a Dodson J. D. 1908. The Relation of Strength of Stimulus to Rapidity of Habit-formation. *Journal of Comparative Neurology and Psychology.* 18: 459–482. DOI:10.1002/cne.920180503.

Zagury-Orly, I. a Durning, S. J. 2021. Assessing Open-book Examination in Medical Education: The Time Is Now. *Medical Teacher*. 43 (8): 972–973. ISSN 0142-159X. DOI:10.1080/014215 9X.2020.1811214.

Zvára, K. 2008. *Regrese.* Praha: MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze. 254 s. ISBN 978-80-7378-041-8.

14 REJSTŘÍK