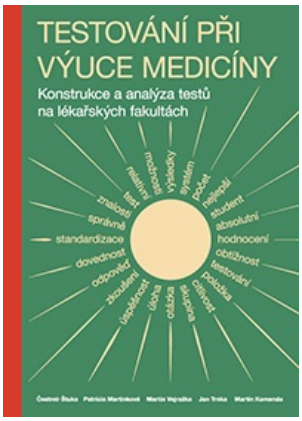
Fórum:Testy

Testování při výuce medicíny

Konstrukce a analýza testů na lékařských fakultách



Štuka, Č., Martinková, P., Vejražka, M., Trnka J., Komenda M.: *Testování při výuce medicíny, Konstrukce a analýza testů na lékařských fakultách*. Praha: Karolinum, 2013.,
ISBN 978-80-246-2369-6

Informační technologie jsou dnes přirozenou součástí výuky medicíny. Díky osobní důvěře mezi pracovníky lékařských fakult zodpovědnými za elektronickou podporu vzdělávání vznikl projekt MEFANET (MEdical FAculties NETwork), který zastřešuje spolupráci fakult v této oblasti. Tato vzdělávací síť dnes sdružuje všechny lékařské a zdravotnické fakulty v České republice a na Slovensku, organizuje odborné konference a podporuje spolupráci a rozvoj výuky s pomocí moderních informačních technologií. Nově se na této platformě rozvíjí spolupráce i v oblasti přípravy, sdílení a analýzy testů.

Moderní metody hodnocení znalostí studentů patří v akademickém světě k velmi diskutovaným tématům. Standardizované testování umožňuje spravedlivě, prokazatelně a objektivně ohodnotit znalosti a dovednosti studentů. Autoři publikace se snaží využít spolupráce mezi fakultami a přispět k rozšíření nových metod testování. Kniha přináší souhrn postupů, doporučení a metod pro tvorbu kvalitních didaktických testů. Část zaměřená na statistiku snímá z analýzy testů nálepku „terra incognita“ a ukazuje nástroje, které jsou k dispozici. Celý text je zaměřen tak, aby čtenář získal přehled o metodách hodnocení výsledků výuky a mohl držet krok s vývojem v rychle se měnícím světě lékařského a zdravotnického vzdělávání.

Poděkování

Autoři děkují svým blízkým za toleranci a podporu a svým spolupracovníkům za inspiraci a pomoc. Zvláštní dík patří recenzentům, jmenovitě PhDr. Andree Pokorné, Ph.D., PhDr. Tamaře Váňové a doc. RNDr. Karlu Zvárovi, CSc., za cenné a obsáhlé připomínky, které svým rozsahem a hloubkou dílo zásadně obohatily.

Druhé vydání elektronická verze knihy se připravuje na adrese: **www.wikiskripta.eu/Testy**. Za všechny Vaše náměty a připomínky k tomuto textu zaslané na adresu testy@wikiskripta.eu Vám budou autoři vděční.

Souhrn

Kniha se zabývá konstrukcí a analýzou testů, se zvláštním zřetelem k potřebám testování znalostí na lékařských fakultách. Je míněna jako praktická pomůcka a návod pro pedagogy, kteří sami znalostní testy vytvářejí a využívají, nebo se k tomu chystají. Kniha se snaží být také uceleným vodítkem pro vedení fakult, které o koncepci testování rozhoduje na vyšší úrovni, například pro potřeby přijímacích testů nebo atestací.

Testování znalostí je klíčovým momentem v procesu vzdělávání. V mnoha případech rozhoduje o dalším osudu studenta, o tom, zda je či není přijat na danou vysokou školu či zda získá či nezíská akademický titul. Testy jsou ale důležité také pro samotnou vzdělávací instituci a celou společnost: na jejich základě mohou být vybíráni ti nejvhodnější adepti pro studium medicíny a pro výkon lékařské praxe. Význam správně sestavených testů není zanedbatelný ani na úrovni jednotlivých vyučovaných kurzů a samotného vzdělávacího procesu: často právě forma a obsah testu rozhoduje o tom, na které znalosti či dovednosti se student zaměří a naučí se je. Dobře připravené testy jsou proto důležitým nástrojem kvalitní výuky.

Kniha se podrobně věnuje způsobům hodnocení znalostí, do detailu rozebírá možnosti, které poskytuje písemné či elektronické zkoušení. Popisuje tvorbu testů a jednotlivé druhy položek, jejich výhody i rizika použití. Velký prostor je věnován novým formátům otázek, které umožňují zkoušet nejen izolovanou znalost faktů, ale i míru pochopení látky. V zorném poli však není jen tvorba otázek, ale i celý cyklus přípravy testu: stanovení cílů výuky a plánování testu, příprava položek, jejich oponování a pilotního odzkoušení, vyhodnocování kvality položek a celého testu, standardizace a převod výsledků testů do formy klasifikace.

Pro přípravu kvalitních otázek a testů je nezbytná koordinovaná spolupráce týmu odborníků, jejichž vynaložený čas se promítá do prostředků vynaložených na položku. Jako cesta ke snížení nákladů na tvorbu položek a testů jsou v publikaci popsány možnosti zabezpečeného sdílení otázek pomocí položkových bank včetně příkladů úspěšných sítí pro sdílení testových položek a jejich organizačních a technických řešení. Probrány jsou i vybrané softwarové nástroje pro testování a pro hodnocení kvality testů a jejich položek.

Kniha čtenáři zodpoví, kdy zkoušet ústně a kdy písemně, kdy je vhodné počítačové testování a jak jej lze zavést. Poradí, jak správně nastavit klasifikační stupnici, poskytuje návod, jak numericky zhodnotit obtížnost i kvalitu testu a jeho položek, nebo pomůže odhalit, zda některé položky či celé testy nebyly studentům předem známy. Přestože je zaměřená především na testování znalostí v oblasti medicíny, metody a principy obsažené v publikaci naleznou své uplatnění na celém širokém poli znalostního i psychologického testování.

Spolupráce autorů z více vysokých škol na tomto tématu odráží společné úsilí zavést moderní metody testování do praxe vlastních pracovišť i poskytnout odrazový můstek ostatním, kteří se o tuto tématiku zajímají.

Student Assessment in Medical Education

Construction and analysis of tests at medical faculties

Summary

This book is about the construction and analysis of tests with a specific emphasis on testing needs of medical schools. It should serve as a practical aid and manual for teachers who prepare tests and use them to assess students in an educational setting. This book should also become a guideline for the leadership of medical schools in their high-level decisions about tests for entrance or board exams.

Testing of students' knowledge is a key moment in the educational process. In many cases it determines their future paths such as whether they will be admitted to a study programmes or whether they graduate. Tests are also important for educational institutions themselves and for the society as a whole as they help choose the best candidates for medical study programmes and for medical practice. Appropriately composed tests also play a significant role in individual study courses and the whole educational process as the form and content of a test often determines on which areas students will focus and which skills they will master. Well-designed tests are therefore important tools for good education.

Our book presents a detailed treatment of knowledge assessment methods including possibilities of written and electronic testing. It describes various types of test items, their advantages and their risks. A significant portion of the text is devoted to new question formats that allow the testing of conceptual understanding over isolated facts. The book doesn't deal only with the construction of test questions but describes the whole cycle of test preparation including the setting of teaching goals, test blueprinting, appropriate item construction, item review and test piloting. It looks at item and test quality analysis, standardisation and mapping of test results onto commonly used marking schemes.

The preparation of well-designed questions and tests requires a team of experts and therefore also significant amounts of time and other resources. One way of decreasing the resource-intensiveness of item and test construction is their secure sharing through item banks, which are also discussed including examples of successful networks for test item sharing with relevant organisational and technical solutions. The book also covers available software tools for student knowledge assessment and for test and item quality analysis.

This book will explain when to use written or oral exams, where computer-based testing may be appropriate and how to implement it. It will help you set up a good marking scheme, quantitatively analyse the difficulty and quality of your tests and of individual items and find out if any items or whole tests may have been leaked to students beforehand. Although we focus on knowledge testing in medicine the methods and principles described here are applicable to the broader area of educational and psychological testing.

Jaké otázky vám tato kniha pomůže zodpovědět?

- **Zkoušet ústně, nebo písemně?**

  kapitola 1.3 Jak zkoušet

- **Zkoušeli bychom písemně, ale nemáme testové otázky.**

  kapitola 4 Typy otázek a jejich vytváření

- **Zkoušeli bychom písemně, ale nemáme čas a sílu to zavést.**

  kapitoly 2 Cyklus přípravy testu, 9 Realizace testů

- **Zkoušíme písemně, ale nevíme, jestli dobře.**

  kapitola 8 Analýza výsledků a hodnocení kvality testu

- **Zkoušíme písemně, ale chceme to dělat lépe.**

  kapitoly 3 Plánování testu, 5 Oponentura otázek, 7 Standardizace a normování testu

- **Jsou naše testy příliš obtížné nebo snadné?**

  kapitoly 8.1 Popis a grafické zobrazení výsledků, 8.3.1 Obtížnost položky

- **Jak poznám špatnou otázku?**

  kapitoly 5 Oponentura otázek, 6 Pilotování testu, 8.3 Položková analýza

- **Jak poznám, že test byl studentům předem znám?**

  kapitoly 8.1 Popis a grafické zobrazení výsledků, 9.1 Zabezpečení testů

- **Jak správně nastavím klasifikační stupnici?**

  kapitola 7.6 Klasifikace studentů

- **Jsou (počítačové) testy vhodné pro všechny studenty a všechny předměty?**

  kapitoly 1.3 Jak zkoušet, 11.4 Testování handicapovaných studentů

- **Rádi bychom testovali moderněji, ale nemáme vhodný software.**

  kapitola 9.2 Počítačové testování

Dříve než začnete

Hodnocení studentů je běžnou součástí práce vysokoškolského pedagoga. Může se zdát, že na něm není nic složitého. Jak ale vyzkoušet velké množství studentů v krátkém čase? Nebo jak prokázat objektivitu a reprodukovatelnost zkoušení? V životě pedagoga tak přicházejí chvíle, kdy je třeba vykročit ze zaběhnutého rámce a vydat se na dobrodružnou cestu neznámým terénem testování. Ačkoli se tato cesta může zpočátku jevit jako džungle pojmů, standardů a houští statistických metod, nemusí být zcela neschůdná. Pojďme se společně do této džungle vypravit.

Než se ale vydáme na cestu, měli bychom si odpovědět na otázky **proč**, **jak**, **co** a **s kým** chceme vlastně zkoušet.

Proč

Velmi často zkoušíme, abychom zjistili, zda student dostatečně zvládl náplň daného předmětu, případně zda může či nemůže postoupit do dalšího stupně studia. Tento typ zkoušení se označuje jako výstupní neboli **sumativní** a jeho primárním výstupem je **hodnocení** výkonu studenta (splnil/nesplnil, známka, počet bodů či umístění v rámci testované skupiny).

Různé formy zkoušení však také mohou sloužit jako zdroj **zpětné vazby** pro učitele či studenty. Například nás může zajímat, nakolik studenti v průběhu výuky vstřebávají a chápou předkládaný obsah, či které oblasti jim činí největší potíže a kde tedy jako učitelé musíme přidat. Můžeme také studenty testovat proto, aby oni sami našli svoje slabé stránky a mohli na nich zapracovat. Tomuto typu zkoušení se říká průběžné, nebo též **formativní**.

V rámci jednoho předmětu, kurzu či ročníku je možné a vhodné tyto dva typy zkoušení kombinovat, zejména na začátku studia a u rozsáhlejších předmětů. Účel zkoušení nám také pomůže určit požadovanou úroveň znalostí – jinak přísní budeme u jednoho z mnoha průběžných testů a jinak u závěrečné zkoušky, která může rozhodnout o ukončení studia.

Co

Jednou z nejdůležitějších otázek, které je vhodné zodpovědět, je, co přesně chceme zkoušet. Jednoduchá odpověď by mohla znít, že v biochemii budeme zkoušet biochemii a v patologii – co jiného než patologii. Z hlediska plánování formy zkoušení se však vyplatí na věc podívat detailněji.

V průběhu výuky na vysoké škole se učitelé snaží studentům předávat směs znalostí, dovedností a postojů namixovaných v různých poměrech podle konkrétního předmětu. V biochemii se tak studenti učí intermediáty Krebsova cyklu (znalosti), chemické výpočty, a třeba i základy experimentální práce v laboratorním praktiku (dovednosti, případně i správné postoje k poctivému nakládání s experimentálními daty). V kurzech komunikace či etiky získávají dovednosti nutné k dobré komunikaci s pacienty, svými rodinnými příslušníky nebo kolegy, osvojují si profesionální postoje nutné k řešení složitých situací a jistě získají i teoretické poznatky o komunikačních technikách a etických systémech. Zásadní otázkou tedy je, co vše musí studenti prokazatelně zvládnout, aby daný předmět absolvovali, a co tedy budeme chtít zkoušet. Obvykle to nebude celý obsah kurzu, ale jen jeho centrální část. „Co se považuje za důležité“ by se mělo odvíjet od představy, jak má vypadat absolvent celého studia, a od potřeb dalších navazujících předmětů.

Úrovně znalostí a dovedností

Stejně jako musíme být schopní definovat obsahovou náplň zkoušky či testu, je také třeba přesně odpovědět na otázku, **jakou úroveň znalostí nebo dovedností** chceme testovat. Zkoušíme-li určité odborné téma, můžeme po studentovi požadovat, aby prokázal [1]

- **Znalost** (student **zná**; v anglicky psané literatuře *knowledge*, úroveň *knows*)

  > *Zkoušíme diagnostiku plicní embolie. Za **znalost** považujeme, že student má nastudované teoretické údaje o plicní embolii, zná nejčastější zdroje embolizace, rizikové faktory, patofyziologii změn, k nimž embolie vede atd.*

- **Porozumění** (student **ví jak**; *competence*, *knows how*). Zkoušený dokáže znalosti z předchozí úrovně **zapojit** do kontextu.

  > *Dokáže určit, jaké výsledky zobrazovacích metod jsou kompatibilní s diagnózou plicní embolie, jaké jsou očekávané výsledky jednotlivých laboratorních a klinických vyšetření apod.*

- **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.

  > *Určí diagnózu nebo vysloví podezření na plicní embolii na základě předloženého popisu konkrétního případu, rentgenových snímků a výsledků laboratorních vyšetření.*

- **Činnost** (v praxi **provádí** správně veškeré potřebné úkony; *action*, *does*). Této úrovně by měl dosáhnout např. kandidát u státní závěrečné zkoušky nebo u atestace.

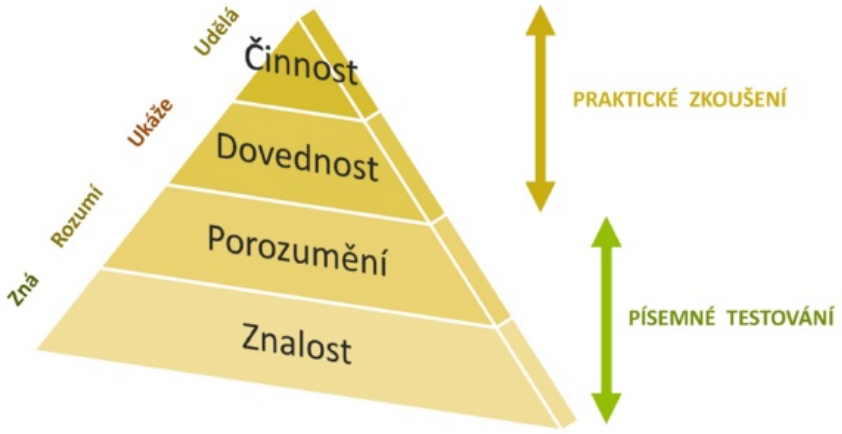
  > *Dokáže v praxi od pacienta odebrat anamnézu, fyzikálně jej vyšetřit, správně zajistit, ordinovat adekvátní vyšetření, interpretovat jejich výsledky, ordinovat správnou léčbu atd.*

💡 **Citát: Konfucius (4. stol. př.n.l.)**

V teoretických předmětech se od studenta lékařství bude vyžadovat pouze *znalost* a *porozumění*. Student jiného oboru, který se připravuje na dráhu vědeckého pracovníka, by ale měl ve stejném „teoretickém“ předmětu dosáhnout všech úrovní dovedností.

Znalosti a *porozumění* medika získané v teoretických předmětech by nicméně měly být aplikovány později v klinických oborech. Mohou tedy být podkladem pro *dovednosti* a *činnosti* zkoušené ve vyšším ročníku v rámci jiné části studia.

Výše uvedené čtyři úrovně znalostí a dovedností se používají při výuce medicíny; často se schematicky znázorňují jako tzv. Millerova pyramida (viz obr. 1.1). Toto pojetí vychází z obecnějšího konceptu, tzv. Bloomovy taxonomie výukových cílů [2].



Obr. 1.1 **Millerova pyramida úrovní znalostí a dovedností** umožňuje přehledně zobrazit úrovně vhodné pro písemné testování.
Písemné testování (stejně jako např. testování pomocí počítače) lze použít pro hodnocení znalostí a porozumění. Vyšší úrovně dovedností je třeba zkoušet jinými způsoby.
(Upraveno podle [3]*).*

Jak

Pokud máme jasno v tom, jaké znalosti, dovednosti či postoje chceme zkoušet, můžeme přemýšlet o vhodných **formách zkoušení**. Zhruba je můžeme rozdělit na *ústní* a *písemné*. V tomto textu jako svébytnou skupinu vyčleníme ještě *praktické zkoušení*, které hraje v medicíně velmi významnou roli. Každá z uvedených tří forem má své výhody a nevýhody.

Písemné či **počítačové zkoušení** je velmi vhodné pro hodnocení velkého množství studentů a velkého rozsahu látky. Jeho hlavní síla spočívá v hodnocení znalostí od přehledového zkoušení velkých souborů faktů, přes porozumění jejich souvislostem až po simulované řešení problémů či klinických situací. Principiální výhodou je jeho zpětná přezkoumatelnost a velké možnosti statistické analýzy výsledků. Z metodického hlediska je do značné míry lhostejné, zda se zkoušení provádí na papíře nebo elektronicky na počítači; v dalším textu tedy nebudeme tyto dvě možnosti rozlišovat. O písemných a elektronických formách zkoušení budou detailně pojednávat následující kapitoly.

Ústní zkoušení je vhodnější na zjišťování schopnosti řešit problémy, zejména ty hůře strukturované, a tedy podobné reálným situacím. Nevýhody ústního zkoušení zahrnují především velkou časovou a personální náročnost, problematickou standardizaci, obtížnost srovnání jednotlivých zkušebních komisí a termínů a nemožnost zpětného přezkoumání. Výhodou je naopak možnost eliminovat případné nedorozumění v pochopení zadání otázky či odpovědi na ni vzájemnou komunikací studenta a pedagoga.

Praktické zkoušení budeme v tomto textu považovat za specifickou problematiku, které se budeme věnovat jen okrajově. Praktické zkoušení hraje významnou úlohu zejména v klinické části medicínského studia. Mělo by být zásadní součástí závěrečných zkoušek a později i zkoušek konaných např. v rámci specializačního vzdělávání. Praktická zkouška má často několik součástí, při nichž se hodnotí výkon zkoušeného při určité činnosti. Často ale zkoušený odpovídá i na otázky, takže i praktická zkouška mívá prvky písemného či ústního zkoušení. Více informací o praktickém zkoušení naleznete v příloze.

Konkrétní formu zkoušení zvolíme podle toho, *co* chceme zkoušet, jaký je *rozsah* zkoušené látky, *kolik studentů* je třeba ohodnotit, jaké k tomu máme personální či technické *podmínky* a jak *spravedlivě či přesně* potřebujeme v dané situaci zkoušet. Pět studentů pravděpodobně vyzkoušíme mnohem rychleji ústně než písemně, pokud započteme čas nutný na přípravu kvalitního testu. Naopak někdy můžeme být nuceni použít písemné zkoušení z praktických důvodů (nedostatečné personální zajištění), i když v daném případě nemusí jít o optimální formu.

Obecně se dá říci, že **písemné či počítačové testování je vhodné pouze pro zkoušení znalostí a porozumění**. Vyšší úrovně dovedností je třeba hodnotit principiálně jinými metodami (tedy již zmíněným praktickým zkoušením). Jistě není třeba zdůrazňovat, že znalosti a porozumění jsou nutnou, nikoliv však postačující úrovní pro úspěšného absolventa mnoha částí studia; písemné testování tak nutně nemůže být jedinou metodou hodnocení, v určité fázi na něj musí navazovat jiné přístupy.

Moderní technologie a využití počítačů hrají stále větší roli ve vzdělávání, a v medicínském vzdělávání zvlášť. Počítačová podpora hodnocení studentů se rozvíjí již půl století, prakticky od nástupu optického rozpoznávání papírových dotazníků. Jsou čtyři pádné důvody, proč počítačové testování studentů používat: **efektivnost, průkaznost, spolehlivost a přesnost**. [4], [5]

Tab. 1.1 Srovnání písemného a ústního zkoušení

	Písemné a počítačové zkoušení	Ústní zkoušení
Výhody	▪ Rychlá administrace při velkém počtu studentů ▪ Menší potřebný počet zkoušejících (výhoda při velkém počtu studentů) ▪ Rychlé zpracování výsledků ▪ Objektivita ▪ Archivovatelnost	▪ Snazší příprava ▪ Lze testovat i zapojování poznatků a komplexnější dovednosti ▪ Vhodné pro malé počty studentů ▪ Zkoušený získává okamžitou zpětnou vazbu ▪ Možnost eliminace nedorozumění v zadání
Nevýhody	▪ Větší nároky na přípravnou fázi ▪ Nevhodné pro zkoušení komplexních dovedností a činností ▪ Zkoušený dostává zpětnou vazbu o výkonu až s časovým odstupem po zkušebním aktu (netýká se automaticky vyhodnocovaných počítačových testů, kdy student obdrží zpětnou vazbu okamžitě) ▪ Malá efektivita při malém počtu studentů ▪ Chybí osobní kontakt (v některých případech naopak výhodou)	▪ Nízká míra objektivity, špatná standardizovatelnost ▪ Větší riziko lidské chyby ▪ Obtížná přezkoumatelnost

V dalším textu se budeme věnovat téměř výhradně písemnému testování. Mnoho postupů pro práci s písemnými testy se ale obdobně používá i pro kvalitně vedené ústní zkoušení.

S kým

Předpokladem efektivního a relevantního zkoušení je jeho dobrá organizace. U předmětů s větším rozsahem, významem v rámci studijního programu a s velkým počtem studentů je prakticky nezbytné **vytvořit tým**, který se bude organizaci zkoušení systematicky věnovat, a to včetně zpětné kontroly jeho kvality.

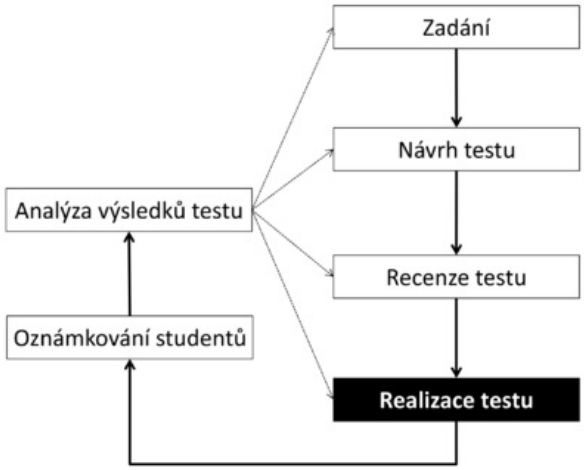
Výhody zkušebního týmu jsou zřejmé: rozloží se nápor práce a umožní se účinná kontrola kvality a vzájemná podpora. Týmová spolupráce je také nezbytná pro standardizaci testů (viz **kapitola 8 Standardizace a normování testu**). Některé zahraniční univerzity do zkušebních týmů zařazují i externí členy z jiných institucí, čímž posilují nestrannost komisí a zajišťují srovnatelnost zkušebních standardů mezi jednotlivými vzdělávacími institucemi.

Tip: Kvalitní testování nezajistí jednotlivec – sestavte aspoň malý tým

Cyklus přípravy testu

Podívejme se, co nás při přípravě, realizaci a vyhodnocení testů může potkat. Předpokládejme, že jsme se rozhodli otestovat znalost skupiny studentů pomocí písemného testu. Příprava písemného testu je náročnější než samotné ústní zkoušení; musíme tedy pro takové rozhodnutí mít nějaký důvod. Může to být potřeba vyzkoušet v omezeném čase velké množství studentů, či potřeba zajistit spolehlivé a reprodukovatelné hodnocení.

Nejjednodušší (tzv. nestandardizovaný) písemný test lze sestavit *ad hoc*, pouze na základě zkušeností vyučujícího. Není na tom nic špatného, pokud je účelem testu pouhé poskytnutí zpětné vazby studentům nebo vyučujícím. Má-li však být výstupem klasifikace nebo rozhodnutí se závažnějšími důsledky (např. rozhodování o postupu studenta do dalšího studia), měla by být přípravě testu věnována patřičná pozornost, aby bylo hodnocení validní, objektivní a reprodukovatelné.



Obr. 2.1 Intuitivní schéma testového cyklu.

Jak by měl vypadat **cyklus přípravy testu**? Jeho základní kroky odhadneme i intuitivně: Máme-li rozmyšlené **zadání** testu, můžeme podle něj test **navrhnout**. Vytvoříme otázky, které si během **recenze** necháme zkontrolovat kolegy. Poté můžeme test **realizovat**, studenty **oznámkovat** a statisticky zhodnotit i samotný test během jeho **analýzy**. Poučíme se, promítneme **zpětnou vazbu** do celého cyklu přípravy a můžeme se pustit do přípravy dalšího testu. Toto intuitivní schéma je na obrázku 2.1.

Projděme nyní jednotlivé kroky podrobněji. Ještě detailněji pak budou postupně popsány v následujících kapitolách. Zvídavý čtenář si může rovněž rozšířit obzory zhlédnutím šestiminutového videa (http://www.ets.org/s/understanding_testing/flash/how_ets_creates_test_questions.html) o cyklu přípravy otázek vytvořeného společností Educational Testing Service (ETS) (http://www.etsglobal.org/Cz/Cze/O-nas/Firma/Kdo-jsme) [6].

Zadání

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

Návrh a příprava testu

Návrh testu je dalším klíčovým bodem celého procesu. Je třeba stanovit, kolik otázek bude test obsahovat z každého tematického okruhu a jaké typy otázek 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 otázek a poměru zastoupení jednotlivých témat v připravovaném testu. Podle anglického pojmenování dříve užívaných modrých kopií stavebních plánů se tomuto **plánování testu** říká *blueprinting*.

Samotná **tvorba testových úloh** patří k odborně i časově náročnějším etapám přípravy testu a je vhodné se na ni teoreticky připravit. V minulosti postupně vznikla celá řada formátů testových úloh, z nichž mnohé byly následně opět opouštěny a skončily na „pohřebišti testových formátů“ (viz Příloha 1). Pozornosti čtenáře doporučujeme formát otázek s jedinou nejlepší odpovědí (*single-best answer*, SBA), který je v současnosti jednou z nejpoužívanějších forem otázek s mnohočetným výběrem odpovědi (*multiple-choice questions*, MCQ).

Při tvorbě testu lze použít i otázky vytvořené dříve. Ty lze schraňovat v tzv. **bance úloh**. Otázky lze pak také sdílet s dalšími skupinami nebo institucemi.

Recenze testu

Má-li být test kvalitní, je nezbytnou součástí jeho přípravy i **oponentura otázek**, při níž se odstraní nahodilé chyby či omyly autorů testu, nejednoznačné či jinak problematické formulace apod. Při oponentuře otázek jsou položky předloženy k posouzení skupině odborníků (např. metodika přípravy testů programu Rogo doporučuje nejméně 5–9 osob), kteří podle připraveného formuláře procházejí testové úlohy a ověřují kvalitu jejich formulace.

Při opakované rutinní tvorbě testů je oponentura součástí samotné tvorby otázek před jejich zařazením do **položkové banky**.

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 a tak dále. Položky, jejichž psychometrické vlastnosti jsou známé, se nazývají *kalibrované*. Protože je pilotní testování organizačně náročné (musíme vždy zajistit skupinu testovaných kvalitativně srovnatelnou s cílovou skupinou, vytvořit jim přiměřenou motivaci atd.), používá se často jako pilotní testování až samotný první běh testu. Známý výsledek pilotního testu převedený do podoby **kalibrovaných položek** je podmínkou pro další efektivní používání nových testových úloh.

Důležitým krokem v této etapě je i nastavení meze, pod kterou nesmí znalost studenta klesnout, aby mohl být považován za úspěšného absolventa kurzu. Tedy např. frekventant kurzu první pomoci nesmí být považován za úspěšného absolventa, pokud nezvládne základní kardiopulmonální resuscitaci. Potřebujeme tedy nastavit jakési absolutní standardy a tento krok se proto nazývá **absolutní standardizace**.

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 férovost testu. S tím souvisí potřeba omezit možnost úniku testových otázek, zajistit identifikaci účastníků testu, dozor během testu a rovné podmínky testu pro všechny účastníky.

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. Ebelovou 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.

Analýza výsledků testu

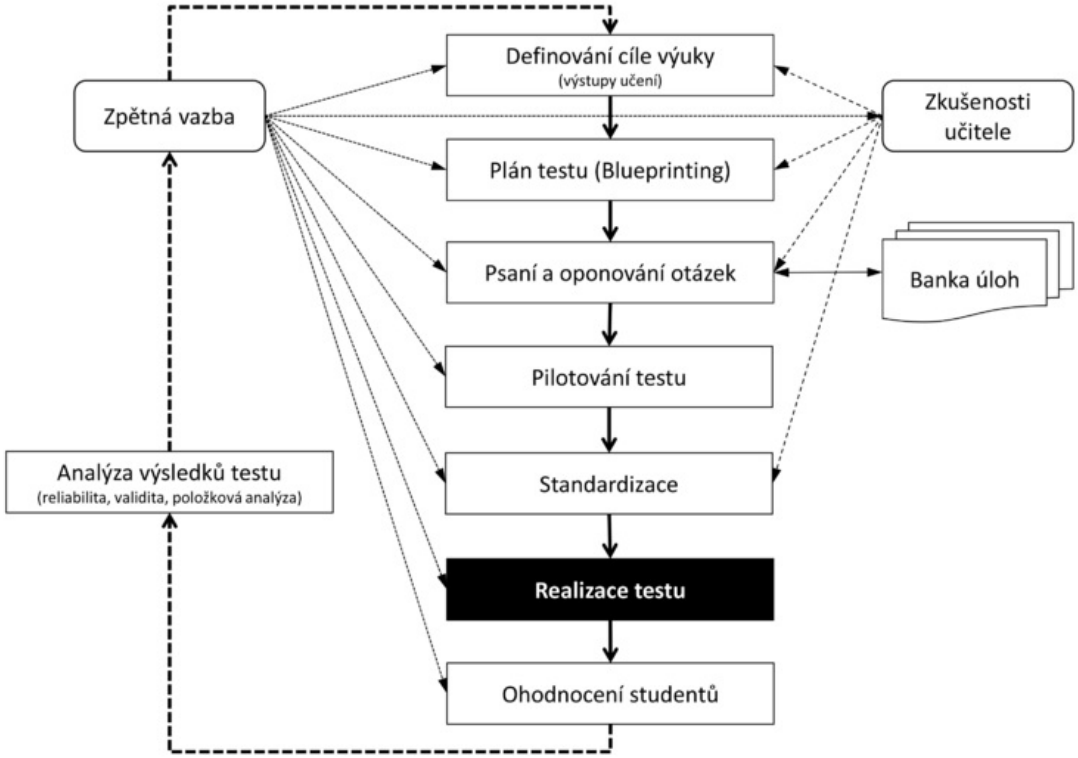
Test je nástroj a jako každý nástroj má konkrétní vlastnosti, které můžeme popsat. Chování testu a jeho položek můžeme hodnotit pomocí *analýzy výsledků testu.* U testu jako celku nás zajímá především jeho spolehlivost (**reliabilita**) a zda měří to, co by měřit měl (**validita**). U jednotlivých položek testu pak **položkovou analýzou** zkoumáme jejich **obtížnost** a **citlivost** s cílem vyloučit nevhodné (nebo prozrazené) položky z dalšího používání.

Optimální je zhodnotit kvalitu testu ještě *před* jeho ostrým nasazením v rámci pilotního testování. Vlastnosti testu je poté potřeba ověřit na cílové skupině testovaných při ostrém nasazení. Při opakovaném použití testu je užitečné porovnávat výsledky v jednotlivých bězích testu.

Zpětná vazba

Výsledky analýzy testu jsou součástí zpětné vazby, díky níž můžeme zlepšit jednotlivé kroky testového cyklu a připravit další, dokonalejší běh testu.

Po tomto myšlenkovém exkurzu můžeme schéma testovacího cyklu nakreslit podrobněji. Dostaneme tak schéma testového cyklu v souladu s doporučením AMEE [7], viz obr. 2.2. Poznamenejme, že začátek testového cyklu vyžaduje **zkušenost učitele**. V dalších krocích, počínaje pilotním testováním, již není účast zkušeného učitele podmínkou a při analýze a hodnocení testu je vhodné spolupracovat se statistikem.



Obr. 2.2 Schéma testového cyklu. (Upraveno podle [7])

V následujících kapitolách se budeme věnovat jednotlivým krokům testového cyklu podrobněji.

Plánování testu (blueprinting)

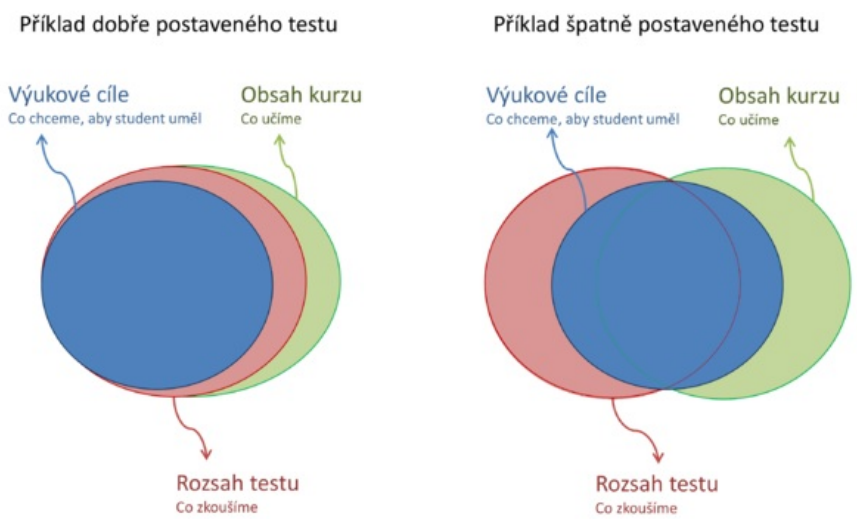
Účelem **plánování testu** (anglicky *blueprinting*) je definovat obsah testu a skladbu použitých metod a forem zkoušení.

V této fázi přípravy testu tedy hledáme odpověď na otázky

- co přesně se bude zkoušet,
- jak velký podíl testu se bude věnovat kterému tématu,
- jakým typem úloh se bude zkoušet určitý typ znalostí či dovedností.

Jinými slovy, při plánování testu se snažíme navrhnout, kolik bude test obsahovat položek ze kterého okruhu a jakého typu tyto položky budou. Zvažujeme přitom význam jednotlivých témat v rámci celého předmětu. Pečlivým naplánováním testu se vyhneme situacím, kdy se po většinu kurzu výuka věnuje určitým problémům, které učitelé považují za zvlášť důležité, v testu se však k těmto tématům objeví jen málo otázek a zbytek se zabývá jinými okruhy. Testovaní jsou pak zaskočeni a výsledky obvykle neodpovídají ani názoru vyučujících. Naopak dobrý plán testování umožní objektivně a srovnatelně klasifikovat i několik skupin studentů pomocí různých variant testů.

Připomeňme na tomto místě, že klíčovou informací pro plánování testu by mělo být, **co má absolvent umět** (tj. „cíle výuky“, *learning objectives*). Tomu by měl odpovídat jak **obsah a rozsah kurzu**, tak i **obsah a rozsah testu**. V praxi se obsah a rozsah znalostí a dovedností vyučovaných, zkoušených a žádoucích vždy poněkud liší. Při plánování testu bychom tedy měli vycházet z cílů výuky, je však třeba přihlížet i ke skutečné náplni kurzu. Výsledky testu se samozřejmě mohou a mají stát zpětnou vazbou pro zdokonalování kurzu v dalších bězích.



Obr. 3.1 Cíle výuky, skutečně vyučovaná látka a obsah a rozsah testu se v praxi vždy poněkud liší. Velkým rozdílům mezi cíli výuky, skutečně probranou látkou a náplní testu má zabránit plánování testu. V dobře naplánovaném testu jsou všechna požadovaná a vyučovaná témata zastoupena proporcionálně, test je vyrovnaný. Nevznikají nežádoucí „překvapení“ vyplývající z toho, že studenti jsou testovaní z něčeho jiného, než očekávali.

Plán testu se nejčastěji vytváří jako dvourozměrná mapa či tabulka [8]. Řádky tabulky odpovídají zpravidla obsahovým celkům, které se mají zkoušet. Sloupce odpovídají kontextu nebo aspektu dané problematiky. Do jednotlivých políček se 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 prázdný nebo téměř prázdný řádek ani sloupec. Pomocí takto sestavené mapy lze snadno dosáhnout toho, že počet otázek věnovaných určitému tématu odpovídá jeho významu a že se adekvátně zkoušejí všechny aspekty určitého problému.

Ukažme si příklad plánování testu z lékařské chemie. Nejprve vytvoříme tabulku, jejíž řádky budou odpovídat tématům podle sylabu předmětu a sloupce běžnému členění výkladu k těmto tématům:

Tab. 3.1 Příklad plánu testu – krok 1

	Struktura	***Fyzikálně-chemické vlastnosti***	***Reaktivita***	***Medicínský význam***
Přechodné prvky				
Nepřechodné prvky				
Anorganické sloučeniny kyslíku				
Anorganické sloučeniny dusíku a síry				
Alifatické uhlovodíky				
Cyklické a heterocyklické sloučeniny				
atd.				

Nyní do jednotlivých políček doplníme počet úloh, které by se měly pro danou kombinaci řádku a sloupce objevit v testu. Snažíme se přitom, aby tyto počty odpovídaly významu látky v rámci předmětu. Není třeba vyplnit všechna políčka, v celé tabulce by ale neměly vznikat větší nevyplněné oblasti.

Tab. 3.2 Příklad plánu testu – krok 2

	Struktura	***Fyzikálně-chemické vlastnosti***	***Reaktivita***	***Medicínský význam***
Přechodné prvky	*-*	*0-1 úloha*	*2 úlohy*	*1 úloha*
Nepřechodné prvky	*-*	*0-1 úloha*	*2 úlohy*	*1 úloha*
Anorganické sloučeniny kyslíku	*2 úlohy*	*1 úloha*	*1 úloha*	*-*
Anorganické sloučeniny dusíku a síry	*2 úlohy*	*1 úloha*	*1 úloha*	*0-1 úloha*
Alifatické uhlovodíky	*2 úlohy*	*1 úloha*	*0-1 úloha*	*2 úlohy*
Cyklické a heterocyklické sloučeniny	*2 úlohy*	*2 úlohy*	*2 úlohy*	*2 úlohy*
atd.				

Nakonec zvolíme vhodné typy testových úloh. V tomto případě bude test složený jednak z úloh, v nichž student volí jedinou nejlepší odpověď (single best answer, SBA), a jednak z úloh s krátkou tvořenou odpovědí (short answer question, SAQ) – v nich můžeme po studentovi vyžadovat např.

nakreslení chemického vzorce nebo napsání chemické rovnice.

Tab. 3.3 Příklad plánu testu – krok 3

	Struktura	***Fyzikálně-chemické vlastnosti***	***Reaktivita***	***Medicínský význam***
Přechodné prvky	-	*0-1 SBA*	*2 SBA*	*1 SAQ*
Nepřechodné prvky	-	*0-1 SBA*	*2 SBA*	*1 SAQ*
Anorganické sloučeniny kyslíku	*1 SBA* *1 SAQ*	*1 SBA*	*1 SAQ*	-
Anorganické sloučeniny dusíku a síry	*1 SBA* *1 SAQ*	*1 SBA*	*1 SAQ*	*0-1 SAQ*
Alifatické uhlovodíky	*2 SAQ*	*1 SBA*	*0-1 SAQ*	*2 SBA*
Cyklické a heterocyklické sloučeniny	*2 SAQ*	*2 SBA*	*1 SBA* *1 SAQ*	*2 SBA*
atd.				

Tip: Terminologické zastavení: Otázka nebo položka?

Typy otázek a jejich vytváření

Otázky s mnohočetným výběrem odpovědí (MCQ)

Nejrozšířenější formou otázek v písemných testech jsou **otázky s mnohočetným výběrem odpovědí** (*multiple choice questions*, **MCQ**). Zkoušený vybírá jednu či více odpovědí z nabídnutého seznamu možností. Velkou předností těchto typů úloh je objektivita hodnocení – testy je možné bodovat i automatizovaně, získaný počet bodů nezávisí na subjektivním posouzení hodnotitelem.

Dosud bylo zkonstruováno a pojmenováno několik desítek podtypů otázek typu MCQ, ale jen několik z nich se rozšířilo. V této kapitole se proto budeme podrobněji věnovat jen zlomku popsaných podtypů formátu MCQ. Některé další typy jsou popsány v Příloze 1.

Poznamenejme, že terminologie otázek s mnohočetným výběrem odpovědí není jednotná. V této publikaci se důsledně držíme názvosloví popsaného v následujícím textu. Termín *multiple choice questions*, *MCQ*, však bývá v jiných textech často používán pro podskupinu, kterou zde vymezujeme úžeji jako *multiple true/false questions*, *MTF*.

.

Otázky s mnohočetným výběrem odpovědi typu „ano/ne“ (MTF)

V literatuře se tyto položky dnes nejčastěji označují jako „multiple true/false“ (**MTF**; podle amerického National Board of Medical Examiners **typ X**). Jedná se o vůbec nejpoužívanější testový formát, a jak ukážeme dále, jde o formát **nadužívaný**. Položka začíná „kmenem“ se zadáním otázky a následuje několik možností (obvykle čtyři až šest). Testovaný má označit jednu či více správných odpovědí.

Katalyzátor

1. *způsobí, že reakce probíhá jiným mechanismem*
2. *způsobí, že v průběhu reakce vznikají jiné meziprodukty*
3. *stejnou měrou urychlí dopřednou i vratnou reakci*
4. *zvýší rovnovážnou konstantu reakce*

V praxi se setkáme s řadou variant těchto položek: liší se minimálním a maximálním počtem správných odpovědí (žádná správná až všechny správné; právě jedna správná; nejméně jedna správná) a různými způsoby bodování. Celá řada studií se věnovala otázce, jaký způsob hodnocení tohoto typu úloh dává nejlepší výsledky; víceméně můžeme konstatovat, že rozdíly nejsou nikterak přesvědčivé (např. [9] [10]).

V každém případě platí, že každou položku MTF můžeme rozepsat do několika otázek s odpovědí ano/ne.

Výše uvedenou položku můžeme přepsat jako sadu čtyř samostatných tvrzení:

- *Katalyzátor způsobí, že reakce probíhá jiným mechanismem ANO/NE*
- *Katalyzátor způsobí, že v průběhu reakce vznikají jiné meziprodukty ANO/NE*
- *Katalyzátor stejnou měrou urychlí dopřednou i vratnou reakci ANO/NE*
- *Katalyzátor zvýší rovnovážnou konstantu reakce ANO/NE*

Odlišnost bývá (nikoli však nutně) v různém způsobu hodnocení: většina způsobů hodnocení MTF shrnuje odpovědi na dílčí otázky do jednoho souhrnného výsledku, což bývá přísnější, než hodnotí-li se každá z nabídnutých odpovědí samostatně. Jak jsme ale už uvedli výše, rozdíl nebývá v praxi významný.

V posledních dvaceti letech se tento typ položek dočkal značné kritiky [11] [12]. V zásadě se dá říci, že je poměrně obtížné sestavit větší počet těchto položek tak, aby byly jednoznačné a současně nebyly příliš snadné.

V praxi se často stává, že odpověď považovaná za „správnou“ za určitých podmínek neplatí. Odpovídá-li pak na stejnou otázku několik recenzentů, často zvolí různá řešení – podle toho, zda jako správnou označí striktně jen možnost, která platí bez výjimky vždy, nebo např. možnost, která platí ve více než 95 % případů, více než 60 % případů apod. Jinými slovy, individuálně se liší vnímání toho, s jakou pravděpodobností musí nabídnutá odpověď platit, aby byla označena jako správná. Opravdu jednoznačně zadané otázky zase mnohdy bývají pro vysokoškolské testování příliš snadné.

Cystická fibróza

1. *je onemocnění s incidencí asi 1:2000*
2. *je letální zpravidla před dvacátým rokem věku*
3. *u mužů způsobuje neplodnost*
4. *je autozomálně recesivně dědičná*

*Tato otázka je problematická, neboť na možnosti 1, 2 a 3 nelze jednoznačně odpovědět ani „ano“, ani „ne“; pouze 4. možnost je jednoznačná. Pokud bychom tuto otázku předložili skupině expertů, jejich odpovědi by se lišily. Incidence cystické fibrózy není 1:2000 ve všech etnických skupinách; recenzenti by pravděpodobně požadovali doplnění otázky. Problematická je také formulace „incidence je **asi** 1:2000“. Podobně sporné jsou i odpovědi na 2. a 3. nabízenou možnost.*

Defekt septa síní bývá u dětí spojen s

1. *systolickým šelestem*
2. *plicní hypertenzí*
3. *Fallotovou tetralogií*
4. *cyanózou*

I tato, na první pohled „nezávadná“ otázka, je sporná. Odpovědi se totiž liší podle závažnosti vady, věku pacienta a podle toho, zda onemocnění bylo či nebylo léčeno. Při tomto zadání si testovaný vytvoří určitý předpoklad a z něj při volbě odpovědí vychází. Položíme-li otázku několika zkoušeným (popřípadě expertům – recenzentům), bude každý vycházet z jiných předpokladů a odpoví jinak. Výsledek testu obsahujícího MTF tedy neodráží pouze znalost testované látky, ale i těžko odhadnutelné psychické pochody související s interpretací otázky a nabízených možností. (Upraveno podle [13])

Další nevýhodou MTF položek je, že často testují pouze vybavení izolovaného údaje.

Tip: Otázky s mnohočetným výběrem odpovědi typu „ano/ne“ (MTF) přinášejí řadu těžko odhadnutelných problémů. Je-li to možné, nepoužívejte je.

Otázky s jedinou nejlepší odpovědí (SBA)

Otázky s jedinou nejlepší odpovědí (*single best answer*, **SBA**; podle amerického National Board of Medical Examiners **typ A**) na první pohled připomínají výběrové otázky MTF s jednou správnou odpovědí – tedy formát široké veřejnosti dobře známý např. z testů pro získání řidičského průkazu. Liší se od nich zdánlivou drobností: z nabízených možností může být i několik správných; jedna možnost je ale výrazně lepší, než všechny ostatní, a testovaný má právě tuto *jedinou nejlepší odpověď* označit. Další důležitý rozdíl spočívá v tom, jakou hloubku znalostí testujeme. U otázek MTF se často jedná o jednoduché spojení dvou termínů (např. porfyrie – hem, viz níže uvedený příklad), zatímco u otázek SBA je k úspěchu potřeba zapojit znalostí více.

32letý muž přichází pro 4 dny trvající, postupně progredující slabost končetin. Dosud byl zdráv, před 10 dny však prodělal infekci horních cest dýchacích. Je afebrilní, arteriální tlak má 130/80 mmHg, tepovou frekvenci 94/min. Dýchání je mělké s frekvencí 42/min. V orientačním neurologickém nálezu dominuje symetrická slabost mimických svalů a svalů horních i dolních končetin. Čití je intaktní. Hluboké šlachové reflexy nelze vybavit. Zánikové jevy jsou negativní.

Která z následujících diagnóz je nejpravděpodobnější?

1. *akutní diseminovaná encefalomyelitida*
2. *syndrom Guillain-Barré*
3. *myasthenia gravis*
4. *poliomyelitis*
5. *polymyositis*

Povšimněte si, že žádná z navrhovaných diagnóz není v daném případě zcela vyloučena. Přesto je varianta č. 2 výrazně lepší odpovědí na otázku, než všechny ostatní. Pokud tuto otázku předložíme skupině expertů, na odpovědi 2 se všichni shodnou. (Upraveno podle [13])

Otázky s jedinou nejlepší odpovědí se dnes považují za jeden z nejlepších nástrojů pro písemné testování v medicíně. Ukazuje se, že jsou **časově nejefektivnější** ze všech používaných formátů [14]. Vyhodnocování odpovědí je snadné a dobře automatizovatelné. Pro zpracování výsledného hodnocení testu není třeba používat žádných složitějších metod (např. vážení odpovědí, korekce na hádání odpovědí apod.)[15].

Skutečnost, že v tomto testovém formátu nemusejí být správné odpovědi ani distraktory („nesprávné“ možnosti) platné či naopak neplatné bez výjimky ve všech případech, odpovídá reálnému životu. Otázky s jedinou nejlepší odpovědí umožňují komplexnější testování znalostí a porozumění než mnohé jiné formáty. To je jejich velká přednost ve srovnání např. s MTF, které testují spíš izolované znalosti [16].

Ukažme si rozdíl mezi „klasicky“ konstruovanou výběrovou otázkou s jednou správnou odpovědí (tedy v našem textu spadající do kategorie otázek MTF) a otázkou s jedinou nejlepší odpovědí (SBA): sestavíme oba typy otázek tak, aby se „ptaly na totéž“.

Příklad otázky ve formátu MTF s jednou správnou odpovědí

Akutní intermitentní porfyrie je důsledkem defektu biosyntézy

1. *kolagenu*
2. *kortikosteroidů*
3. *mastných kyselin*
4. *glukózy*
5. *hemu*
6. *tyroxinu*

Na otázku odpoví správně studenti, kteří si pamatují, že akutní intermitentní porfyrie nějak souvisí s metabolismem hemu. Tuto izolovanou znalost (tedy spojení termínů porfyrie a hem) bude mít i mnoho slabších studentů; distraktory (tj. nesprávné možnosti) jsou navíc poměrně snadno odhalitelné.

Samotná znalost izolované informace ovšem příliš nepomůže při diagnostice onemocnění v praxi. Student potřebuje mnohem více dalších znalostí a dovedností, ty však jednoduché výběrové otázky tohoto typu nemohou postihnout.

Příklad tématicky stejné otázky ve formátu s jedinou nejlepší odpovědí (SBA)

Dosud zdravý 33letý muž přichází pro epizody slabosti a silných křečovitých bolestí břicha. Obtíže se začaly objevovat asi před půl rokem. Podobné epizody se vyskytovaly i v širším příbuzenstvu (teta a bratranec). Během epizody bolestí má vzedmuté břicho, peristaltika je obleněná. V neurologickém nálezu dominuje slabost svalů obou paží.
Příčinou obtíží je pravděpodobně defekt některé biosyntetické dráhy. Které?

1. *pro kortikosteroidy*
2. *pro mastné kyseliny*
3. *pro glukózu*
4. *pro hem*
5. *pro tyroxin*

Povšimněme si, že otázka ve formátu SBA má stejné nabídnuté odpovědi, ale kmen otázky je nyní v podobě klinického medailónku (kazuistiky), který uvádí nejvýznamnější anamnestická data a klinické nálezy. Povšimněme si rovněž, že termín akutní intermitentní porfyrie není vůbec zmíněn. Zkoušený musí sám k této diagnóze na základě klinického medailónku dospět; současně by měl vědět, že toto onemocnění je způsobené poruchou syntézy hemu.

Otázka ve tvaru SBA vede studenta k zapojení komplexnějších schopností než jednoduché výběrové otázky, které testují pouze izolované znalosti. (Upraveno podle [13])

Jakkoli lze otázky s jedinou nejlepší odpovědí považovat za jeden z nejmodernějších testových formátů, nejsou v medicíně ničím novým. Americký National Board of Medical Examiners označuje tyto otázky podle pořadí jako „typ A“ a používá je od r. 1951.

Tip: Většina testů by měla být postavena převážně na otázkách s jedinou nejlepší odpovědí.

Podrobnější informace naleznete na stránce Vytváření otázek s jedinou nejlepší odpovědí.

Rozšířené přiřazovací otázky (EMQ)

Rozšířené přiřazovací otázky (*extended-matching questions*, **EMQ**; podle amerického National Board of Medical Examiners **typ R**) patří k nově zaváděným testovým formátům. Společně s otázkami s jedinou nejlepší odpovědí (SBA) se dnes EMQ doporučují jako jeden z hlavních prostředků pro písemné testy. Původně se používaly pro testování klinických znalostí a dovedností, dnes však nacházejí stále větší uplatnění i v teoretických předmětech.

Kmen rozšířených přiřazovacích otázek bývá poměrně dlouhý – zpravidla jej tvoří **klinický medailónek** nebo **scénář**. Testovaný vybírá jedinou nejlepší odpověď z většího množství nabízených možností. Otázky, které se týkají stejné oblasti, zpravidla využívají jednu společnou sadu možných odpovědí.

- *Okruh: Bolesti břicha – diagnostika*
- *Otázka:* ***Vyberte nejpravděpodobnější diagnózu pro každý z níže uvedených případů***
- *Nabídnuté odpovědi:*

```
Možnosti:
A. Aneurysma abdominální aorty          K. Nefrolitiáza
B. Appendicitis                         L. Mesenteriální lymfadenitida
C. Obstrukční ileus                     M. Trombóza mezenteriální arterie
D. Cholecystitida                       N. Ruptura ovariální cysty
E. Karcinom tlustého střeva             O. Pankreatitida
F. Zácpa                                P. Pelvic inflammatory disease
G. Divertikulitida                      Q. Peptický vřed
H. Mimoděložní těhotenství – ruptura    R. Perforovaný peptický vřed
I. Endometrióza                         S. Pyelonefritida
J. Kýla                                 T. Torze ovaria
```

- *Kmen otázky: 25letá žena přichází pro náhle vzniklou bolest v pravém dolním břišním kvadrantu, která se postupně zhoršuje. Pociťuje nauseu, ale nezvrací. Před začátkem bolesti byla zdráva. Při vyšetření je v pravém dolním kvadrantu hmatná hluboko uložená rezistence, která je palpačně citlivá. Peristaltika je zachována. Per vaginam je vpravo hmatná tuhá rezistence o průměru asi 7 cm. Hematokrit je 32 %, počet leukocytů v periferní krvi je 18 tisíc / mm^3. Aktivita amylázy v séru je v referenčních mezích, zkouška na okultní krvácení do stolice je negativní.*

Odpověď:

Položka může mít více kmenů otázky (scénářů) ke stejné sadě nabídnutých odpovědí.

🔎 *Podrobnější informace naleznete na stránce Fórum:Testy/EMQ.*

Otázky typu „vyber N“

Otázky typu „vyber N odpovědí“ (anglicky *pick N*) jsou podobné otázkám s jedinou nejlepší odpovědí (SBA) nebo rozšířeným přiřazovacím otázkám (EMQ). Liší se v tom, že zkoušený má z nabídnutého seznamu možných odpovědí vybrat dvě až pět možností.

U 50letého muže se postupně začíná objevovat zmatenost, dezorientace a porucha krátkodobé paměti. Má parézu levé dolní končetiny. V krevním nátěru je mikrocytóza a bazofilní tečkování erytrocytů. Vyberte dvě nejpravděpodobnější příčiny:

```
A. Diabetická polyneuropatie          F. Roztroušená skleróza
B. Huntingtonova choroba              G. Parkinsonova choroba
C. Laterální míšní syndrom            H. Gliom pontu
D. Encefalopatie při otravě olovem    I. Tabes dorsalis
E. Meduloblastom                      J. Wernickeova encefalopatie
```

Pokud by v tomto případě měl zkoušený vybrat jedinou nejlepší odpověď, byla by otázka sporná: nejpravděpodobnější je diagnóza B, avšak dosti pravděpodobná je i možnost H. Další navrhované možnosti jsou výrazně horší odpovědí. Má-li zkoušený zvolit dvě možnosti, je řešení poměrně jednoznačné.

Často se tento typ používá při zkoušení volby postupu v určité situaci, kdy se např. ordinuje několik vyšetření současně, nebo se provádí několik léčebných opatření najednou či téměř najednou.

V zadání je třeba definovat očekávaný počet odpovědí. Tím se z otázky typu mnohočetného výběru *ano/ne* stane otázka typu výběru nejlepší odpovědi.

Úlohy s otevřenou odpovědí

Testovaná osoba odpovídá volně psaným textem. Pro hodnocení úloh s otevřenou odpovědí je obtížné použít automatizované postupy – odpovědi zpravidla musí být čteny a bodovány vyškoleným hodnotitelem nebo zkoušejícím. Na druhou stranu tento typ úloh umožňuje komplexnější posouzení testovaných schopností.

Otázky s krátkou tvořenou odpovědí (SAQ)

Otázky s krátkou tvořenou odpovědí (*short-answer questions*, SAQ) patří v praxi mezi nejpoužívanější formát úloh, ať už při písemném nebo ústním zkoušení. Používají se i v průběhu seminářů, stáží a přednášek. Tvoří je otázka, na kterou **lze odpovědět jedním nebo několika málo slovy**. Na otázku často existuje **několik správných odpovědí**.

- *Jaký je poslechový nález u tohoto pacienta?*
- *Kterou zkouškou prokážete v moči přítomnost bílkoviny?*

Variantami tohoto typu úloh je například doplnění popisků u anatomického schématu, doplnění chemické rovnice, matematický výpočet, doplnění chybějících slov v jazykovém testu a mnoho dalších.

Vlastnosti otázek s krátkou tvořenou odpovědí, pokud jde o jejich použití v testech a spolehlivost hodnocení, jsou někde **mezi výběrovými otázkami a eseji** (viz dále) [14][17]. **Hodnocení** otázek s krátkou odpovědí je **obtížnější než u výběrových otázek, ale spolehlivější než u eseje**. Ve srovnání s výběrovými otázkami je jistou výhodou otevřenost odpovědi (testovaný nemůže odpověď uhádnout), v zásadě ale tento typ otázek **testuje stejné úrovně znalostí a dovedností, jako výběrové otázky** [17]. Stejně jako u výběrových otázek zde platí, že otázky s krátkou odpovědí **nejsou vhodné pro testování komplexnějších znalostí**, hlubšího porozumění problému a pochopení složitějších vztahů. Pokud ovšem studenti vědí, že budou testováni otázkami s krátkou odpovědí (a nikoliv výběrovými otázkami), lépe se naučí faktografické údaje [17].

🔎 *Podrobnější informace naleznete na stránce Vytváření otázek s krátkou odpovědí pro písemné testy a jejich hodnocení.*

Esej

Zkoušení psaním eseje má mnoho vlastností podobných ústním zkouškám. Hodnocení výkonu je do značné míry subjektivní; písemná forma ovšem umožňuje zachovat záznam o zkoušce a v případě potřeby se k němu vrátit, popřípadě opakovat bodování jiným hodnotitelem. Zadáním je obvykle otázka nebo (častěji) soubor otázek, které mají být v eseji pojednané, případně jde o široce zadané téma. V některých případech může testovaná osoba používat určené, nebo i libovolné informační zdroje. Pro psaní eseje je vždy stanoven čas a může být vymezen i rozsah odpovědi.

🔎 *Podrobnější informace naleznete na stránce Fórum:Testy/Esej/Podrobnosti.*

Modifikovaný esej (MEQ)

Modifikovaný esej (*modified assay question, MEQ*) se ve světě používá např. pro zkoušení při postgraduálním vzdělávání v oblasti praktického lékařství. Pomocí tohoto typu úloh lze posoudit analytické uvažování, interpretaci nálezů a klinické rozhodování. V pregraduálním vzdělávání se naopak tento typ úloh příliš neosvědčuje[14].

Otázka začíná krátkým klinickým scénářem nebo charakteristikou. Následují otázky, na které uchazeč odpovídá několika větami.

38letá pacientka, povoláním učitelka na základní škole, byla přijata na všeobecné interní oddělení pro únavu a tachykardii k dalšímu vyšetření.

- *Jaké jsou tři nejpravděpodobnější příčiny jejích obtíží?*
- *Napište pět otázek, které položíte pacientce a které vám pomohou mezi zvažovanými příčinami rozhodnout.*

🔎 *Podrobnější informace naleznete na stránce Fórum:Testy/MEQ/podrobnosti.*

Hlavní doporučení pro volbu typu otázek

Každý test musí být sestaven podle potřeb konkrétního kurzu; při konstrukci je vždy vhodné použít postupy popsané v kapitole 3 věnované plánování testu. Pokud jde o volbu typu položek testu, lze vycházet z následujících doporučení:

- V kontextu vysoké školy je většinou nejvýhodnější založit test na otázkách s jedinou nejlepší odpovědí (SBA).
- Testy pro menší skupiny studentů je možné doplnit o otevřené otázky. Efektivní je především použití otázek s krátkou tvořenou odpovědí (SAQ). V testech pro malé skupiny studentů, u kterých se nepředpokládá opakované použití, může být tento typ otázek nejvýhodnější.
- Další formáty jsou vhodné jako doplněk nebo ve specifických situacích:
  - V testech orientovaných na klinické uvažování je možné použít rozšířené přiřazovací otázky (EMQ), popřípadě otázky s výběrem několika odpovědí (*pick N*).

- Pro zkoušení hlubšího porozumění problematice může být výhodné použít modifikovanou esej (MEQ) nebo i eseje. Je však třeba počítat se značnými časovými i odbornými nároky na hodnocení testů.

V každém případě je třeba připomenout, že písemné testování je vhodné pro zkoušení **znalostí** a **porozumění** (viz kapitola **1 Dříve než začnete**). Komplexnější dovednosti a činnosti je třeba zkoušet prakticky. Písemné testování může být vhodným předstupněm, který ověří, zda student vůbec má základní předpoklady k tomu, aby praktickou zkoušku, která je časově i organizačně náročnější, splnil.

Oponentura otázek

Nedílnou součástí přípravy testů je **oponentura**. Je rozdělena do několika fází, které se vždy zaměřují na specifickou oblast. Jejím cílem je odhalení nedostatků, které zpravidla testy ve své počáteční podobě obsahují. Provedení oponentury je totožné pro všechny formy zkoušení. Stejný postup tedy lze použít při elektronických i písemných variantách testování. Motivací k provádění revize je objektivní zajištění správnosti, optimalizace testu a odstranění subjektivních vlivů. I když bývá oponentura 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. Teprve potom přichází na řadu pilotní administrace testu na malé skupině studentů.

Oponentura, podobně jako příprava kompletní testové agendy, je založena na týmové spolupráci. Několik zainteresovaných odborníků nezávisle na sobě posuzuje vhodnost jednotlivých otázek a společnými silami se snaží o odstranění všech nedostatků, které by mohly při praktické realizaci vadit. **Týmová spolupráce** tedy **hraje** při oponování databanky úloh zcela **klíčovou roli**. Subjektivní pohled autora dané otázky může mnohdy negativně ovlivnit formulaci zadání, případně také nabízených odpovědí, a tedy i správnost celé otázky. Konstruktivní oponentura několika vyučujících zvyšuje především objektivitu otázek a významně tak přispívá ke zkvalitnění celého testu.

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

Obsahová revize

```
Jsou odpovědi správně a přesně formulované? Nejsou distraktory diskutabilní?
```

Položky testu je třeba nejprve zkontrolovat po obsahové stránce. Tuto obsahovou revizi by měli provádět nejlépe ostatní spoluautoři, popřípadě pedagogové ze stejného oboru. Musíme zdůraznit, že u každého z typů otázek (s otevřenou odpovědí a s výběrem odpovědi) se provádí obsahová revize jiným způsobem. Cílem je vždy nalezení nesprávných nebo nepřesných formulací v zadání otázek; u položek s výběrem odpovědí se obdobně hledají nepřesnosti či chyby i v jednotlivých nabízených odpovědích.

V dosud nejrozšířenějších otázkách s mnohočetným výběrem odpovědí typu MTF se často chybuje při definici nesprávných odpovědí (distraktorů). Obecně se doporučuje kontrolovat zejména

- přesnost formulace zadání/kmene otázky,
- zda jsou jednotlivé distraktory v každé otázce formulovány tak, aby za žádných okolností, v žádné interpretaci ani v žádném uvažovaném případě nemohl být distraktor správnou odpovědí (platí právě pro nejvíce používanou verzi MTF),
- zda položky v testu odpovídají plánu testu (blueprint).

Zopakujme na tomto místě ještě jednou ukázku položky s problematicky zvolenými distraktory:

Cystická fibróza

1. *je onemocnění s incidencí asi 1:2000*
2. *je letální zpravidla před dvacátým rokem věku*
3. *u mužů způsobuje neplodnost*
4. *je autozomálně recesivně dědičná*

*Jde o typický příklad nejednoznačné otázky (některé z distraktorů mohou být v určitém případě považovány na správnou odpověď). Na možnosti 1, 2 a 3 nelze jednoznačně odpovědět ani „ano“, ani „ne“; pouze 4. možnost je jednoznačná. Pokud bychom tuto otázku předložili skupině expertů, jejich odpovědi by se lišily. Incidence cystické fibrózy není 1:2000 ve všech etnických skupinách; recenzenti by pravděpodobně požadovali doplnění otázky. Problematická je také formulace „incidence je **asi** 1:2000“. Podobně sporné jsou i nabízené odpovědi 2. a 3.*

V rámci revize obsahu je velmi vhodné, aby zadání otázek a nabízené odpovědi zkontrolovali jak spoluautoři celého testu, tak 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. Formulace alternativních odpovědí (distraktorů) je pro většinu pedagogů jedna z nejobtížnějších činností při vytváření databanky položek. Je velmi

náročné vhodně definovat nesprávné odpovědi tak, aby nebyly příliš zavádějící ani přehnaně snadné. 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.

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.

Revize férovosti

```
Testují otázky pouze požadovanou konkrétní znalost a nic jiného?
```

Férovost bývá často spojována s rovností podmínek při řešení testu. Jak napovídá obrázek 5.1, pouhé zajištění rovných podmínek nemusí být dostačující. Jak tedy férovost testu a jeho položek posuzovat?

Obr. 5.1 Rovnost podmínek (zdroj http://teacherstraining.com.au/friday-funny-standardised-testing/)

Každá položka by měla testovat právě požadovanou vědomost, znalost či schopnost a nic jiného. Pokud jsou k zodpovězení otázky nutné znalosti a dovednosti, které z jakéhokoli důvodu nebyly srovnatelně dostupné všem testovaným osobám, tedy pokud všichni testovaní neměli totožnou 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 a především pochopitelné. Ačkoli chtěl autor otázky ověřit určitou znalost, současně v tomto případě nechtěně testuje jazykovou vybavenost a přehled v odborné terminologii. V této souvislosti může být další komplikací také testování pozornosti studentů prostřednictvím "chytáků v zadání", případně používání podobných zkratek nebo pojmů.

Položka by neměla zvýhodňovat žádnou skupinu podle věku, pohlaví, původu, společenského a ekonomického postavení, víry, rasy, mateřského jazyka, atd. Studenti z různých skupin *se shodnou úrovní znalosti* by měli na danou otázku odpovídat správně se stejnou pravděpodobností.

> *V testech SAT ve Spojených státech se vyskytla otázka, v níž měli dotazovaní najít dvojici pojmů, mezi nimiž je vztah shodný jako mezi pojmy „běžec“ a „maratón“. Správná odpověď byla „veslař“ a „regata“. Při analýze testů se ukázalo, že na tuto otázku odpovídali prokazatelně hůře černí studenti (22 % správných odpovědí), než jejich bílí kolegové (53 % správných odpovědí), ačkoli v jiných otázkách tomu tak nebylo. Otázka předpokládala „samozřejmou“ znalost sportu bohatých, která však nebyla v různých etnických skupinách populace distribuována rovnoměrně.* [18]

Základní doporučení a pravidla tvorby testových položek týkající se férovosti položek jsou obsažena v manuálu ETS Guideline for Fairness Review [19].

Neférovost položek lze často odhalit důkladnou revizí samotného zadání. Někdy ji však neodhalí ani zkušený autor. Lze ji také rozpoznat při analýze výsledků testu, kde se k této problematice vrátíme.

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 otázky a ověřit, zda jsou dostatečně čitelné, srozumitelné a formálně i typograficky jednotné. Složitá větná souvětí, dvojité zápory a krkolomná zadání otázek je vhodné přepracovat do jednodušší formy tak, aby student nemohl ve formulaci zabloudit. Zadání otázky i samotné odpovědi 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 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.

> *Při redakční revizi můžeme odhalit i gramaticky nebo graficky návodné formulace otázek (tzv. sugestivní zadání):*
>
> *Místem narození Jana Amose Komenského byl:*
>
> 1. *Uherský Brod*
> 2. *Nivnice*
> 3. *Komňa*

4. Brno

Formulář pro recenzenty úloh

Z praktického hlediska je výhodné vybavit recenzenty formulářem, který je oponenturou testové položky „provede“. Tím, že recenzent odpovídá na jednotlivé otázky ve formuláři, musí se testovou položkou zabývat z různých pohledů. Není třeba striktně vyžadovat, aby každá testová položka zcela vyhověla ve všech sledovaných parametrech; oponent by však případné odchylky měl v každém konkrétním případě považovat za opodstatněné (což zpravidla napíše v komentáři). Příklad formuláře pro recenzenty úloh najdete níže, v přílohách je k dispozici k tisku a k editaci .

Tab. 5.1 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 neobsahuje chytáky (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		

Pilotování testu

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. 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 rozlišit mezi dobrými a slabými účastníky testu a které umožňuje získat subjektivní i objektivní zpětnou vazbu od testované skupiny. Pretestování je srovnatelné s kroky, které se provádějí při vyvozování závěrů z „ostrého“ testování. Zatímco pro samotný pilotní běh testu stačí menší skupina studentů (například 20 [20]) 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 [21].

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 testu. Názory každého člena zvolené skupiny je nutné brát v úvahu a zvážit jejich 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 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
- diskuze ve frontální výuce (v případě menšího množství studentů, při větším počtu se tato varianta stává neefektivní)
- poznámky v testu nebo tzv. přemýšlení nahlas (tzv. „think aloud“, viz [22]), kdy jsou studenti žádáni, aby při řešení testu komentovali nebo zaznamenávali své myšlenkové pochody

Objektivní zpětná vazba

Objektivní zpětná vazba je důležitá pro svou nepopiratelnost, která vychází z matematického zpracování výsledků testu. Závěry objektivní zpětné vazby jsou podložené a dávají jasné indicie k případné modifikaci nevyhovujících testových položek. Mezi nejznámější a nejhojněji užívané metody patří:

- zhodnocení **obtížnosti** testových úloh (identifikace snadných a obtížných úloh, nevyhovujících otázek, 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.

Standardizace a normování testu

Standardizace testu znamená zajištění rovnosti podmínek testovaných, porovnatelnosti jejich výsledků navzájem a porovnatelnosti výsledků z různých běhů testu tak, aby zkoušení bylo spravedlivé, objektivní a reprodukovatelné. Standardizované testy nabízejí všem respondentům stejný test za stejných (nebo přiměřeně rovných) podmínek, a jsou proto vnímány jako spravedlivější než jiná hodnocení, která používají nesrovnatelné otázky a podmínky pro studenty skládající zkoušky v různých termínech nebo u různých examinátorů.

Přínosy standardizace

Jednou z hlavních předností standardizovaného testování je, že výsledky mohou být objektivně dokumentovány a mají dostatečný stupeň spolehlivosti (reliability) a správnosti (validity). Výsledky standardizovaného testování jsou zobecnitelné a opakovatelné, což je odlišuje od školního hodnocení, které je závislé na jednotlivém učiteli. Bez standardizovaného testování by bylo obtížné objektivizovat rozdíly ve vzdělávání jednotlivých škol či učitelů.

Dobře navržený standardizovaný test 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

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 učení“ [23]. Kritici standardizovaných testů poukazují na uniformitu takového vzdělávacího modelu a produkování absolventů „jako na montážní lince“ [24]. Další námitkou je, že nadužívání a zneužívání standardizovaných testů poškozuje výuku, neboť zužuje osnovy. Použití standardizovaného testování bez ohledu na cíle výuky totiž způsobuje, že co není testováno, se neučí; způsob zkoušení se pak stává vzorem toho, jak předmět učit. Do jaké míry jsou tyto výhrady relevantní pro testování v rámci studia na lékařských fakultách, je ovšem zatím nezodpovězená otázka. Příznivci standardizovaného testování reagují, že nejde o kritiku standardizovaného testování jako takového, ale o kritiku špatně navržených testů.

Rozsah standardizace

Standardizace může být chápána v několika různě širokých pojetích [25]:

1. Základní standardizace se týká **rovnosti podmínek**, podoby materiálů a procedur testování (tzv. *standardizace I*). Takto byla chápána např. standardizace podmínek literátských zkoušek ve staré Číně.
2. Standardizace v širším pojetí se zabývá **nastavením mezí** pro úspěšné absolvování testu. Výkon testovaného je porovnáván s výkony populace, pro kterou je test určen [26], nebo s absolutními kritérii, požadavky na minimum znalostí, které si má absolvent kurzu odnést (tzv. *standardizace II*).
3. V nejširším pojetí lze standardizaci chápat jako naplnění všech standardů testování, tedy kromě výše zmíněných také zajištění vysoké **validity** a **reliability** testu, jak o tom pojednává kapitola *8 Analýza výsledků a hodnocení kvality testu* (tzv. *standardizace III*).

Standardizace jako zajištění rovnosti podmínek

Následující medailónek ukazuje, že první typ standardizace (ve smyslu vymezení rovných podmínek) má starověké kořeny.

Historický medailonek – standardizované testování ve staré Číně

*První historické důkazy o standardizovaném testování nacházíme ve starověké Číně. Hannibalův současník Čchin S'-chuang, který vládl malému, zaostalému státu Čchin, dobyl a podmanil si všechna okolní království. Jeho následníci z dynastie Chan, aby udrželi dobytá území pod kontrolou, zrušili v prvním století př. n. l. výsadní postavení aristokracie a státní správu svěřili centralizovanému byrokratickému aparátu, do nějž bylo třeba vybírat kompetentní a loajální úředníky. Uchazeči o státní službu skládali přísné písemné **císařské**, nebo též **literátské zkoušky**. Tento systém výběru úředníků byl pak používán 2000 let.* [27] *Za panování dynastie Suei (581–618) byl systém císařských zkoušek definitivně zformalizován a standardizován a přijat jako jediná metoda výběru kandidátů. Testování bylo písemné, anonymizované a velmi důkladně chráněné.*

Obr. 7.1 Imperiální zkoušky v Číně na soudobé ilustraci (11. století)

Prevence proti nedovoleným pomůckám

Hned u vchodu do střeženého areálu čekala kandidáty důkladná zdvojená osobní prohlídka. Prohlídku vykonávali čtyři vojáci současně. Pokud při druhé prohlídce bylo nalezeno něco podezřelého, byli potrestáni i vojáci, kteří prováděli první prohlídku. Nebyly povoleny žádné osobní předměty vyjma psacích pomůcek. Dohlížitelé dostávali za odhalení nelegálních pomůcek odměnu tří uncí stříbra a podvádějící kandidáti byli vylučováni. Nicméně tendence podvádět byla vzhledem k významu zkoušek silná a bohatí kandidáti si například opatřovali miniaturní přepisy klasiků na součástech oděvu.

Prevence nedovolené pomoci

Po celou dobu zkoušek byli kandidáti uzavřeni v pečlivě hlídaném areálu. Testový arch byl úředně orazítkovaný a kandidáti byli trvale pod dozorem. Aby se snížila možnost opisování, obcházel po první hodině jeden z dohlížitelů kandidáty a časovým razítkem vyznačil, kam do té doby dopsali. Pokud byl tento kontrolovaný úsek příliš krátký, nahlíželo se na test jako na podezřelý, a to i kdyby byl později celý vyplněn. Práce byly rovněž porovnávány navzájem a hledala se nápadná podobnost. Identita uchazečů byla ověřována osobním svědectvím garanta, který zkoušeného znal, a mohla být ověřována i dodatečným porovnáváním písma, pokud vznikla pochybnost. Zkoušený nesměl opustit svoji celu více než jednou, nesměl nechat spadnout svoje papíry na zem, nesměl mluvit, dívat se na ostatní. Testový list měl pro případ drobnějších přestupků prostor na tři „kárná“ razítka. Pokud kandidát tento počet vyčerpal, byl ze zkoušek vyloučen.

Prevence zvýhodnění examinátorem

Neméně důkladná byla opatření proti protěžování hodnotitelem. Nejen, že byly testy pro hodnocení anonymizovány, ale byly dokonce celé přepisovány, aby nemohly být identifikovány podle rukopisu. Opisovači neměli k dispozici barvu inkoustu, kterou používali studenti, aby nemohli do původních prací zasahovat, a tak dále [28]*. Z výčtu tehdejších bezpečnostních opatření lze čerpat inspiraci ještě dnes.*

Za dynastie Song (960–1279) dosáhl zkouškový systém jako způsob výběru státních úředníků svého vrcholu. Systém byl zrušen až při vlně reforem v roce 1905. [29], [30]

Standardizované testování se dostává do Evropy počátkem 19. století prostřednictvím bývalých koloniálních správců, mezi kterými proslul Thomas Taylor Meadows svým (celkem trefným) varováním, že britské impérium přijde o všechny kolonie, pokud nebudou koloniální úředníci vybíráni v nestranné soutěži a nezávisle na původu. [31]

Tradiční západní pedagogika, která vycházela spíše z klasických řeckých kořenů, používala pro posouzení studentů hlavně nestandardizované hodnocení ve formě psaných esejů. K hromadnému nasazení standardizovaného testování došlo v rámci Britského impéria nejdříve v koloniální Indii, kde byli touto cestou vybíráni zaměstnanci, aby se zabránilo korupci a protekci. Na konci 19. století bylo standardizované testování jako metoda přijato i v kontinentální Británii a později i v dalších západních zemích.[32]

Pro podporu reprodukovatelnosti podmínek a postupů testování, skórování a interpretace testových výsledků existuje několik pomůcek a nástrojů

- K zajištění reprodukovatelnosti testů realizovaných více pedagogy, na více školách, či v delším časovém období bývá testovým týmem vytvářen **metodický materiál pro hodnotitele** (označuje se např. jako *pokyny pro hodnotitele*, *příručka k testu*, *pokyny pro zkušební komisi*, *metodické pokyny pro hodnotitele*, *pokyny pro organizaci zkoušky* apod. [33], [34], [35], [36]). Dává pedagogovi přesný návod na přípravu, provedení a vyhodnocení testu, aby byla zajištěna reprodukovatelnost výsledků.
- V případě testů důležitých pro další rozhodování o studiu nebo profesním uplatnění může tým připravující test vydat příručku obsahující instrukce a zadání pro studenty, nazývanou odborně *testový sešit*. Příkladem může být příručka pro přípravu studentů na písemnou práci z českého jazyka v rámci státních maturit [37].

Vyrovnávání obtížnosti testů

Pokud je test administrován opakovaně, například pokud slouží k ověřování úrovně vědomostí potřebné pro výkon nějaké odbornosti nebo povolání, může být vznesen požadavek na zajištění vzájemné porovnatelnosti jednotlivých běhů testu. Součástí standardizace se pak stává **vyrovnávání obtížnosti testů** (též harmonizace testů). Vyrovnávání

obtížnosti (angl. Equating) je statistický proces umožňující přepočítat hodnocení studentů z jednotlivých běhů (forem) testu tak, aby výsledky studentů dosažené v jednom běhu mohly být porovnávány s výsledky studentů v jiných bězích testu [38].

Lze k tomu použít řadu metod založených buď na klasické testové teorii (CTT), nebo na teorii odpovědi na položku (IRT) [39].

Základem části těchto metod je tzv. *kotvení testu*. Do testu se zařadí určitý počet úloh, které jsou ve všech verzích stejné. Tyto tzv. *kotvící položky* pak slouží ke vzájemnému porovnání 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 [40].

Pro vyrovnávání obtížnosti testů a jejich škálování na základě IRT je k dispozici volně dostupný software IRTEQ (http://www.hantest.net/irteq) [41].

Standardizace jako stanovení norem

Samotný výsledek konkrétního testu nemá žádnou vypovídací hodnotu o tom, jak si respondent stojí v porovnání s ostatními. Pouhý počet bodů nám neříká nic o tom, jestli student dosáhl nadprůměrného výsledku, nebo naopak zapadl do beznadějného podprůměru. Získá-li pak student ve dvou různých testech stejný počet bodů, může to znamenat v jednom testu vynikající výkon, zatímco v testu druhém pouze výkon průměrný. Teprve na základě porovnání dosaženého počtu bodů se standardy nebo výkony ostatních jsme schopni jednotlivce adekvátně posoudit. Standardizovanými testy se tedy snažíme o vyjádření výsledků jednotlivých respondentů buď vzhledem k výsledkům reprezentativního vzorku (typicky se jedná o stovky studentů) [42], nebo vzhledem ke kritériím – konkrétním znalostem, které absolvent kurzu musí mít.

Nejjednodušší metody standardizace jsou založeny na určení procenta respondentů, kteří dosáhli v daném testu horšího výsledku než daný student. Tento postup se používá často například při vyhodnocení přijímacího řízení. Ke každému bodovému zisku je přiřazeno *percentilové pořadí*, které zhruba uvádí, kolik procent respondentů dosáhlo výsledku horšího než testovaný uchazeč. Lze tak velmi snadno posoudit relativní pořadí konkrétního jedince v celé skupině respondentů.

Standardizaci testu – ve smyslu objektivizace hodnocení výsledku studenta v testu – lze rozdělit na tři přístupy

- **Relativní standardizace** je založena na analýze rozdělení získaných dat a porovnává výsledky respondentů mezi sebou.
- **Absolutní standardizace** je založena na dosažení konkrétních kritérií, tedy například toho, kolik správně zodpovězených otázek daný respondent vyprodukoval. Příkladem je stanovení hranice 70 % správně zodpovězených otázek pro úspěšné složení testu. [43]
- **Kombinovaná standardizace** je pak kombinací absolutní hranice mezi úspěšným a neúspěšným studentem (tzv. *pass mark*) a relativního rozdělení známek v pásmu úspěšnosti např. podle percentilů, směrodatné odchylky, apod.

Tab. 7.1 Schematický příklad ilustrující hodnocení studentů při relativní a absolutní standardizaci.
Studenti mají zodpovědět otázku: Co bylo příčinou druhé světové války?

Odpovědi studentů	Hodnocení při absolutní standardizaci	Hodnocení při relativní standardizaci
Student 1: Druhá světová válka byla vyvolána vpádem Hitlera do Polska.	Odpověď je správná.	Tato odpověď je horší než odpověď Studenta 2, ale lepší než odpověď Studenta 3.
Student 2: Druhá světová válka byla vyvolána mnoha faktory včetně hospodářské krize, obecné ekonomické situace, růstu nacionalismu a nevyřešených následků první světové války. Válka v Evropě začala německou invazí do Polska.	Odpověď je správná.	Tato odpověď je lepší než odpověď Studenta 1 a Studenta 3.
Student 3: Druhá světová válka byla vyvolána atentátem na Arcivévodu Ferdinanda.	Odpověď je chybná.	Tato odpověď je horší než odpověď Studenta 1 a Studenta 2.

Rozhodnutí, který typ standardizace pro konkrétní test použijeme, souvisí vždy s účelem testu.

Výhody a nevýhody jednotlivých typů standardizace

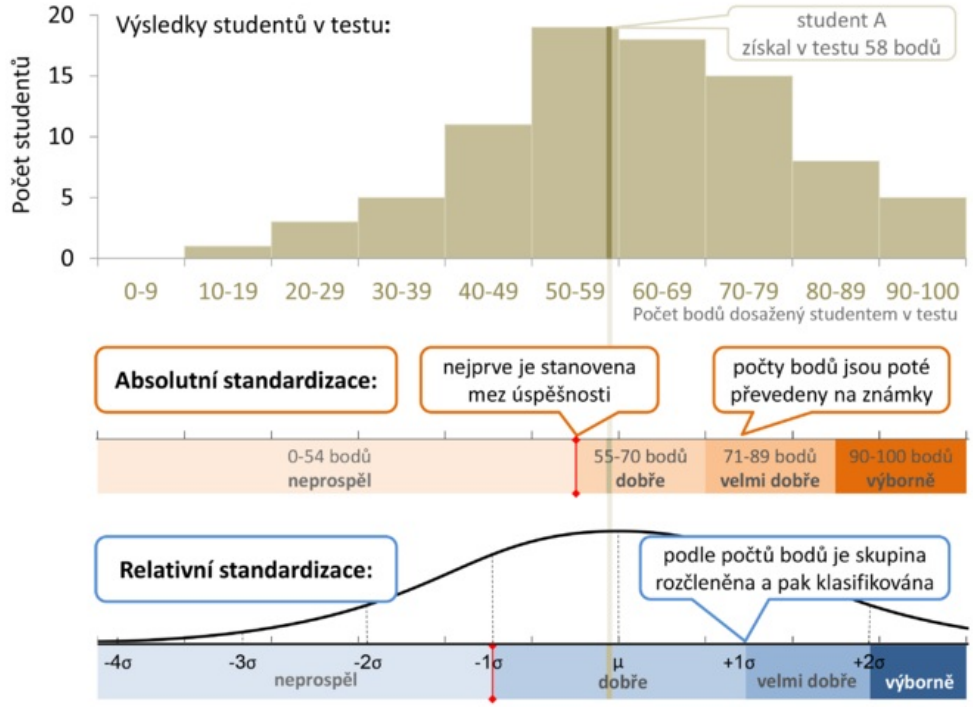
Relativní standardizace se neváže na obsah testu, ale hodnotí jednotlivé účastníky mezi sebou. Výhodou tedy je, že zabraňuje inflaci nejvyšších hodnocení, zřetelně odliší nejlepší studenty a není nutné individuálně standardizovat každý test zvlášť.

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 nemůže v testu uspět, ani kdyby látku uměli sebelépe. 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 oslabením vztahu mezi jejich úsilím a výslednou známkou, protože ta závisí nejen na jejich vlastním výkonu, ale i na výkonu ostatních. 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.

Absolutní hodnocení závisí jen na tom, co se student naučil, nikoli na jeho pozici mezi ostatními. Jeho nevýhodou je nutnost stanovovat kritéria úspěchu pro každý test zvlášť. Musí být nastavena tak, aby rozlišovala mezi studenty, kteří danou oblast dostatečně zvládli, a těmi, jejichž znalosti či dovednosti nejsou dostatečné k dalšímu postupu.

Kombinovaná standardizace spojuje do jisté míry výhody absolutního hodnocení s kompetitivním aspektem hodnocení relativního. Studenti, kteří dosáhli absolutní hranice pro úspěšné složení zkoušky, jsou rozřazeni do skupin a podle dosažených bodů jsou jim přiděleny známky.

Níže je uvedeno schéma, které ilustruje rozdílný přístup k hodnocení při relativní a absolutní standardizaci.



Obr. 7.2 **Příklad hodnocení výsledku studenta v testu z pohledu absolutní a relativní standardizace**.

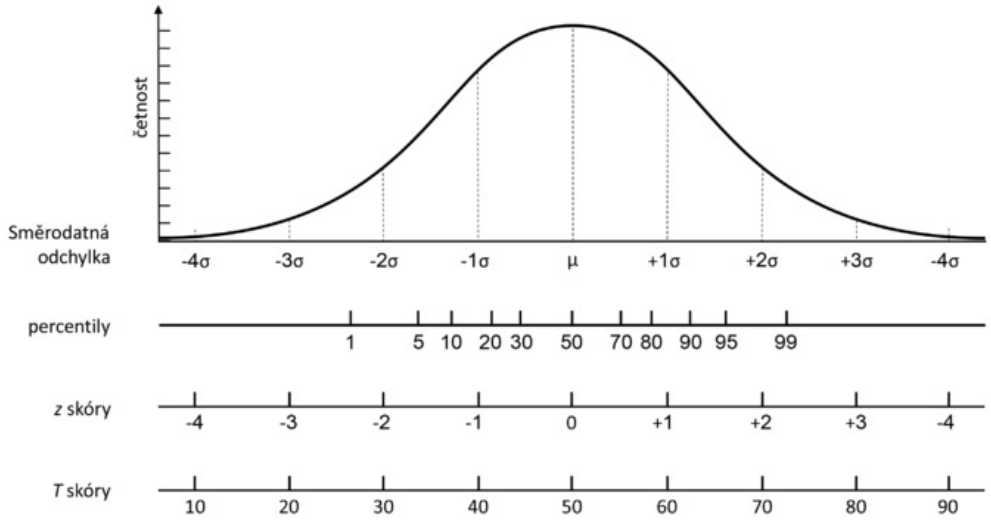
Absolutní standardizace srovnává výkon studenta s hranicí vědomostí vyžadovanou pro absolvování kurzu. Hranice se stanovují na základě expertního odhadu. Student A v testu uspěl, neboť dosáhl počtu bodů, který učitel považoval za minimum nezbytné pro absolvování zkoušky.

Relativní standardizace je založena na porovnávání výsledků studentů mezi sebou – skupina je rozčleněna podle dosaženého počtu bodů a oznámkována. Student A byl klasifikován známkou „dobře“. Relativní standardizaci můžeme provést až po testu, kdy jsou již známé výsledky studentů.

Relativní standardizace

Relativní standardizace je způsob vyhodnocení testu, při němž se výkon testovaného jedince porovnává s výkonem relevantní populace. To znamená, 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.

Nevýhodou relativní standardizace je, že hodnocení jednotlivce nezávisí jen na jeho výkonu, ale i na výkonech ostatních studentů. Relativní standardizace je vhodná pro porovnávání výkonu velkých skupin a neměla by být používána ve skupinách menších než 40 studentů.



Obr. 7.3 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.

> 💡 **Tip: Samotné celkové skóre, resp. celkový počet bodů, může dávat zkreslující obraz o výsledku studenta. Chcete-li znát výsledek studenta vzhledem k testované skupině, použijte některou z metod relativní standardizace.**

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, přičte se polovina studentů, kteří měli výsledek stejný jako daný student, a určí se, jak velkou část tvoří tato skupina. Percentilové pořadí PR_i 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 N_i je kumulativní četnost u daného výsledku, n_i je četnost daného výsledku a n je počet testovaných žáků. Kumulativní četnost vyjadřuje počet studentů, kteří dosáhli daného nebo horšího výsledku.

Uvažujme, že chceme vypočítat percentilová pořadí 30 testovaných studentů, kteří dosáhli následujících výsledků:

1, 5, 5, 8, 8, 8, 8, 15, 15, 15, 15, 16, 16, 16, 21, 21, 23, 23, 23, 23, 24, 25, 27, 28, 28, 28, 28, 28, 28, 28

*Nejprve sestavíme tabulku četností, tedy k danému výsledku (celkovému počtu bodů) uvedeme **četnost** všech studentů se stejným bodovým výsledkem. Poté spočítáme **kumulativní četnosti** jako součet četností v daném řádku tabulky a všech předchozích řádcích. Nakonec dopočítáme pro každý řádek percentilové pořadí dle výše uvedeného vztahu.*

Uvažujeme-li studenta, který získal v didaktickém testu 21 bodů, pak je

- *jeho celkové skóre v pořadí 6. nejhorší, tedy* $i = 6,$
- *četnost daného výsledku je* $n_6 = 2,$
- *kumulativní četnost daného výsledku je* $N_6 = 16,$
- *počet testovaných žáků byl* $n = 30.$

Tedy percentilové pořadí pro studenta s 21 body je

$$PR_6 = 100 \cdot \frac{N_6 - \frac{n_6}{2}}{n} = 100 \cdot \frac{16 - \frac{2}{2}}{30} = 50,$$

jinými slovy mezi 100 studenty by se náš student s 21 body umístil zhruba na 50. místě. Zhruba 50 % studentů testované (referenční) skupiny dosáhlo horšího výsledku než student s 21 body.

Tab. 7.2 Výpočet percentilů

i	***Počet bodů z testu***	***Četnost*** (n_i)	***Kumulativní četnost*** (N_i)	***Percentilové pořadí*** (PR_i)
1	*1*	*1*	*1*	*1,67*
2	*5*	*2*	*3*	*6,67*
3	*8*	*4*	*7*	*16,67*
4	*15*	*4*	*11*	*30,00*
5	*16*	*3*	*14*	*41,67*
6	***21***	***2***	***16***	***50,00***
7	*23*	*4*	*20*	*60,00*
8	*24*	*1*	*21*	*68,33*
9	*25*	*1*	*22*	*71,67*
10	*27*	*1*	*23*	*75,00*
11	*28*	*7*	*30*	*88,33*

z-skóry

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

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

Pokud například studenti dosáhli v testu průměrně $\bar{X} = 50$ *bodů a směrodatná odchylka byla* $SD = 20$*, pak student, který měl 30 bodů, má z-skór roven* $(30 - 50)/20 = -1$*. To znamená, že studentův výsledek je jednu směrodatnou odchylku pod celkovým průměrem. Pokud se celkové skóre řídí přibližně normálním rozdělením, pak platí, že přibližně 68 % studentů má výsledný počet bodů v rozmezí ±1 směrodatná odchylka od průměru, celkem 95 % studentů má výsledek testu v rozmezí ±2 směrodatné odchylky a až 99,75 % studentů v rozmezí ±3 směrodatné odchylky od průměru. Pokud je tedy studentův z-skór roven −1, znamená to, že celkem asi (100 − 68)/2 = 16 % studentů má výsledek testu nižší než zmíněný student.*

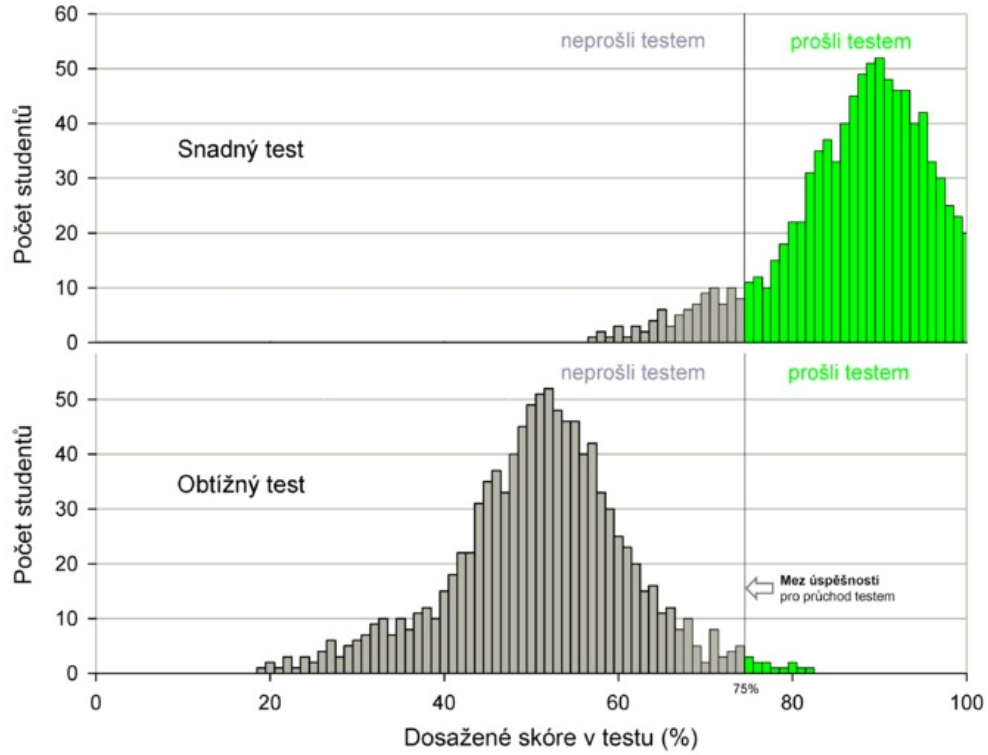
Pomocí z-skóru může vyučující snadno identifikovat žáky výtečné (z > 2) a naopak velmi slabé (z < −2). Snadno může také porovnat studentovy výsledky v různých částech testu.

Podrobnější rozbor dalších metod standardizace (C-škála, škála „stanin“) je k dispozici například v publikaci autorů Jeřábka a Bílka s názvem Teorie a praxe tvorby didaktických testů. **Cite error: Invalid <ref> tag; invalid names, e.g. too many**

Absolutní standardizace

Absolutní 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 řadili mezi relativně lepší uchazeče. 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“, víceméně arbitrární procentuální meze (50 %, 75 % apod.) bez hlubšího zdůvodnění. Přitom některé otázky v testu mohu mít zásadní význam a jiné mohou být jen okrajové.



Obr. 7.4 Dvojice grafů ukazuje, jaké důsledky by mohlo mít nastavení meze úspěšnosti v testu odhadem „od oka“. V případě snadného testu (obrázek nahoře) uspějí při nastavení limitu na 75 % téměř všichni. Stejný limit použitý pro jiný, obtížnější test (obrázek dole) způsobí, že testem projde jen pár žáků.

Abychom se vyhnuli problémům se špatně stanovenou mezí úspěšnosti, je třeba **kvalifikovaně posoudit obtížnost testu**, tedy provést jeho **absolutní standardizaci**. Dobře k tomu poslouží např. Angoffova či Ebelova metoda, které jsou popsány dále.

Existuje celá řada metod pro absolutní standardizaci různých typů hodnocení studentů. Jejich přehled nalezne čtenář například v obsáhlém díle Handbook of Test Development [44].

Většina metod vychází z **expertního posudku** relevantních odborníků, který se může zaměřit buď na jednotlivé položky testu (tedy jak která položka přispívá k rozlišení mezi úspěšným a neúspěšným studentem), nebo naopak na testovanou populaci (tedy zda test umí rozlišit mezi vhodnými a nevhodnými kandidáty).

Zaměřme se nyní podrobněji na dvě metody standardizace – Angoffovu a Ebelovu metodu. K praktickému provedení je v obou případech vhodné zajistit 6–8 odborníků v testovaném oboru.

Angoffova metoda

Patrně nejznámější z položkových metod standardizace je metoda podle **Angoffa**, resp. její modifikace podle Hambletona a Plakeové [45]. Principem metody je expertní odhad hraničního počtu bodů nutného k úspěšnému absolvování testu nebo úlohy. Pracuje se s hypotetickým tzv. minimálně kompetentním studentem. *Minimálně kompetentní student* je takový, který právě splňuje minimální požadavky kladené v daném oboru. Jinými slovy, je to nejslabší student, který by ještě měl testem projít.

Test (nebo položku) posuzuje skupina expertů (obvykle tým učitelů daného oboru či kurzu, kteří zodpovídají za přípravu a hodnocení testů, *examination committee*) 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 je na každém řádku určitá položka z testu a každý sloupec tvoří odhady jednoho experta.

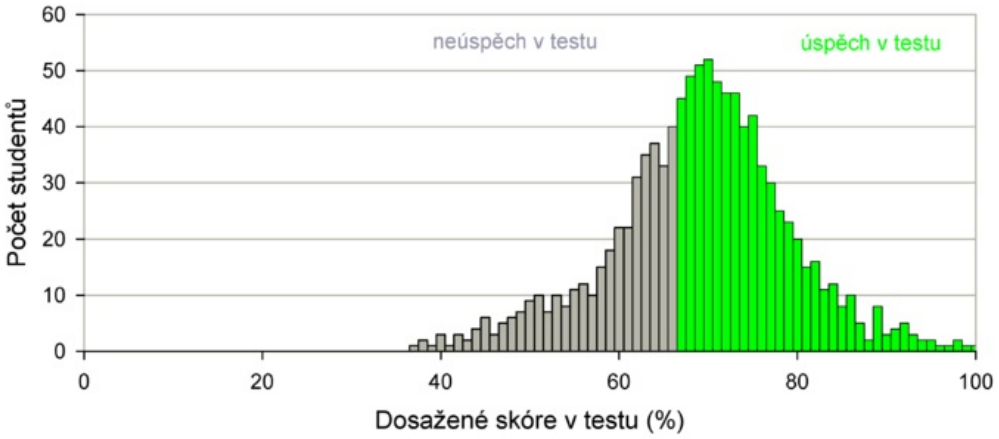
Tab. 7.3 Tabulka expertních odhadů pravděpodobnosti správného zodpovězení otázky minimálně kompetentním studentem

Číslo položky	Expert 1	Expert 2	Expert 3	Průměr
1	0,75	0,60	0,60	0,65
2	0,70	0,50	0,60	0,60
3	0,80	0,60	0,70	0,70
4	0,70	0,60	0,70	0,70
…	…	…	…	…
Průměr				**0,66, tj. 66 %**

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 % pravděpodobnost odpovědět správně. Analogicky u výběrové otázky s pěti možnostmi a jedinou správnou odpovědí bude minimum 0,2.

Po vyplnění tabulky se obvykle posuzuje, zda se experti ve svých odhadech shodli. Položky, v nichž je rozptyl odhadů velký, 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“.



Obr. 7.5 Mez úspěšnosti stanovená pomocí standardizační metody rozdělí soubor testovaných na úspěšné a neúspěšné. Standardizaci umožňují a podporují i některé testovací programy, například Rogo, z nějž můžete obdržet právě takovýto graf.

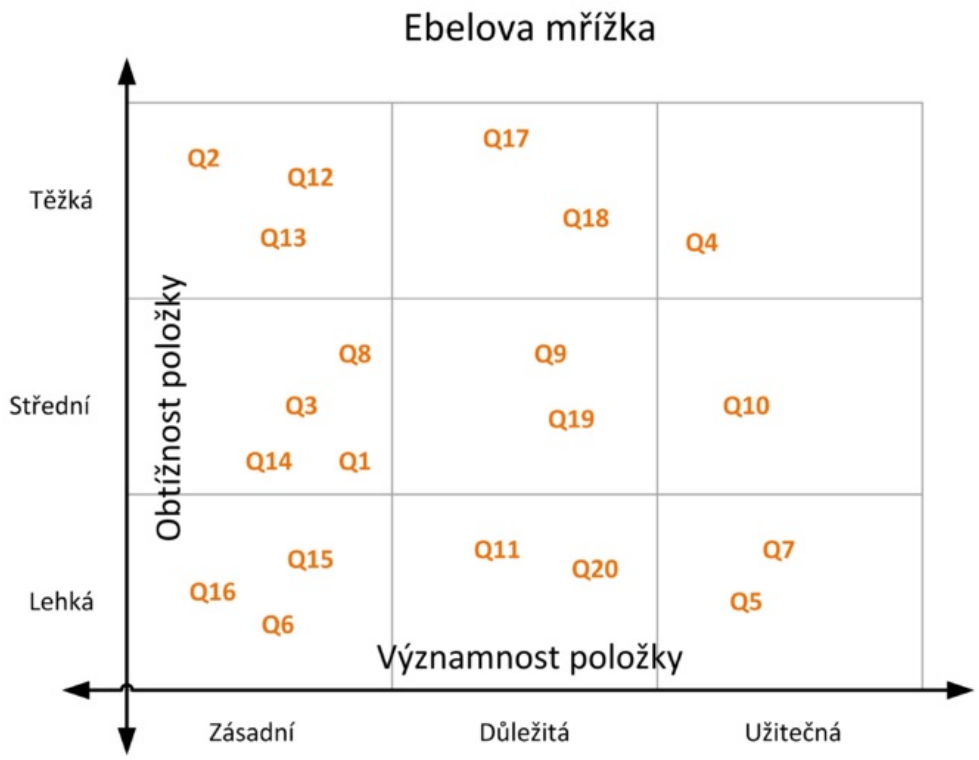
Pro úspěšné použití této metody je nutné, aby zúčastnění experti měli dostatek zkušeností v dané oblasti a poměrně přesnou představu o tom, co studenti v daném kurzu musí zvládnout. Experti si tedy musí umět představit, co minimálně kompetentní student umí, resp. by měl umět. Vzorovou tabulku ve formátu MS Excel, která je navržena v souladu s Angoffovou metodou standardizace najdete v příloze (http://is.muni.cz/www/98951/standardizace-angoffova-metoda.xls).

- *Dodatečná poznámka (CS 2018): Angoff anchor statements: setting a flawed gold standard? (https://www.researchgate.net/publication/319957789_Angoff_anchor_statements_setting_a_flawed_gold_standard)*

Ebelova metoda

Ebelova metoda má dvě modifikace. **Tradiční Ebelova metoda** slouží pro stanovení expertního odhadu minimálního výsledku, kterého by měl student dosáhnout, aby ještě prošel testem [46], [47]. **Modifikovaná Ebelova metoda** slouží pro přípravu obsahově validního testu [48], [49], [50].

V tradiční metodě podle **Ebela** nejprve experti roztřídí jednotlivé otázky do skupin podle dvou kritérií: **významu** a **obtížnosti**. Škála významu (relevance) položky jde od „zásadní“, přes „důležitá“ a „užitečná“ až k „irelevantní“. Položky označené experty jako „irelevantní“ se z dalšího použití vyloučí. Škála obtížnosti je trojstupňová: „těžká“, „střední“ a „snadná“ [51]. Zařazením položek do kategorií podle obou kritérií vznikne **Ebelova mřížka**:



Obr. 7.6 Ebelova mřížka

Předem bývá dohodnuto, že položky, u nichž se nedosáhne definované úrovně shody expertů (např. 80 %), budou diskutovány a případně vyloučeny. Po takovémto rozdělení položek odhadnou experti, jakou část otázek z každé kategorie by měl správně zodpovědět minimálně kompetentní student. Součiny těchto proporcí a počtů otázek v každé kategorii se poté sečtou a toto číslo se vydělí celkovým počtem otázek: výsledkem je hledaná hranice úspěšnosti.

Tab. 7.4 Ebelova metoda – krok 1
Experti rozdělí otázky podle dvou kritérií – **význam** a **obtížnost**

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

Tab. 7.5 Ebelova metoda – krok 2
Experti odhadnou úspěšnost zodpovězení otázkek z každé kategorie minimálně kompetentním studentem

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

Tab. 7.6 Ebelova metoda – krok 3
Vypočtou se parametry pro jednotlivé kategorie:
počet otázek · odhad úspěšnosti MKS

	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 můžeme provést i v tabulce:

Tab. 7.7 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 %**

Uvedený příklad je jen ilustrativní, v praxi bychom měli pracovat s většími počty otázek.

O významu Ebelovy metody svědčí i to, že Ebelova mřížka je součástí některých testovacích programů. Konkrétně je přímo zahrnuta v testovacím programu Rogo, o němž budeme podrobně mluvit v části věnované realizaci testů.

Klasifikace studentů

Klasifikace, čili oznámkování výkonu studentů v testu, je pokračováním či rozšířením standardizace testu. 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 **Cite error: Invalid <ref> tag; invalid names, e.g. too many**.

Pro nastavení relace mezi výkonem v testu a klasifikačními stupni je nutno si ujasnit, co je vlastně v daném případě smyslem vysokoškolského vzdělávání.

První z možných pohledů nahlíží na vysokoškolské medicínské vzdělávání jako na šestiletý test inteligence, který třídí jednotlivce podle intelektuálních schopností a pracovních návyků. Tento přístup odráží zájem potenciálních zaměstnavatelů vybrat nejvhodnější kandidáty na omezený počet míst a pomáhá zajistit, aby na klíčová místa byli vybráni intelektuálně nejschopnější. Při studiu staví tento přístup studenty proti sobě, nechává je mezi sebou soutěžit.

Druhý pohled je velmi 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, intelektuálními nástroji a návyky, které z nich učiní informované a kulturně gramotné členy společnosti.

Tyto dva pohledy na vzdělávání mají své paralely ve dvou hlavních přístupech ke klasifikaci.

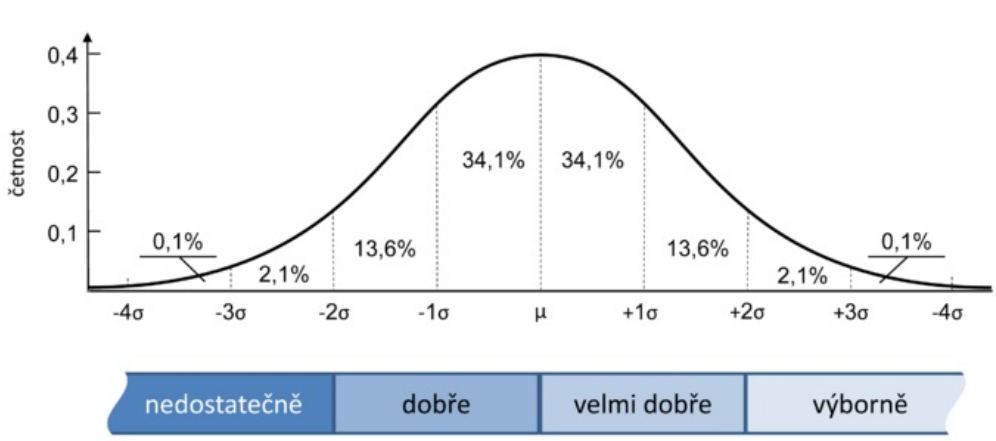
Klasifikace porovnávající výkonnost ve skupině (relativní klasifikace)

Prvním přístupem je klasifikace založená na **relativním výkonu ve skupině**. Stanovuje známku jako funkci pořadí studenta v rámci určité skupiny. Pomocí testu **vytvoříme pořadí studentů a těm rozdělíme klasifikační stupně** podle předem stanovených procentuálních mezí. Relativní klasifikace je založena na předpokladu, že výkonnost všech studijních skupin (napříč prostorem a časem) je v zásadě stejná. Z pohledu studenta obsahuje tento způsob klasifikace zjevnou nespravedlivost, protože hodnocení nezávisí jen na výkonu studenta, ale i na výkonech ostatních. Je tedy možné, že se stejnou mírou znalostí by student byl v jednom roce klasifikován lépe než v roce jiném.

Chceme-li použít hodnocení založené na relativní standardizaci, je třeba učinit dvě rozhodnutí. Nejprve je třeba stanovit, jaký klasifikační stupeň přiřadíme k průměrnému výkonu. Pro často používaný pětistupňový klasifikační systém A, B, C, D, F (1, 2, 3, 4, 5) můžeme intuitivně zvolit C jako klasifikaci odpovídající průměrnému výkonu; není to však jediná možnost.

Dále je nutné rozhodnout o hranicích oddělujících jednotlivé klasifikační stupně. Narazíme při tom na otázky typu „kde je hranice mezi mírně a silně nadprůměrným?“ Rozhodnutí je na zkušenosti a úsudku. Pomoci však mohou obecně platná doporučení.

Pro stanovení konkrétních známek se používá např. z-skór nebo percentilové pořadí podobným způsobem, jako je popsáno v kapitole Relativní standardizace testu. Při čtyřstupňové klasifikační stupnici pak hranicím mezi jednotlivými klasifikačními stupni odpovídají např. z-skóry –2, 0, 2, jak je naznačeno na obrázku 7.7:



***Obr. 7.7* Příklad klasifikace výsledku studenta v testu při použití relativní standardizace**. Skupina se rozdělí podle dosaženého výsledku a oznámkuje podle příslušné klasifikační stupnice – v tomto případě čtyřstupňové.

Klasifikace založená na kritériích (absolutní klasifikace)

Naproti tomu klasifikace založená na kritériích měří úspěšnost studenta **ve vztahu ke kritériím** vyžadovaným pro dosažení toho kterého klasifikačního stupně. Na rozdíl od normativní klasifikace není předem stanoveno rozdělení klasifikačních stupňů a studentovo oznámkování není ovlivněno výkonností ostatních. Pokud bychom teoreticky zkoušeli výrazně nadprůměrnou skupinu studentů, mohou všichni dostat dobré známky, a naopak, pokud by se náhodou sešla velmi slabá skupina studentů, nemusí dobré známky dostat nikdo.

Cílem klasifikace založené na kritériích je zhodnotit objektivní výsledky studenta nezávisle na výsledcích ostatních studentů [52].

- Hodnocení studentů podle **relativní standardizace** nám umožňuje zjistit, kteří studenti se naučili nejvíce, ale nezjistíme, jestli se studenti ve skutečnosti naučili dost.
- Naopak při hodnocení podle **absolutní standardizace** zjistíme, zda-li se studenti naučili dost, ale nemusíme být schopni určit, kteří studenti se naučili nejvíce [53].

Obě metody jsou ve skutečnosti do jisté míry provázané, protože při relativní klasifikaci jsou implicitně zahrnuta i absolutní kritéria, která měli na mysli pedagogové připravující test.

Nejjednodušším způsobem klasifikace je stanovit, kolik procent z celkového počtu bodů je třeba k dosažení určité klasifikace. Například pro známku A požadovat 90 % a více, pro B 80–90 % a tak dále. Problém tohoto přístupu je v arbitrárním nastavení hranic jednotlivých klasifikačních stupňů. Proto je vhodné u důležitějších zkoušek nastavit hranice pomocí expertního odhadu, například podle Angoffovy nebo Ebelovy metody.

Někdy je na závadu, že klasifikace neodráží různorodou obtížnost otázek. Studenti pak mohou dosáhnout vysokého počtu bodů i tím, že správně odpoví na mnoho snadných, méně důležitých otázek.

Propracovanější forma absolutní klasifikace proto 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 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 otázky postihující i schopnost nabyté vědomosti aplikovat.

Pro absolvování obou druhů testu lze nastavit odlišné požadavky na výkon studentů, jak je naznačeno v tabulce 7.8.

Tab. 7.8 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.

Klasifikace v praxi

V kapitole, která se věnuje praktické klasifikaci (známkování) studentských výsledků v testu, se omezíme na diskuzi o testech s uzavřenými položkami s výběrem odpovědi (*multiple choice*). Klasifikace dalších typů testových položek je diskutována v odborné literatuře (např. [54] nebo [55]).

Většina zahraničních univerzit má vypracovaná standardní klasifikační schémata, která pak umožňují srovnávání výsledků jednotlivých studentů uvnitř i mezi obory. Výsledek daného testu či celého souboru písemných i jiných hodnocení je tedy přepočítán na standardní škálu, podle níž jsou potom rozděleny známky.

Jako příklad můžeme vzít klasifikační schémata Univerzity v Edinburgu (http://www.ed.ac.uk/schools-departments/student-administration/exams/regulations/common-marking-scheme). Pro pregraduální lékařské obory je relevantní schéma CMS3 (CMS3:Bachelor of Medicine and Bachelor of Surgery) :

Tab. 7.9 CMS3 schéma

Počet bodů	Známka	Popis
90–100	A	Výborně
80–89	B	Velmi dobře
70–79	C	Dobře
60–69	D	Uspokojivě (Pass)

Přepočítávání pak funguje následovně: tvůrci testu pomocí jedné ze standardizačních metod stanoví minimální hranici pro úspěch v daném testu (tzv. *pass mark*) například na 80 % z maximálního možného počtu bodů. Jelikož známka D (pass) v CMS3 odpovídá 60–69 bodům bude pro účely známkování 80 % z maximálního možného počtu bodů odpovídat hodnocení 60 bodů. Pro známku A bude třeba v testu dosáhnout 95–100 % (tedy 90–100 bodů) z maximálního možného počtu bodů, pro B 90–94 %, pro C 85–89 % a pro D 80–84 %. Jinými slovy, nejprve určíme, kteří studenti v testu uspějí, a pak je mechanicky rozdělíme do jednotlivých klasifikačních stupňů. Přepočet lze matematicky vyjádřit následovně: $Z = 60 + \frac{40}{100 - P} \cdot (p - P)$, kde Z je výsledný počet bodů (z kterého dle CMS3 schématu určíme známku), P je minimální procento nutné k úspěšnému absolvování daného testu (*pass mark*) a p jsou procenta dosažená daným studentem.

Analýza výsledků a hodnocení kvality testu

Dopady výsledků testování mohou mít v praxi zásadní důsledky – např. rozhodnutí o přijetí, rozhodnutí o udělení atestace nebo o udělení titulu. Použijeme-li nevhodné testy, závěry učiněné z jejich výsledků mohou být zavádějící. Kvalita použitých testů a jejich položek je proto zásadní a je důležité ji pravidelně kontrolovat.

Jaké vlastnosti by měl mít kvalitní znalostní test? V první řadě by měl měřit znalost, kterou měřit chceme. Dále by měl měřit co nejpřesněji a výsledky by měly být reprodukovatelné, zadáme-li studentovi jinou verzi téhož testu. Testy a jednotlivé položky by měly být spravedlivé a neměly by zvýhodňovat některé skupiny. Jednotlivé položky by měly mít vhodnou obtížnost a měly by mít schopnost dobře rozlišovat mezi různými úrovněmi znalostí studentů. Jak ověřit, zda náš test všechny tyto vlastnosti splňuje? Mnohé nám napoví **analýza výsledků testu**.

Jak analýzu výsledků provést? Prvním krokem analýzy výsledků by měl být **souhrnný popis** a vhodné **grafické vyobrazení** celkových výsledků všech studentů. Taková sumarizace se často vyžaduje při vykazování výsledků nebo pro sledování vývoje výsledků u stejného testu v průběhu let. Ukážeme si ale, že může být také prvním nástrojem k identifikaci problémů testu, např. při nežádoucím prozrazení („vynesení“) části úloh.

V dalším kroku je důležité prozkoumat vlastnosti testu jako celku, především jeho **reliabilitu** neboli spolehlivost, a jeho **validitu**, tedy zda měří znalost, kterou měřit chceme. Ke kvalitě celého testu přispívají jednotlivé položky (otázky v testu), je proto potřeba sledovat také vlastnosti jednotlivých položek a nabízených distraktorů. Tato **položková analýza** v sobě zahrnuje odhady obtížnosti a citlivosti jednotlivých položek. Je důležitá také pro tvorbu dvou nebo více srovnatelných verzí testu.

V následujícím textu se převážně budeme věnovat klasickým odhadům vlastností testu a jeho položek. Klasické odhady mají však svá omezení. V první řadě tyto odhady závisí na úrovni znalosti testované populace. To je na závadu, používáme-li položky nebo celé testy pro skupiny studentů, které se od sebe výrazněji liší (např. v případě sdílení položkových bank). Nemůžeme pak nekriticky převzít popis určité položky, který například říká, že jde o položku snadnou

– pokud test aplikujeme na jiné skupině studentů, kteří se třeba vzdělávají jiným systémem, může pro ně tatáž položka být naopak velmi obtížná. V případě vyššího počtu testovaných studentů je tak vhodné použít k odhadům vlastností položek a úrovní znalostí studentů složitějších odhadů, tzv. **teorie odpovědi na položku** (*item response theory*, **IRT**). IRT umožní modelovat vlastnosti položky a celého testu pro různé úrovně znalosti studentů.

Nepředpokládáme, že by čtenář měl umět sám provést všechny níže uvedené analýzy. Záměrně uvádíme i složitější analýzy a rozebíráme výhody, aby čtenář získal představu o různých možnostech ověřování kvality testů. Cílem je, aby čtenář byl o metodách informován a věděl, co lze zjistit pomocí dostupného softwaru, nebo ve spolupráci se specialistou – statistikem.

💡 **Tip: Jak na to?**

Popis a grafické zobrazení výsledků

Prvním krokem analýzy by vždy měl být popis výsledků dosažených v testu pomocí souhrnných statistik a vhodného grafické vyobrazení. Souhrnný pohled na výsledky nám dá první představu o vlastnostech testu, může ale také upozornit na jeho problémy. Patrně každého zkoušejícího zajímá, jaký byl nejlepší dosažený výsledek a zda někdo dosáhl maximálního počtu bodů. Jaký je nejhorší výsledek? A pokud má několik studentů test zcela nesprávně, čím je to způsobeno, nenastala jen někde procesní chyba v hodnocení (například obodování testu podle šablony pro jinou variantu)? Celkovou obtížnost testu můžeme dále popsat průměrným počtem bodů. Seřadíme-li výsledky vzestupně, obtížnost můžeme popsat také prostřední hodnotou (tzv. mediánem). Zadáváme-li stejný test opakovaně, je dobré sledovat, jak se průměrný výsledek vyvíjí v letech. Můžeme tak například odhalit, že kvalita studentů se rok od roku snižuje. Nebo naopak, že si vedou studenti rok od roku lépe. Ve druhém případě je ale na místě otázka, zda je to lepšími znalostmi studentů, nebo tím, že test je již prozrazený (vynesený).

Směrodatná odchylka vypovídá o rozptýlenosti výsledků. Jak již víme, přibližně 95 % výsledků bude ležet ± 2 směrodatné odchylky od průměru (viz také kapitola **z-skór studenta**). To ovšem pouze v případě, že se výsledky řídí normálním (neboli tzv. Gaussovým) rozdělením – což by obecně měly, ale je potřeba to také ověřit.

Tabulka popisných statistik pro celkové výsledky může vypadat např. takto:

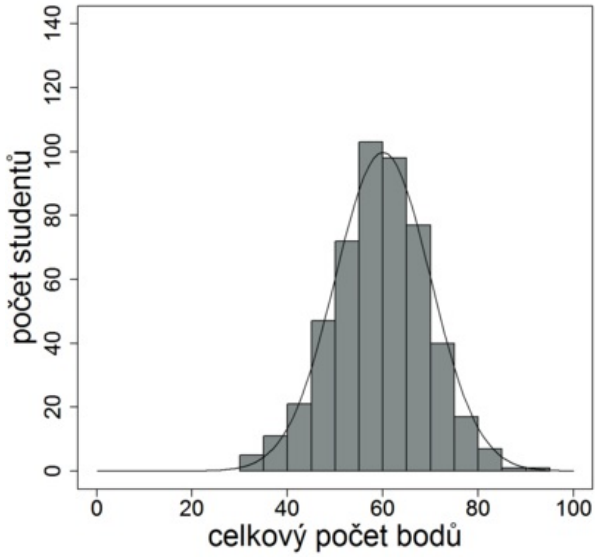
Tab. 8.1 Tabulka popisných statistik

minimální možný počet bodů	0
maximální možný počet bodů	100
dosažené minimum	27
dosažené maximum	99
průměr	68,8
medián	68,0
směrodatná odchylka	15,3

Pokud se průměr od mediánu výrazněji liší, může to indikovat nesymetrii výsledků, potažmo nenormalitu dat. *Šikmost* rozdělení může souviset s celkovou přílišnou jednoduchostí či obtížností testu. Je-li test například příliš snadný a větší množství výsledků se pohybuje u maximálního možného počtu bodů (víc bodů získat nelze), delší „chvost“ je nalevo.

Pro celkový výsledek studenta v testu se odborně používá termín **hrubý skór.** Většinou je hrubý skór počítán jako součet bodů za jednotlivé položky; někdy však může být počítán složitěji.

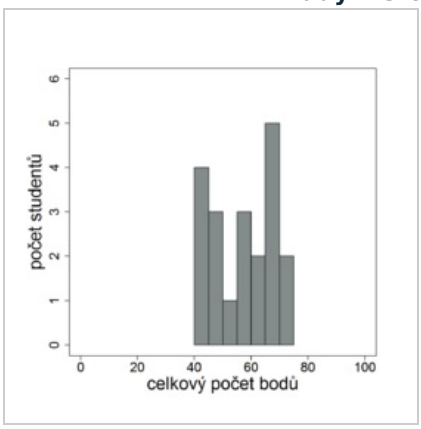
Graficky lze rozdělení hrubých skórů ve skupině testovaných studentů posoudit pomocí histogramu, který jsme zahlédli již v předchozích částech textu. **Histogram** je sloupcový graf, v němž výška sloupců vyjadřuje četnost sledované veličiny v daném intervalu. Lze jej chápat jako odhad hustoty rozdělení vědomostí v populaci. V běžném případě očekáváme, že se rozdělení znalostí řídí normálním rozdělením. Pro posouzení, zda tomu tak skutečně je, lze histogram proložit křivkou normálního rozdělení (přičemž střední hodnotu a rozptyl odhadujeme z dat), viz obr. 8.1.



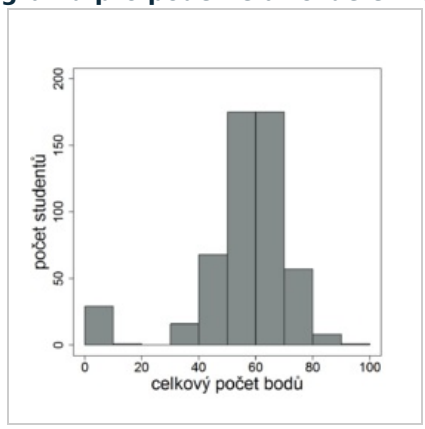
Obr. 8.1 **Histogram celkových skóre u studentů a jím proložená křivka normálního rozdělení**

Histogram nemusí odpovídat křivce normálního rozdělení, máme-li malý počet studentů (viz obr. 8.2a). Máme-li studentů dostatek, histogram by křivce normálního rozdělení odpovídat měl. V opačném případě je to upozorněním na možné problémy: Naznačuje-li graf nějaké neobvyklé výsledky, tzv. odlehlé hodnoty (viz obr. 8.2b, např. několik málo jedinců má počet bodů mnohem nižší nebo mnohem vyšší než ostatní), je potřeba se zamyslet nad jejich příčinami. Nebyl v těchto případech test obodován podle špatné šablony? Alarmující je také dvouvrcholový histogram (viz obr. 8.2c). Naznačuje, že námi testovaní studenti jsou ve skutečnosti směsí dvou nebo více různých skupin s různými vlastnostmi a patrně také s různými podmínkami pro úspěch v testu. Důvodem mohou být různí vyučující, různí cvičící, ale také např. vynesení testu. Takové skutečnosti je pak potřeba brát v potaz i v dalších níže uvedených analýzách (a např. pro případ dvou nebo více skupin je potřeba buď výsledky skupin analyzovat samostatně, nebo použít složitějších modelů).

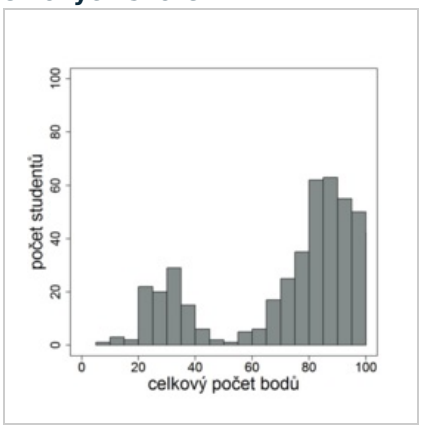
Příklady histogramů pro podezřelá rozdělení celkových skóre



Obr. 8.2a Histogram v případě malého počtu studentů



Obr. 8.2b Histogram v případě odlehlých hodnot



Obr. 8.2c Histogram v případě dvou skupin s různými vlastnostmi

Reliabilita a validita testu

Dobrý didaktický test by měl měřit co nejpřesněji a měl by měřit to, co chceme, aby měřil. Tyto vlastnosti jsou popsány konceptem reliability a validity.

Reliabilita neboli **spolehlivost** vypovídá o tom, nakolik je výsledek testu ovlivněn náhodnou chybou – třeba tím, že student na část otázek nezná odpověď a pouze „tipuje“. Jinými slovy, reliabilita říká, do jaké míry by se shodla dvě nezávislá testování téhož studenta.

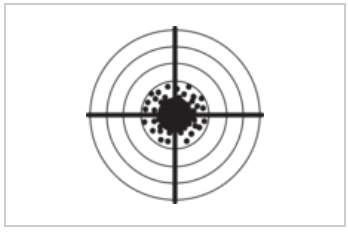
Validita neboli **správnost** popisuje, do jaké míry test měří tu vlastnost (znalost), kterou chceme, aby ve skutečnosti měřil. Ptáme se tedy, zda test skutečně zkouší učivo konkrétního předmětu, a ne něco úplně jiného: například znalosti získané jinde, schopnost porozumět složitě formulovaným větným konstruktům nebo schopnost odhadnout, co měl na mysli autor testu složeného z mnoha nejednoznačných otázek.

Reliabilita je nutným předpokladem validity, z reliability ale validita přímo neplyne. Tyto dva koncepty a vztah mezi nimi lze dobře ilustrovat na příkladu střílení na terč.

Ilustrace pojmů reliability a validity na příkladu střílení na terč

Obr. 8.3a Nízká **reliabilita**, tudíž nízká **validita**

Obr. 8.3b Vysoká **reliabilita** a nízká **validita**



Obr. 8.3c Vysoká **reliabilita** i **validita**

Reliabilita vypovídá o tom, nakolik jsou střely rozptýlené. Validita pak vypovídá o tom, jak často střely trefují cíl. Padají-li výstřely příliš daleko od sebe, spolehlivost (reliabilita) střelce je malá. Nelze pak mluvit ani o validitě, jsou-li totiž výstřely příliš rozptýlené, cíl trefí málokdy.

Dopadají-li střely blízko sebe, lze mluvit o vysoké spolehlivosti střelce (podobně didaktický test vnímáme jako spolehlivý, pokud bychom při nezávislých administracích daného testu opakovaně naměřili u stejného studenta stejné hodnoty). To nám však ještě nezaručuje, že střelec trefuje správně cíl, může totiž spolehlivě trefovat zcela jiné místo (stejně jako didaktický test může zcela přesně měřit jinou vlastnost, než kterou chceme měřit).

Spolehlivost je tedy nutným předpokladem validity, není ale předpokladem jediným. Přesný střelec musí spolehlivě trefovat střed terče. Právě tak validní didaktický test musí spolehlivě měřit tu znalost, kterou chceme, aby měřil.

Odhady reliability (spolehlivosti) testu

Jsou výsledky testu konzistentní? Reliabilita neboli spolehlivost měří, nakolik jsou výsledky zopakovatelné, nakolik *výstřely padají blízko sebe*. Je tak jednou z fundamentálních vlastností testu. Neexistuje však pouze jediný správný způsob jejího odhadování [57]. Použijeme-li kterýkoliv z přístupů odhadu reliability, měli bychom rozumět principům a teoriím, na kterých je založen, jakož i omezením, která s sebou předpoklady daného přístupu přinášejí [7]. Jaké jsou tedy možnosti odhadu reliability?

Odhad metodou test-retest

Přímo z definice reliability se nabízí myšlenka zadat test studentům dvakrát po sobě a měřit závislost mezi dvěma výsledky pomocí korelačního koeficientu. Tento odhad reliability je často používán např. pro odhad měřicích přístrojů (tlaku, váhy apod.). V případě didaktického testování je ale jeho využití velmi omezené. Studenti si totiž otázky pamatují a mají tendenci odpovídat konzistentně, což by odhad reliability nadhodnotilo. Na rozdíl od opakovaného střílení na terč nebo měření tlaku proto nelze dvě administrace stejného znalostního testu považovat za nezávislé. V případě delšího časového intervalu se zase studenti správná řešení mohou naučit (nebo je zapomenout). Korelace mezi dvěma časově vzdálenějšími výsledky pak spíš vypovídají o **stabilitě znalosti v čase**.

Odhad pomocí paralelních forem testu

Abychom vyloučili vliv paměti při použití metody test-retest, lze vytvořit dvě tzv. paralelní formy testu. Stručně řečeno to jsou dva testy, které měří danou znalost stejným způsobem, mají stejné průměry celkových skóre, stejné směrodatné odchylky a stejné korelace s jinými testy, a to v jakékoliv populaci. Paralelní formy lze vytvořit např. změnou číselných údajů v zadání jednotlivých příkladů apod. Odhad reliability je pak opět založen na korelaci mezi dvěma celkovými skóre. Tento způsob odhadu reliability lze v didaktickém testování nadmíru doporučit. Jak lze tušit, omezením je obtížnost vytvoření dvou verzí testu, které by byly skutečně paralelními formami. Odhad pak může do jisté míry odrážet, nakolik jsou dvě verze testu skutečně ekvivalentní. Dalším omezením může být neochota respondentů k dvojímu testování. Výsledky v druhém testu také mohou být ovlivněny zvýšenou únavou studentů.

Odhad metodou split-half

V mnoha případech není možné testovat studenty vícekrát, a je tedy zapotřebí odhadovat reliabilitu z jediné administrace testu. Jedním z přístupů je rozdělení testu na dvě poloviny a zkoumání korelace mezi celkovými skóre v těchto dvou polovinách. Pokud test rovnou konstruujeme jako dvojice paralelních položek, můžeme tento přístup vnímat jako odhad reliability paralelních forem, které však mají poloviční délku. Reliabilita je závislá na počtu položek (testy s více položkami mívají reliabilitu vyšší), je proto nutná úprava odhadu na skutečnou délku testu pomocí tzv. *Spearmanovy-Brownovy formule* [58], [59]. Výslednému odhadu se říká **split-half reliabilita**.

Pokud test nekonstruujeme přímo jako dvojice odpovídajících si položek, může být oříškem, které z možných rozpůlení zvolit. Nabízí se rozdělení na první a druhou polovinu, na sudé a liché položky, nebo lze položky rozdělit náhodně. Každé rozpůlení dá ale jiný odhad spolehlivosti. Jedním z řešení je rozdělit test postupně na všechny možné poloviny a získané odhady zprůměrovat, odhadu se pak říká **průměrný split-half koeficient**. Číselně blízké a přitom výpočetně o dost jednodušší je rozdělit test na nejmenší možné části (tedy na jednotlivé položky) a odhadovat reliabilitu pomocí korelací mezi nimi. To je základem myšlenky Cronbachova alfa.

Reliabilita jako vnitřní konzistence položek – Cronbachovo alfa

Nejčastěji používaným odhadem reliability je **Cronbachovo alfa**, jehož vzorec lze najít v příloze. Je oblíbený z několika důvodů: jeho výpočet je jednoznačný, srozumitelný a je implementován do všech statistických programů. Cronbachovo alfa měří závislosti mezi jednotlivými položkami, je tudíž mírou vnitřní konzistence testu. Za jeho používáním coby

odhadu spolehlivosti stojí představa, že všechny položky testu měří jedinou vlastnost, síla jejich závislosti je vysoká a jediné rozdíly jsou tudíž způsobené chybou měření.

Cronbachovo alfa je pojmenováno po Lee Cronbachovi, který jej proslavil v roce 1951 [60]. Hodnotu rovnou jedné dostaneme, pokud jsou položky svázány lineárně (v takovém případě by ovšem stačila položka jediná). Malá hodnota naopak vypovídá o nízké vnitřní konzistenci položek, nebo nízké spolehlivosti [61].

- **Bacha na Cronbacha!**

Psychometři upozorňují na mnohá omezení a dezinterpretace Cronbachova alfa [62]. Jedním z omezení je již naznačená skutečnost, že je pouze dolní mezí reliability a silně ji podhodnocuje v případě, kdy položky nepopisují stejnou oblast vědomostí. Alternativní koeficienty vhodné pro případ složitější struktury znalostí vyžadovaných testem jsou popsány v článku [63].

Cronbachovo alfa také nefunguje správně, pokud jsou vedle náhodných chyb v měření obsaženy další chyby, např. efekt posluchárny, ve které byl test skládán (skládají-li studenti test v různých posluchárnách, lze si představit, že v některých posluchárnách studenty ruší hluk z ulice), efekt zkoušejícího (zkouší-li více examinátorů) apod. S takovými situacemi se vypořádává tzv. teorie zobecnitelnosti [64], [65], která používá složitějších modelů analýzy rozptylu.

Cronbachovo alfa vychází z modelu normálně rozdělených položkových skóre. Je tedy na místě otázka, nakolik správné je jeho použití v případě položek typu ano/ne. Některá zobecnění byla navržena v článcích [66] a [67].

Nakonec upozornění, které platí pro všechny typy odhadů reliability, avšak pro Cronachovo alfa obzvlášť [68]: Již v definici reliability je obsažena důležitá vlastnost, a sice že je závislá na homogenitě testovaných jedinců. Budeme-li testovat skupinu studentů s podobnými znalostmi (např. jednu studijní skupinu), odhad reliability bude nižší než v případě, kdy budeme testovat skupinu méně homogenní – např. studenty různých ročníků. Odhadujeme-li reliabilitu testu, je proto nutné předem stanovit, pro jakou skupinu je test určen (jeden obor, celý ročník, celou školu) a odhady je pak nutné počítat z odpovídajících dat.

Demonstrujme význam Cronbachova alfa 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 otázek 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 otázky z testu testují totéž. Cronbachovo alfa se sníží.

Mluvíme-li o vnitřní konzistenci testu, měli bychom si uvědomit, že nezávisí jen na samotných otázkách, 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.

Shoda v rozhodnutí o úspěchu/neúspěchu

Dosud jsme se reliabilitou zabývali v kontextu testů relativního výkonu, které mají za úkol rozlišit mezi jednotlivými studenty. Uvědomme si nyní, že odhad reliability založený na korelaci mezi výsledky dvou testů (např. dvou paralelních forem) nebere v potaz obtížnost těchto testů, a může tak nadhodnotit spolehlivost z perspektivy testu absolutního výkonu. Představme si, že testujeme dvěma testy pět studentů a obdržíme tyto výsledky:

Tab. 8.2 Tabulka ukázkových výsledků pěti studentů ve dvou

výsledků pěti studentů ve dvou testech

	počet bodů v prvním testu	počet bodů v druhém testu
1. student	18	48
2. student	45	75
3. student	33	63
4. student	48	78
5. student	51	81

Pearsonův korelační koeficient je roven 1, mezi dvěma výsledky je přímo lineární závislost (druhý je vždy právě o 30 bodů vyšší než první). Reliabilita (coby korelace paralelních forem) je tedy vysoká. Budeme-li však rozhodovat o složení zkoušky na základě dosažení 50 bodů, první test úspěšně absolvuje 1 student z 5, zatímco v druhém to budou 4 studenti z 5. Pro popsání shody v rozhodnutí o úspěchu/neúspěchu proto nelze použít dosud zmíněné nástroje. Místo toho se odhadují tzv. **indexy shody**. Jedním z nich je například **Cohenovo kappa**. To se počítá z procentuální shody mezi testy: v našem případě by testy rozhodly o (ne)úspěchu shodně jen u prvního a posledního studenta, tedy ve 40 % případů. Cohenovo kappa poměřuje procento shody s pravděpodobností náhodné shody, vychází tak v našem případě ještě o něco menší, a to 0,12 (viz také příloha). Hodnoty kappa vyšší než 0,7 ukazují na velmi dobrou shodu, hodnoty mezi 0,4 a 0,7 na dostatečnou shodu, menší hodnoty, stejně jako v našem případě, na nedostatečnou shodu. Jak plyne z uvedeného příkladu, dva testy mohou mít pro posouzení relativního výkonu vysokou míru ekvivalence, a přesto mohou mít jeho dvě verze nízkou míru shody co do hodnocení absolutního výkonu. Pro odhadování spolehlivosti je proto nutné rozlišit, za jakým účelem je test zadáván a použít pak správné míry spolehlivosti.

Shoda posuzovatelů, zkoušejících nebo komisí

Pokud hodnotí výkon studentů různí posuzovatelé (např. při ústním zkoušení, při hodnocení esejí apod.), je nutné zajistit jejich srovnatelnost. Jak tedy odhadnout, nakolik dva zkoušející hodnotí stejně? Prvním krokem může být spočítat průměrná hodnocení, která jednotliví zkoušející udělili. Pokud např. jeden zkoušející dává v průměru známku 2,4 a druhý 3,6, každý student bude vědět, kterého zkoušejícího si vybrat (bude-li mít tu možnost). Průměrná hodnocení ale mohou být ovlivněna tím, jaké studenty ten který zkoušející hodnotil. Pokud například první zkoušející hodnotil pouze v předtermínu (dá se tedy očekávat, že šlo spíše o snaživější a lépe připravené studenty) a druhý naopak hodnotil jen studenty, kteří měli problémy s obdržením zápočtu, mohlo být hodnocení prvního zkoušejícího ve skutečnosti přísnější než hodnocení druhého. Proto jediným možným způsobem posouzení shody posuzovatelů je zajistit (alespoň na části zkoušených) nezávislé hodnocení od obou zkoušejících. Shodu mezi dvěma výsledky pak už můžeme odhadovat pomocí výše popsaných metod, např. pomocí koeficientu kappa (jde-li o hodnocení úspěchu/neúspěchu), nebo dalších indexů shody.

Odhady validity (správnosti) testu

Měří test to, co chceme, aby měřil? Snad nejdůležitější vlastností testu je, aby měřil skutečně tu znalost nebo dovednost, kterou měřit chceme. Pokud test měří něco jiného, mohou být jeho výsledky zcela nesprávně interpretovány. Míru, do jaké test měří skutečně to, co má, nazýváme **validita**. Přístupů, jak měřit validitu, je celá řada. Tyto přístupy se vzájemně doplňují a tvoří tak různé zdroje důkazů o validitě (podrobněji viz [25]). Jednodušeji pak osvětluje různé druhy validity kniha [69]. Zde pojednáme jen některé z nich.

Validita obsahová

Obsahová validita hodnotí vztah mezi obsahem testu a oborem, jehož znalost má měřit. Zkušený pedagog by měl posoudit, nakolik otázky obsažené v testu pokrývají zkoušenou látku, nakolik reprezentativní je zastoupení otázek z jednotlivých okruhů, zda všechny otázky spadají do zkoušené látky a nezkoušejí spíše nějakou jinou schopnost, vlastnost nebo znalost. Posouzení obsahové validity je zčásti zpětným pohledem na plán testu a zhodnocením jeho kvality, ale také posouzením, zda byl plán testu skutečně dodržen.

Validita kriteriální

Na prokázání skutečnosti, že daný test skutečně měří znalost, kterou měřit chceme, samotná obsahová validita nepostačí. Ta může prokázat pouze skutečnost, že byla vyvinuta snaha pokrýt obsah daného kurikula. Pro doložení validity testu je obecně nutné ještě nějaké další kritérium měřeného atributu, nezávislé na našem didaktickém testu. Validita, které se v tomto kontextu říká validita kriteriální, se pak odhaduje zkoumáním **vztahu výsledku testu k danému nezávislému kritériu**. Jak lze tušit, často největším problémem prokazování kriteriální validity je právě nalezení vhodného kritéria.

- Máme-li k dispozici jiný, již ověřený test, můžeme zjišťovat **závislost mezi naším testem a** zmíněným **ověřeným testem**. Tomuto druhu kriteriální validity se pak říká **validita souběžná** (*concurrent validity*). V praxi se odhaduje tak, že skupině studentů zadáme jak nový test, tak test ověřený (dosud používaný) a měříme závislost mezi dvěma celkovými skóre např. pomocí korelačního koeficientu.
- Představu o tom, do jaké míry náš test **předpovídá budoucí hodnoty nějakého kritéria**, nám dává další typ

kriteriální validity, tzv. **validita predikční**. 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.
- Dále nás může zajímat, zdali daný test **přináší novou informaci nad tu, kterou získáme jiným prováděným testem**, 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 poskytne jeho středoškolský prospěch. Např. studie [70] 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 (v roce přijetí nebo dříve) dokonce na 30 %. Všechny zmíněné efekty byly v modelu signifikantní (tedy statisticky průkazné), prokázala se tak jejich přírůstková validita.

Statistické metody použité k **odhadu kriteriální validity** testu závisí především na typu proměnné, která vystupuje jako kritérium. Lze-li tuto proměnnou chápat jako spojitou (např. počet bodů v jiném testu, nebo úspěšnost měřená průměrným prospěchem v prvním roce vysokoškolského studia), můžeme k odhadu použít Pearsonův nebo jiný korelační koeficient. Obecněji lze závislost na kritériu uchopit pomocí **regresní analýzy** [71]. Odhadnutý regresní model se na základě našich dat snaží pro každý výsledek testu předpovědět hodnotu kritéria (počtu bodů v referenčním testu, průměrný prospěch v následném studiu, úspěšnost studia apod.). Testováním významnosti jednotlivých složek modelu lze prokázat validitu testu nebo jeho přírůstkovou validitu nad již používané testy. Podrobněji se odhadu validity (v kontextu přijímacích zkoušek na vysoké školy) věnuje Zvára v publikaci [72].

Položková analýza

Jak poznám špatnou otázku? Špatnou otázku může zkušený pedagog odhalit už v rámci oponentury testu. Nejlépe ji ale identifikují sami studenti – tím, jak na ni odpovídají. Vyplatí se proto odpovědi studentů analyzovat a hodnotit z nich vlastnosti položek. Výhoda je, že takovou *položkovou analýzu* můžeme (obzvlášť v rámci elektronického testování) snadno automatizovat, nemusí ji tedy provádět sám pedagog. Postačí, pokud výsledkům položkové analýzy rozumí a umí z nich vyvodit správné závěry.

Položkovou analýzu je vhodné provést už před ostrým nasazením testu, zadáme-li položky vybranému vzorku studentů v rámci pilotního testování. Při pilotním testování je potřeba zajistit, aby testovaný vzorek studentů co nejlépe odpovídal populaci studentů, pro které je test určen. Položková analýza by měla být prováděna také po ostrém testování na cílové skupině studentů.

U testových úloh sledujeme především jejich **obtížnost** a jejich **citlivost** neboli rozlišovací schopnost. Obtížnost popisuje, jaké procento testované skupiny odpoví správně. Citlivost říká, nakolik je položka schopná rozlišit mezi lepšími a horšími studenty. Studenty přitom na lepší a horší dělíme většinou podle celkového výsledku v testu, lze ale usuzovat také na základě jiného kritéria, např. na základě úspěšnosti v následném studiu apod. Na správně formulovanou položku by lepší studenti měli odpovídat správně častěji než studenti slabší.

Obtížnost a citlivost jdou ruku v ruce. Velmi obtížné a velmi snadné úlohy mívají malou rozlišovací schopnost, proto by jich obzvláště v testu, který má za cíl rozlišit testované studenty (např. přijímací testy), nemělo být příliš mnoho. Jednu nebo dvě velmi snadné úlohy lze použít v úvodu testu za účelem uklidnění žáků a zvýšení jejich motivace. Obtížnost dalších položek by v optimálním případě měla mít vzrůstající tendenci.

Další důležitou vlastností položky je její **férovost**. Položka by neměla zvýhodňovat některé skupiny studentů. Stejně chytří a stejně dobře připravení studenti různých skupin (děleno dle pohlaví, etnika apod.) by měli správnou odpověď na položku volit zhruba stejně často.

Položka je také podezřelá, pokud ji vynechalo větší množství studentů. Pokud se takové položky vyskytují na konci testu, může být kromě obtížnosti položek důvodem také příliš krátká doba stanovená k řešení testu (blíže viz **kapitola 8.3.4 Analýza vynechaných odpovědí**).

U výběrových úloh s distraktory (tj. „nesprávnými odpověďmi“) je potřeba zkontrolovat také vlastnosti jednotlivých nabízených odpovědí. Distraktory, které nejsou dostatečně atraktivní (volilo je příliš malé procento studentů), nebo které dostatečně nerozlišují mezi horšími a lepšími studenty, je potřeba nahradit jinými.

Podívejme se nyní blíže na jednotlivé nástroje položkové analýzy.

Obtížnost položky

```
Není položka příliš snadná nebo příliš těžká?
```

Obtížnost testových úloh typu ano/ne se obvykle odhaduje jako podíl respondentů, kteří na danou úlohu odpověděli správně. Odhadu se říká index obtížnosti a značí se P:

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

Index obtížnosti nabývá hodnot mezi 0 a 1. Je tím blíže 1, čím víc studentů na položku zodpovědělo správně. Spíš bychom tedy měli mluvit o „snadnosti položky“. Někteří autoři proto zavádějí ještě **hodnotu obtížnosti.** Hodnota obtížnosti je daná procentuálním zastoupením žáků, kteří na položku odpověděli nesprávně, jde tedy o doplněk indexu obtížnosti:

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

V případě bodované položky, za kterou lze dosáhnout 0 až x_{MAX} bodů, se index obtížnosti obecněji definuje pomocí aritmetického průměru $\bar{x}$ bodových hodnocení všech žáků v dané úloze:

$$P = \frac{\bar{x}}{x_{MAX}}.$$

Velmi obtížné (P < 0,2) a velmi snadné (P > 0,8) úlohy mívají malou rozlišovací schopnost, proto by jich v testu nemělo být příliš mnoho. Malé množství velmi snadných úloh lze použít v úvodu testu za účelem uklidnění žáků a zvýšení jejich motivace. Obzvláště pro testy, jejichž cílem je rozlišit testované studenty (např. přijímací testy), jsou optimální položky s obtížností okolo 0,5.

Je potřeba zdůraznit, že zde zavedený odhad obtížnosti je závislý na testované populaci. Budeme-li obtížnost odhadovat na základě odpovědí skupiny chytřejších studentů, bude se položka jevit jako snazší. Výše uvedený klasický odhad je proto platný vždy jen pro vzájemně podobné skupiny studentů. Hodí se např., pokud test každoročně dáváme srovnatelným skupinám studentů. Problém nastává, pokud položky využíváme např. pro různé studijní obory, na více školách apod. Obecnější přístup, kdy se obtížnost položky definuje v závislosti na schopnostech testovaných studentů, používá tzv. teorie odpovědi na položku.

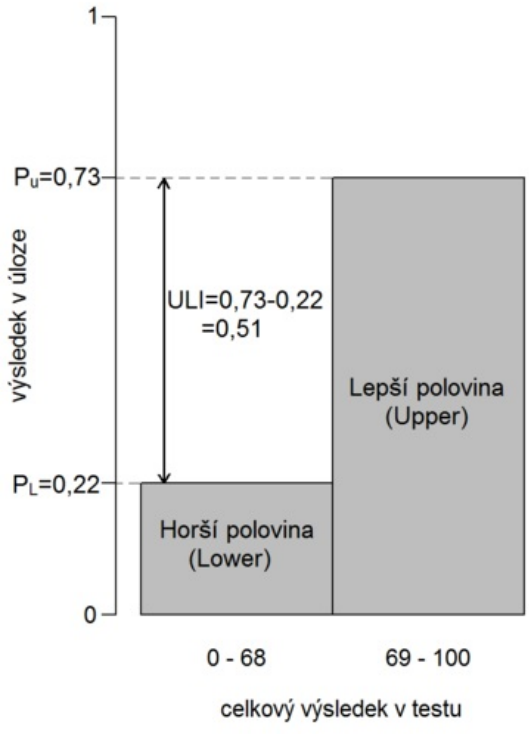
Citlivost položky

Rozlišuje položka dobře mezi lepšími a horšími studenty?

Citlivost neboli **diskriminační schopnost** vypovídá o schopnosti položky rozlišovat mezi lepšími a horšími studenty. Jedním ze způsobů odhadu citlivosti je tzv. **upper-lower index (ULI)**. Počítá se jako rozdíl úspěšnosti v položce mezi skupinou lepších (U – *upper*) a horších (L – *lower*) studentů:

$$ULI = P_U - P_L,$$

viz také obr. 8.4. Studenti jsou zpravidla děleni do dvou podobně velkých skupin (optimálně na polovinu) na základě celkového počtu bodů v testu.



Obr. 8.4 **Index ULI.** Studenti byli rozděleni na dvě podobně velké skupiny podle celkového výsledku v testu. Citlivost konkrétní položky v testu můžeme popsat indexem ULI: na danou otázku správně odpovědělo 73 % studentů z „lepší“ poloviny a 22 % z „horší“ poloviny. ULI se definuje jako rozdíl mezi těmito dvěma čísly, tj. v tomto případě 51 % = 0,51.

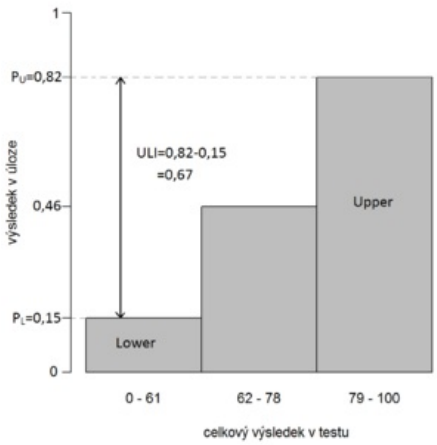
ULI nabývá hodnot mezi −1 a 1. Hodnoty ULI blízké 1 mají úlohy, které dobří žáci řeší správně a slabší žáci řeší nesprávně. Nulovou hodnotu mají úlohy, které obě skupiny žáků řeší stejně dobře. Úlohy se záporným indexem ULI zvýhodňují slabší studenty. Mohou to být úlohy s komplikovaným zadáním, které se lepší žáci neúspěšně pokoušejí řešit,

zatímco horší je s větším štěstím správně uhodnou. Úlohy se zápornou citlivostí by se v testu neměly objevovat. U položek příliš snadných nebo příliš obtížných (kterých by obzvláště v rozlišovacím testu nemělo být mnoho) lze očekávat nízkou rozlišovací schopnost. Byčkovský v [71] dále uvádí, že

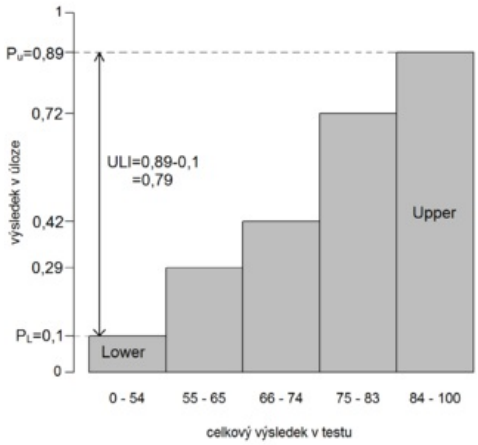
- pro položky s obtížností mezi 0,2 a 0,3 a obtížností mezi 0,7 a 0,8 je za podezřelé potřeba považovat úlohy s citlivostí ULI < 0,15,
- pro úlohy s obtížností mezi 0,3 a 0,7 se za podezřelé považují již položky s rozlišovací schopností ULI < 0,25.

Jelikož prostřední hodnoty celkového skóre v testu může dosáhnout současně více studentů, vyvstává otázka, jak s těmito studenty při dělení na polovinu naložit. Je možné zařadit je do jedné ze skupin (ale do které?), nebo rozdělit (ale jak?) a část zařadit mezi lepší a část mezi horší, nebo jednoduše vynechat. V některých zdrojích se ULI také definuje jako rozdíl mezi nejlepší a nejhorší třetinou studentů (viz obr. 8.5a), čímž se prostředním hodnotám často zcela vyhneme. Lze obecně uvažovat i jiné podíly, např. krajní pětiny studentů (viz obr. 8.5b). Systém Rogo počítá ULI na základě dolních a horních 27,5 % studentů. Indexy ULI vycházejí v případě menších porovnávaných skupin zpravidla větší (jedná se pak o rozdíl v úspěšnosti ve dvou více rozdílných skupinách), a proto by pro ně měla být požadována o něco větší hodnota než při klasickém rozdělení na dvě části. I pro tyto definice však platí, že podezřelé jsou hodnoty záporné a hodnoty blízké 0.

Grafy: Další možná pojetí ULI



Obr. 8.5a Index ULI jako rozdíl mezi nejlepší a nejhorší třetinou



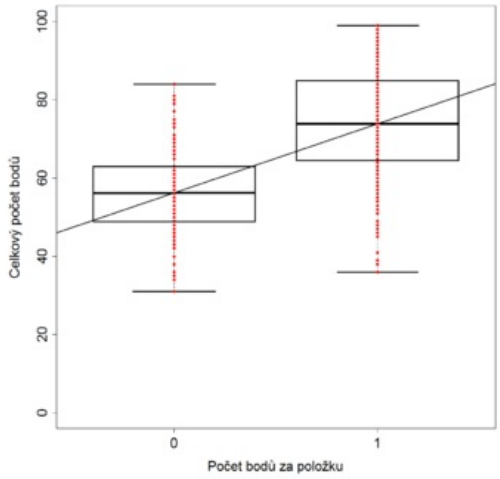
Obr. 8.5b Index ULI jako rozdíl mezi nejlepší a nejhorší pětinou

Jiným přístupem pro hodnocení citlivosti binární položky (tj. položky typu ano/ne) je porovnat průměrný bodový výsledek u skupiny, která položku zodpověděla správně ($\bar{X}_S$), s bodovým výsledkem u skupiny, která položku zodpověděla nesprávně ($\bar{X}_N$). Na dobře fungující položku by měli správně odpovídat vesměs dobří studenti, průměrný výsledek u studentů ze skupiny S by proto měl být vyšší než u studentů ze skupiny N. Významnost **rozdílů průměrných bodových výsledků** $\bar{X}_S - \bar{X}_N$ lze testovat *dvouvýběrovým t testem* [22], ekvivalentně také testem nenulovosti korelačního koeficientu mezi počtem bodů za položku (zde 0 nebo 1) a celkovým počtem bodů.

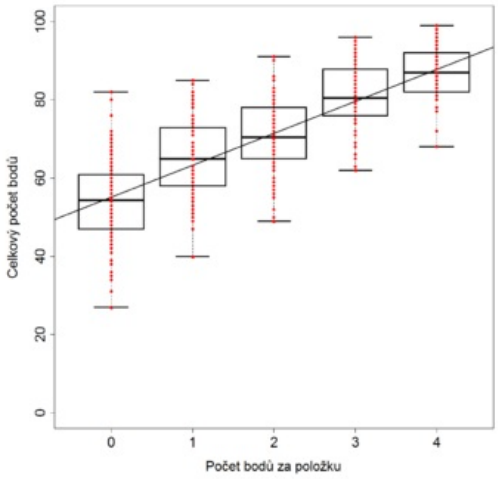
Zobecněním na bodované položky je tzv. **index RIT**. Jde o Pearsonův korelační koeficient (značený často r) mezi položkou (*item*, I) a počtem bodů v celém testu (T). Pro položku, která dobře rozlišuje mezi horšími a lepšími studenty (měřeno počtem bodů v testu) je koeficient RIT blízký jedné. Pro položky, které mezi lepšími a horšími studenty nerozlišují, je tento koeficient blízký nule. Záporné koeficienty pak upozorňují na položky, které zvýhodňují slabší studenty.

Jelikož počet bodů za položku ovlivní celkový počet bodů, hodnoty nejsou nezávislé. Formálně správnější je proto hodnotit korelaci mezi položkou a *součtem bodů za zbylé položky* (angl. *rest*). Tento Pearsonův korelační koeficient je číselně blízký předchozímu a nazývá se **index RIR**.

Graf: Diskriminační index RIT



Obr. 8.6a Diskriminační index RIT pro binární položku. Index vyjadřuje, nakolik jsou body rozpýlené od proložené přímky.



Obr. 8.6b Diskriminační index RIT pro bodovanou položku. Index vyjadřuje, nakolik jsou body rozpýlené od proložené přímky.

Používanou charakteristikou konzistence položek je také **Cronbachovo alfa testu po vynechání položky**: nekonzistentní položky poznáme tak, že po jejich vynechání reliabilita testu měřená Cronbachovým alfa stoupne.

I zde je na místě připomenout, že zavedené indexy citlivosti jsou závislé na testované skupině studentů. Vztaženo k již zmiňovanému příkladu z přijímacích zkoušek, budeme-li analyzovat data všech studentů, kteří se zkoušky zúčastnili, díky menší homogenitě testovaných lze předpokládat, že položky budou studenty rozlišovat lépe. Budeme-li analyzovat pouze data skupiny studentů, kteří byli přijati ke studiu, rozlišovací schopnost bude u této homogennější skupiny patrně nižší. Pro obecnější pojetí opět odkazujeme na kapitolu věnovanou teorii odpovědi na položku, kde je diskriminační schopnost zavedena ve vztahu k celkové schopnosti studenta.

Citlivost jako validita položky

Dosud jsme zkoumali vztah úspěšnosti v položce k celkovému výsledku. Uvědomme si, že tato citlivost odráží, do jaké míry jsou položky vnitřně konzistentní – má tedy co do činění s reliabilitou testu.

Pokud by nás zajímala citlivost položky vzhledem k nějakému jinému kritériu, např. závěrečné známce z předmětu, celkové úspěšnosti ve studiu nebo úspěšnosti v zaměstnání, zajímáme se tak potažmo o validitu dané položky. Položka přitom, stejně jako celý test, může dobře rozlišovat mezi studenty co do celkového počtu bodů v testu, ale nemít žádný vztah ke zkoumanému kritériu (např. budoucí úspěšnosti ve studiu).

Citlivost ve smyslu validity položky můžeme reprezentovat také pomocí indexu ULI, studenty však na horší a lepší dělíme podle zkoumaného kritéria (tedy např. dle průměrné známky v následujícím studiu, úspěšnosti apod.) Lze také použít korelačních koeficientů mezi položkovým skóre a zkoumaným kritériem. Obecněji lze závislost uchopit pomocí regresních modelů, jak bylo zmíněno v části věnované validitě testu.

Validita jednotlivých položek nám pak může pomoci vypátrat, které typy položek způsobují nízkou validitu celého testu.

Analýza vynechaných odpovědí

Pokud vysoké procento studentů na určitou položku neodpoví, může to ukazovat na její obtížnost, ale také **nesrozumitelnost**. Za podezřelé je nutné považovat všechny položky, které vynechalo více než 20 % respondentů. Pokud se na konci testu objevuje souvislá řada nezodpovězených položek, lze předpokládat, že je studenti nestihli vyřešit z **časových důvodů**. V takovém případě je potřeba zvážit zkrácení testu nebo prodloužení doby na řešení.

Počet vynechaných odpovědí souvisí také s hodnocením testu. Vyšší procento vynechaných odpovědí lze předpokládat u testů, které penalizují nesprávnou odpověď (nepenalizují-li vynechání odpovědi ještě více).

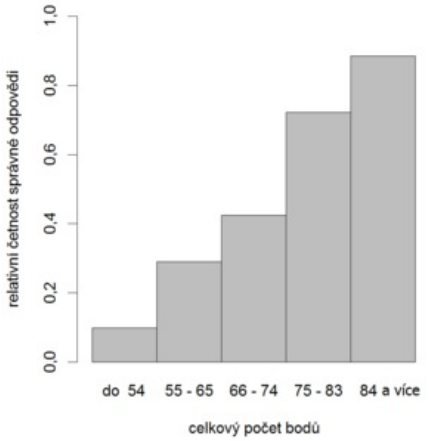
Grafická interpretace vlastností položky

Na vlastnosti položky se můžeme podívat detailněji, rozdělíme-li studenty podle úspěšnosti na několik skupin a úspěšnost v dané položce zobrazíme pro každou skupinu zvlášť. Tato metoda je jakýmsi úvodem k přístupu tzv. *teorie odpovědi na položku* (*item response theory*, IRT), který bude prezentován v následující části.

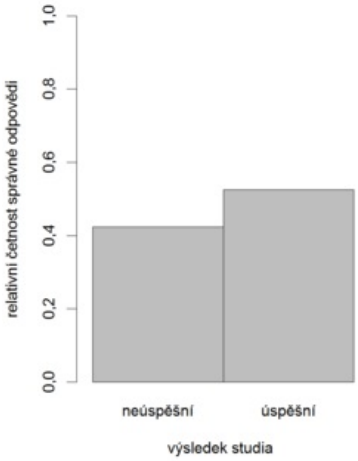
Někdy se stává, že kritérium, podle kterého dělíme studenty na lepší a horší, je jen dvouhodnotové. Stává se to například, když hodnotíme citlivost položky vzhledem k celkové úspěšnosti následného studia (tj. „dokončil studium / nedokončil studium“). Menší počet skupin použijeme také, máme-li malý počet respondentů. V tom případě bývá lépe studenty rozdělit jen na dvě nebo tři skupiny.

U správně zkonstruované položky by relativní četnost správné odpovědi měla růst s celkovým skóre (tj. celkovým počtem bodů), resp. měla by být vyšší pro úspěšné studenty než pro neúspěšné. Tedy čím lépe si studenti vedli v dané položce, tím lépe by si měli vést v celém testu nebo v nastávajícím studiu. Tak je tomu v následujícím příkladu (obr. 8.7a, 8.7b).

Graf: Grafické zobrazení vlastností položky

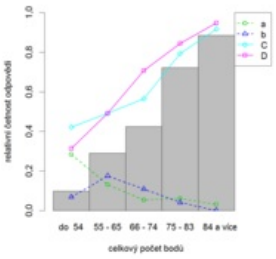


Obr. 8.7a Citlivost dle celkového skóre

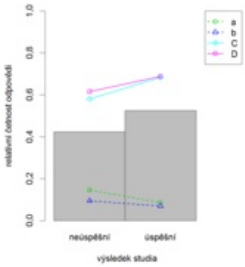


Obr. 8.7b Citlivost dle úspěšnosti

Pro přehlednost jsou v ukázkách správné odpovědi vyznačené velkými písmeny, distraktory jsou označeny písmeny malými. Pokud se jedná o položky s více nabízenými odpověďmi, je na místě vykreslit také relativní četnosti jednotlivých odpovědí (na následujících obrázcích jsou správné odpovědi zobrazeny plnou čarou a distraktory přerušovanou). Četnost volby distraktorů by měla být nižší u lepších studentů, přerušovaná čára by tedy měla mít klesající tendenci. Správné odpovědi by měli častěji volit lepší studenti, proto by čáry popisující správné odpovědi, stejně jako sloupcové grafy popisující celkovou úspěšnost v položce, měly také růst. Tak tomu je i v případě položky CH64 (obr. 8.8a, 8.8b).



Obr. 8.8a Citlivost dle celkového skóre



Obr. 8.8b Citlivost dle úspěšnosti

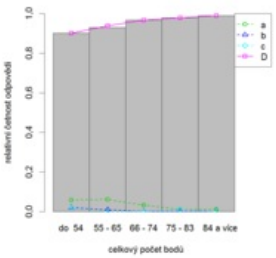
Znění testové otázky CH64
Hydratací kyseliny fumarové vzniká kyselina:
a) maleinová
b) malonová
C) jablečná
D) hydroxyjantarová

Pozn.: V uvedených příkladech jsou správné odpovědi naznačeny velkým písmenem volby odpovědi (zde C, D).

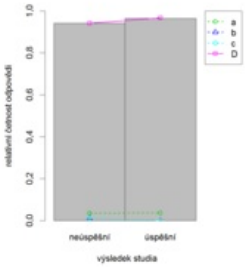
```
Příklad snadné položky
```

Jak je patrné i z grafického znázornění (obr. 8.9a, 8.9b), příliš snadné položky (stejně jako položky příliš těžké) mají nutně malou rozlišovací schopnost. Je to způsobeno skutečností, že na snadnou položku odpoví nesprávně jen velmi malé procento studentů, na velmi obtížnou položku zas odpoví jen velmi malé procento studentů správně.

V případě, že studenti volí pouze jedinou správnou/nejlepší odpověď, bude tato jediná plná čára korespondovat s vrcholky sloupcového grafu. Toto nastává jednak v případě, kdy studenti dopředu vědí, že se jedná o položky s jedinou nejlepší odpovědí (SBA), ale také v případě, kdy ze zadání položky s vícečetnou správnou odpovědí (např. MTF) sami vyvodí, že pouze jediná z nabízených variant může být správná. Tak tomu bylo i u položky CH10.



Obr. 8.9a Citlivost dle celkového skóre



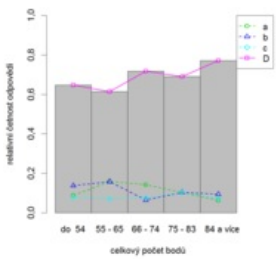
Obr. 8.9b Citlivost dle úspěšnosti

Znění testové otázky CH10
Chemická rovnováha (rovnovážný stav) v reakčním systému je charakterizována:
a) neustále proměnnou koncentrací výchozích látek a produktů
b) neměnnou koncentrací výchozích látek a proměnnou koncentrací produktů
c) proměnnou koncentrací výchozích látek a neměnnou koncentrací produktů
D) neměnnou koncentrací výchozích látek a produktů

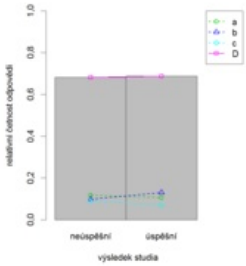
```
Příklad špatně rozlišující položky
```

Položka CH97 má také zřejmě jedinou správnou odpověď. Přestože její obtížnost je vyšší než u předchozí položky, i tato položka velice špatně rozlišuje mezi lepšími a slabšími studenty. Špatná rozlišovací schopnost této položky může být způsobena tím, že k dosažení správné odpovědi musí studenti provést správně hned několik úkonů a využít několika

nezávislých znalostí a dovedností (znalost vzorce oxidu sírového, znalost atomové hmotnosti kyslíku, sestavení správné stechiometrické rovnice, správný výpočet), přičemž v každé z nich je možné udělat chybu.



Obr. 8.10a Citlivost dle celkového skóre

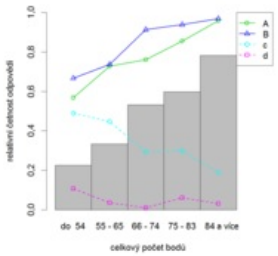


Obr. 8.10b Citlivost dle úspěšnosti

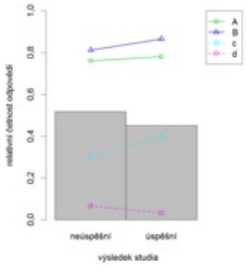
Znění testové otázky CH97
Vypočtěte hmotnost síry teoreticky potřebné k výrobě 16 tun oxidu sírového (A_r S =32):
a) 12,8 t
b) 8 t
c) 3,2 t
D) 6,4 t

```
Příklad málo validní položky
```

Položka CH55 má dvě správné odpovědi a dva distraktory, z nichž distraktor d je velice málo atraktivní. Stejně jako položka CH64 i CH55 velice dobře rozlišuje mezi studenty v testu úspěšnými a neúspěšnými, což je částečně dáno i její optimální obtížností. CH55 ale není příliš vhodná pro předpověď úspěšnosti ve studiu: neúspěšní studenti ji řešili dokonce o něco lépe než studenti úspěšní. Pro zvýšení validity testu bychom tuto položku měli z testu odebrat.



Obr. 8.11a Citlivost dle celkového skóre

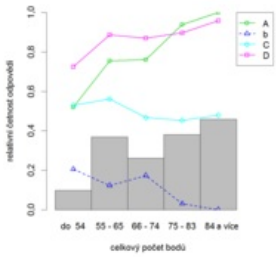


Obr. 8.11b Citlivost dle úspěšnosti

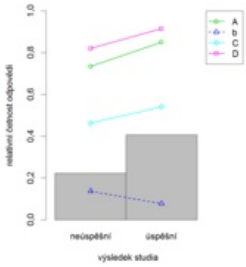
Znění testové otázky CH55
Pod pojmem racemát rozumíme:
A) směs enantiomerů v poměru 1 : 1
B) směs optických antipodů v poměru 1 : 1
c) směs pravotočivých a levotočivých látek v poměru 1 : 1
d) směs epimerů v poměru 1 : 1

```
Příklad validní položky málo citlivé k celkovému výsledku v testu
```

CH28 je těžká položka, obzvláště s ohledem na těžko odhalitelnou správnou odpověď C. Dva správné názvy chemické sloučeniny respondenti patrně nečekali. Položka ovšem poměrně dobře rozlišuje mezi studenty úspěšnými a neúspěšnými ve studiu.



Obr. 8.12a Citlivost dle celkového skóre

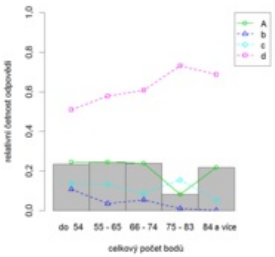


Obr. 8.12b Citlivost dle úspěšnosti

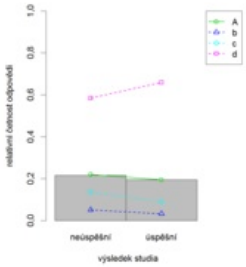
Znění testové otázky CH28
Vyberte správné tvrzení. CS_2 je:
A) nepolární rozpouštědlo
b) polární rozpouštědlo
C) sulfid uhličitý
D) sirouhlík

```
Příklad chybné položky
```

Obzvláště těžká je také položka CH44, kde distraktor d volí více studentů, než správnou odpověď A. Tu, zdá se, respondenti spíše hádali. Distraktor d dokonce volí studenti tím častěji, čím vyššího skóre v testu dosáhli. To budí podezření, že by položka mohla být chybná. Vskutku: správná odpověď je přibližně 0,008, v nabídnutých možnostech tedy chybí, avšak možnost d (autory testu označená jako nesprávná) se jí nejvíce blíží.



Obr. 8.13a Citlivost dle celkového skóre

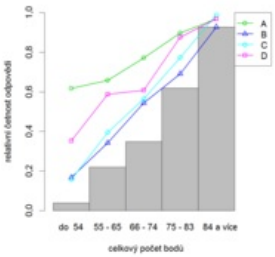


Obr. 8.13b Citlivost dle úspěšnosti

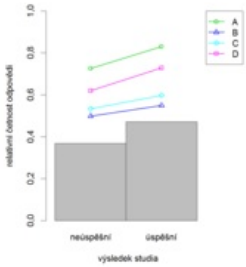
Znění testové otázky CH44
Kyselinu dusičnou o w = 5 % a o hmotnosti 10 g zředíme vodou o hmotnosti 55 g. Jaký je hmotnostní zlomek HNO_3 ve výsledném roztoku:
A) 0,04
b) 0,5
c) 0,55
d) 0,01

```
Příklad správně fungující položky
```

Výtečná diskriminační schopnost položky CH89 souvisí kromě optimální obtížnosti patrně také s faktem, že všechny nabízené odpovědi jsou správné. Zaškrtnout všechny nabízené odpovědi chce jistou dávku znalosti nebo odvahy.



Obr. 8.14a Citlivost dle celkového skóre



Obr. 8.14b Citlivost dle úspěšnosti

Znění testové otázky CH89
K hydrofilním vlastnostem bílkovin přispívá:
A) hydroxyskupina
B) sulfhydrylová skupina
C) guanidylová skupina
D) aminoskupina

Souhrnně jsou psychometrické vlastnosti zmíněných položek a příslušné závěry uvedeny v tabulce 8.3. Položky jsou v ní řazeny dle obtížnosti od nejsnazší úlohy. Shrneme-li doporučení pro položkovou analýzu, postup je následující: V první řadě je potřeba odhadnout obtížnost položky, u příliš snadných ani příliš obtížných položek totiž nelze čekat rozumnou citlivost. U položek s více nabízenými odpověďmi je podobně potřeba zkontrolovat také relativní četnost volby jednotlivých distraktorů. O něco důležitější než obtížnost položek je pak jejich citlivost, tedy schopnost rozlišit mezi horšími a lepšími studenty (počítáno dle celkového výsledku v testu nebo dle jiného vnějšího kritéria). Nakonec se je potřeba podívat také na procento vynechaných odpovědí a zvážit, zda souvisí s obtížností položky nebo spíše s jejím pořadím v testu. Pokud je velké množství vynechaných odpovědí na konci testu, a zvláště pokud jejich obtížnost není příliš velká, je potřeba zvážit zkrácení testu.

Tab. 8.3 Příklady přijímacího testu z chemie

ID	obtížnost	Volba jednotlivých alternativ v %				citlivost podle skóre	citlivost podle úspěšnosti	% vynechaných	závěr slovy
		A	B	C	D				
CH10	**0,95**	**4**	**1**	**1**	95*	0,10	0,02	0,0	snadná položka, neatraktivní distraktory
CH97	0,68	11	11	8	68*	**0,13**	0,00	2,4	špatně rozlišující položka
CH55	0,49	77*	84*	35	**5**	0,49	**-0,07**	0,2	nevalidní položka, neatraktivní distraktor D
CH64	0,47	12	8	63*	65*	0,68	0,10	0,4	správná položka
CH89	0,42	78*	52*	56*	67*	0,74	0,10	0,2	správná položka
CH44	0,21	21*	**4**	11	**62**	**-0,06**	-0,02	1,2	podezřelá položka, distraktor D volen častěji než správná odpověď

Pozn.: správná odpověď je označena *

Pokusme se nyní shrnout závěry, které lze z této položkové analýzy vyvodit:

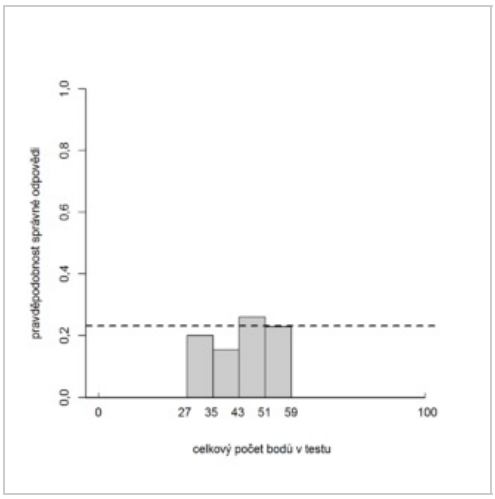
- Žádná z analyzovaných položek nebyla vynechána více než 5 % respondentů. Z analyzovaných položek byla nejčastěji vynechána položka CH97 (2,7 % respondentů). Její umístění na konci testu nás může vést k zamyšlení, zda není časový interval příliš přísný, procento vynechaných odpovědí je ale malé.
- Položka CH10 je velice snadná, její distraktory jsou velmi neatraktivní a správnou odpověď volilo až 95 % studentů. Položka proto také špatně rozlišuje mezi lepšími a slabšími studenty, má téměř nulovou diskriminační schopnost.
- Neatraktivní jsou distraktory v položce CH97. Velice neatraktivní je také distraktor b v položce CH44. Naproti tomu

distraktor d volilo v položce CH44 dokonce více respondentů než správnou odpověď A. Jak je vidět z odhadů citlivosti, položku řešili lepší studenti hůř než studenti slabší (kteří mohli správné řešení hádat). Tato položka je z tohoto důvodu velice podezřelá a jak by vyplynulo z revize textu, také nesprávně zadaná.
- Kromě zmíněné snadné položky CH10 a nesprávně zadané položky CH44, velice špatně mezi lepšími a slabšími studenty rozlišuje také položka CH97.
- Položka CH55 patří mezi položky, které mají uspokojivou citlivost vůči celkovému skóre, řešili ji ale lépe studenti ve studiu neúspěšní než studenti ve studiu úspěšní. Vynecháním této položky bychom mohli přispět ke zvýšení predikční validity celého testu.
- Položky CH64 a CH89 jsou příklady správně fungujících položek.

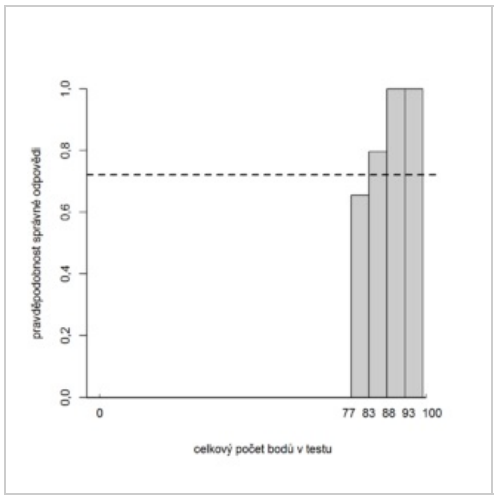
Shrneme-li, klasická položková analýza nám může pomoci rychle odhalit slabiny jednotlivých položek a nabízených odpovědí. Hlavní potíží klasických odhadů obtížnosti a citlivosti je jejich závislost na rozdělení znalosti u testovaných studentů. Zadáme-li položku několika skupinám studentů s různou úrovní znalosti, výsledný odhad obtížnosti (relativní četnost správné odpovědi) se může ve skupinách značně lišit. Podobně pokud odhadujeme citlivost položky u několika skupin s různou mírou homogenity, klasický odhad citlivosti se může ve skupinách značně lišit. S tímto nedostatkem se částečně vypořádává grafické hodnocení vlastností položky, kdy zobrazíme relativní četnost správných odpovědí *v závislosti na celkovém počtu bodů* – tj. v závislosti na jakési míře znalostí studenta. Při grafickém zobrazení již můžeme vnímat *obtížnost* jako relativní četnost správné odpovědi u skupiny s „průměrnou“ znalostí a *citlivost* jako *sklon* sloupcového grafu. To jsou hlavní myšlenky teorie odpovědi na položku, která nám takto obecněji pomůže vlastnosti položky také vyjádřit číselně.

Teorie odpovědi na položku

Klasické odhady obtížnosti, citlivosti, ale i spolehlivosti jsou silně závislé na rozdělení znalostí u testované skupiny. Představme si dvě skupiny studentů – horší a lepší. Použijeme-li klasického odhadu obtížnosti, ve skupině lepších studentů se položka bude jevit jako snadná a naopak ve skupině slabších studentů se bude jevit jako obtížná (viz obr. 8.15a, 8.15b).



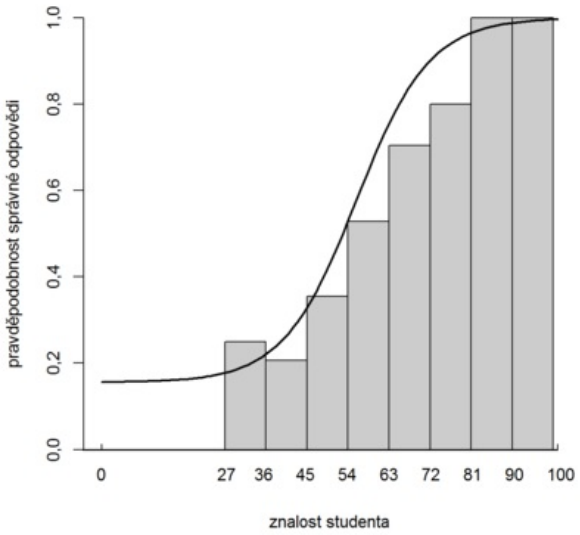
Obr. 8.15a Klasický odhad obtížnosti položky pro horší skupinu



Obr. 8.15b Klasický odhad obtížnosti položky pro lepší skupinu

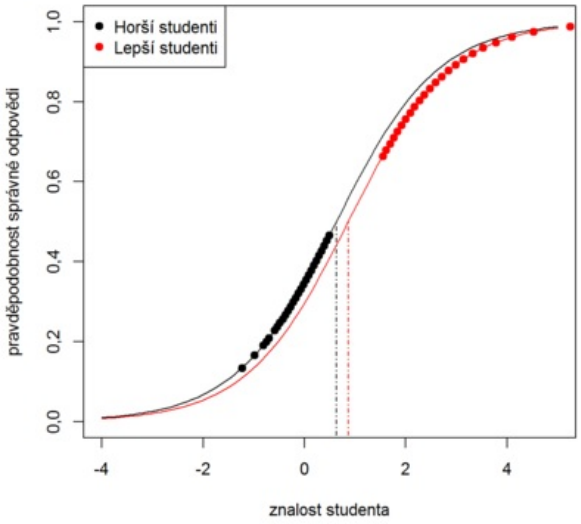
To může být problém, pokud používáme položku pro různé skupiny studentů, např. sdílíme-li položky s dalšími fakultami. Proto se v případě většího počtu studentů nebo skupin studentů používá modernější **teorie odpovědi na položku** (angl. *item response theory*, **IRT**), která modeluje pravděpodobnost správného zodpovězení položky podmíněně pro různé úrovně znalosti studenta.

Zjednodušeně řečeno, vztah relativní četnosti správné odpovědi na položku a celkového počtu bodů, který jsme v předchozí kapitole ilustrovali sloupcovým grafem, v rámci teorie odpovědi na položku modelujeme spojitou funkcí zvanou **charakteristická funkce položky** (angl. *item characteristic function*, ICF).



Obr. 8.16 Aproximace pravděpodobnosti správné odpovědi v položce pomocí IRT modelu

Pokud tedy budeme odhadovat obtížnost položky z odpovědí u skupiny slabších studentů, IRT odhad vezme v potaz úroveň znalosti studentů a odhaduje spodní část křivky. Pokud získáme výsledky od skupiny lepších studentů, je odhadována horní část křivky. Výsledné křivky (potažmo parametry obtížnosti) odhadnuté dle dvou skupin studentů si nakonec mohou být velice blízké a téměř totožné s křivkou, kterou bychom dostali, pokud bychom odhadovali ze všech dat najednou (viz obr. 8.17).



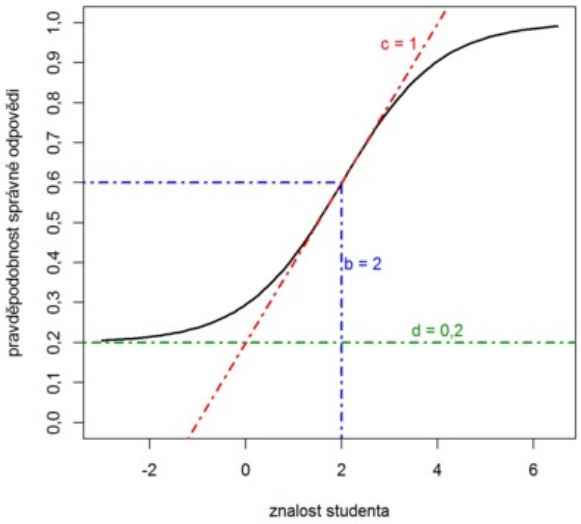
Obr. 8.17 Odhad charakteristické křivky a IRT odhad parametru obtížnosti pro 2 skupiny

Tip: Klasické odhady obtížnosti a citlivosti položky neberou v potaz rozdělení znalostí v testované skupině. Jsou-li položky používány pro velké množství studentů různých znalostních úrovní, je vhodné klasické odhady nahradit IRT odhady.

Odhady vlastností položek pomocí IRT modelů

Nyní nastává otázka, jak nalézt funkci neboli charakteristickou křivku, která popisuje pravděpodobnost správné odpovědi na položku nejlépe. V praxi se většinou předpokládá konkrétní tvar funkce a z dat se odhadují pouze její parametry (vlastnosti položek). Některé nejčastěji používané modely jsou blíže popsány v příloze. Typicky se předpokládá, že charakteristická funkce je esovitého tvaru. Polohu jejího středu na ose x charakterizuje *parametr obtížnosti položky* b. Parametr obtížnosti říká, jakou úroveň znalosti potřebuje student, aby položku zodpověděl právě s poloviční pravděpodobností. Sklon křivky charakterizuje další možný parametr – *parametr citlivosti* a. Ten je intuitivním rozšířením dříve popsaného koeficientu ULI. V případě položek s výběrem z n odpovědí má smysl předpokládat, že student s pravděpodobností $1/n$ správnou odpověď uhádne. Lze pak uvažovat také další parametr – *parametr uhádnutelnosti* d.

Souhrnně je interpretace zmíněných parametrů zobrazena na obrázku 8.18:



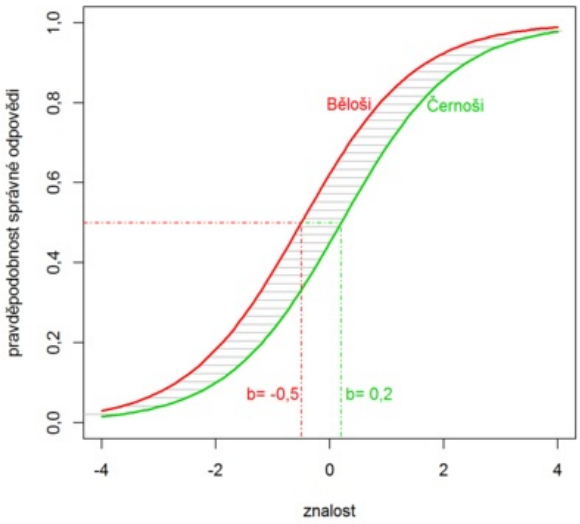
Obr. 8.18 Interpretace parametrů IRT modelu: b obtížnost, c citlivost (diskriminační schopnost), d uhádnutelnost položky

Jak tyto parametry a úrovně znalostí jednotlivých studentů z dat odhadovat? Nevýhodou IRT je, že odhady není možné spočítat tak jednoduše, jako klasické odhady obtížnosti a citlivosti. K odhadu parametrů IRT modelů je zapotřebí použít příslušný statistický software, neboť jde o složitější optimalizační procedury založené na maximalizaci tzv. věrohodnostní funkce. Známe-li parametry položek (např. z předchozí pilotní studie), odhadujeme parametry znalostí studentů metodou *podmíněné maximální věrohodnosti*. V jednoparametrickém IRT modelu (modelu s jediným parametrem – parametrem obtížnosti) tak ke každému celkovému počtu bodů jednoznačně přiřadíme hodnotu znalosti. Ve složitějších modelech již kromě celkového skóre může záviset také na tom, které položky student zodpověděl správně. Známe-li parametry znalosti jednotlivých studentů, metodou *podmíněné maximální věrohodnosti* můžeme odhadnout parametry položek. Pokud parametry znalostí jednotlivých studentů neznáme, ale známe alespoň rozdělení těchto znalostí (většinou předpokládáme normální rozdělení s příslušnou střední hodnotou a rozptylem), můžeme parametry položky odhadnout metodou tzv. *marginální maximální věrohodnosti*. Úroveň znalosti studenta a parametr obtížnosti položky lze odhadovat také společně pomocí tzv. *sdružené maximální věrohodnosti*. Ke všem zmíněným odhadům parametrů je zapotřebí dostatečně velkého vzorku – minimálně stovky, ještě lépe tisíce studentů.

> 💡 **Tip: Který z odhadů vlastností položek použít?**

Využití IRT modelů pro hodnocení férovosti položky

IRT analýza je užitečná např. k vyhodnocení **férovosti položky**, která byla diskutována v části Revize férovosti v rámci Oponentury testu. Zatímco v rámci oponentury testu se snažíme neférovosti položek předejít, na základě odpovědí studentů můžeme případnou neférovost položky odhalit dodatečně. Snažíme se zjistit, zda položka nezvýhodňuje některé skupiny studentů oproti jiným (např. dle pohlaví, etnika, apod.). Prvním krokem může být zjištění relativních četností správných odpovědí na položku ve skupinách (např. u bělochů a u černochů, nebo u mužů a u žen). Rozdílné procento správných odpovědí může upozornit na možnou neférovost, ale může být také zcela matoucí, neboť procento správných odpovědí závisí na úrovni znalostí daných dvou skupin. Zajímá nás proto, zda se relativní četnosti správných odpovědí ve skupinách liší **za podmínky stejných znalostí** studentů. S výhodou proto můžeme použít IRT odhadů dvou charakteristických křivek položky pro dané skupiny (viz obr. 8.19). Odhadneme charakteristické IRT křivky zvlášť pro každou zkoumanou skupinu a porovnáme je. Jako *index rozdílného fungování položky* (*different item functioning*, DIF) se pak bere např. plocha mezi dvěma křivkami, nebo také rozdíl v odhadu parametru obtížnosti ve dvou IRT modelech. Významnost rozdílného fungování položky lze také testovat statistickým testem a jednoznačně tak vytipovat problematické položky.

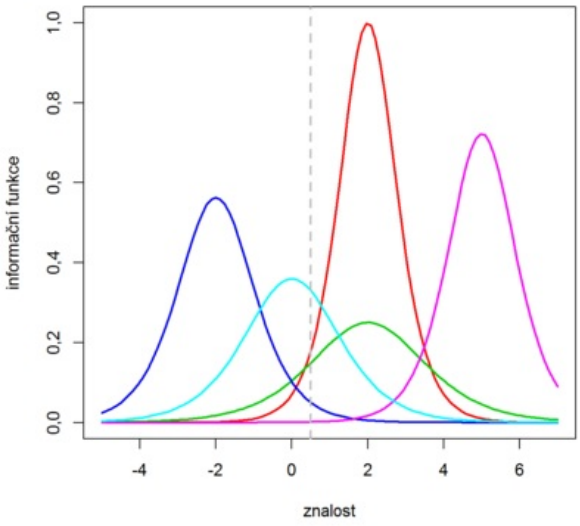


Obr. 8.19 Odhad parametru obtížnosti pro dvě skupiny. Plocha mezi křivkami odpovídá „koeficientu neférovosti“ položky (DIF)

Informační funkce položky

Spolehlivost testu se v klasickém pojetí vyjadřuje jediným indexem. Teorie odpovědi na položku (IRT) naproti tomu poukazuje na skutečnost, že informační přínos jednotlivých položek, potažmo pak celého testu, je závislý na úrovni vědomostí studenta. Informační přínos položky je v IRT popsán tzv. **informační funkcí položky**, která je zvonovitého tvaru a která závisí na parametrech konkrétní položky. Informaci celého testu pak lze vyjádřit jako součet informačních funkcí jednotlivých položek, neboť předpokládáme, že odpovědi na položky jsou na sobě (při dané úrovni znalosti studenta) nezávislé.

Informační přínos položky se odvíjí od její obtížnosti a od její diskriminační schopnosti. Vysoce rozlišující položky mají vysokou a úzkou informační křivku, která významně přispívá informačně, avšak pouze v úzkém rozsahu obtížnosti. Položky s nízkou rozlišovací hodnotou poskytují méně informací, ale v širším rozsahu obtížností. Velikost informace, kterou položka poskytuje při dané úrovni schopnosti studenta může být vhodným kritériem pro rozhodování, zda položku ponechat v testu nebo ji z něj vyřadit. Máme-li přibližný odhad studentových znalostí, mohou pro nás informační funkce položek být vodítkem, kterou položku studentovi zadat (viz obr. 8.20).



Obr. 8.20 **Informační funkce položek s různými parametry obtížnosti a citlivosti**

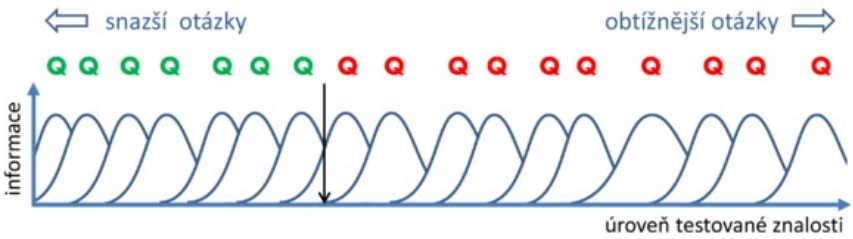
Adaptivní testování a použití IRT v praxi

Informační funkce položky nám může dát vhodné kritérium pro rozhodování, kterou položku použít pro testování studenta s danou úrovní znalostí. Z toho vychází tzv. adaptivním testování. **Adaptivní testování** (*computerized adaptive testing*, CAT) je metoda testování, při níž se **výběr testovacích úloh přizpůsobuje schopnostem testované osoby**. Protože test je tak testovanému doslova „ušitý na míru“, říká se této metodě též *tailored testing*.

Hlavní výhodou adaptivního testování je, že oproti „klasickému“ testu stačí položit podstatně menší počet otázek, a přitom získáme stejně přesný odhad znalosti studenta [73]. Nevýhodou naopak je, že studenti nedostávají stejné otázky a nelze je tedy přímo porovnávat. Z toho mimo jiné vyplývá, že předpokladem zavedení adaptivního testování je vytvoření banky testových úloh (BTÚ) a její naplnění kalibrovanými položkami.

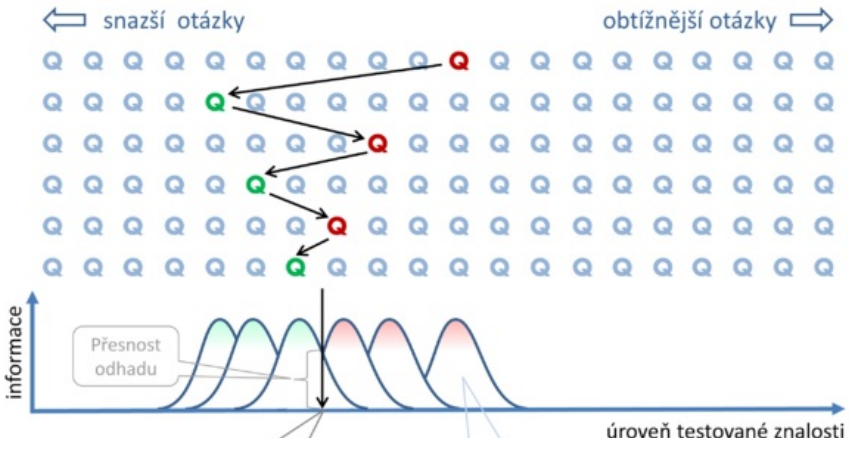
KLASICKÉ TESTOVÁNÍ

Při klasickém testování je třeba otázkami pokrýt celý rozsah možné úrovně znalostí



ADAPTIVNÍ TESTOVÁNÍ

Pro stejně přesný odhad úrovně znalosti testovaného stačí položit menší počet otázek



Obr. 8.21 **Schéma porovnávající klasické a adaptivní testování** pomocí informačních funkcí položek. Položky (otázky) poskytují většinou dobré rozlišení studijního výkonu jen v části rozsahu testovaných znalostí (tam, kde má informační křivka dostatečně vysoké hodnoty). Pokud předem neznáme úroveň znalostí testovaného, musíme mu proto při **klasickém testování** položit otázky pokrývající celý rozsah úrovní znalostí.

V adaptivním testování je úroveň znalosti postupně odhadována v několika krocích a vystačíme proto s podstatně menším počtem otázek pro získání odhadu znalosti s potřebnou úrovní přesnosti.

Poznámka k obrázku: Informační funkce položky jsou zde pro názornost idealizovány, ve skutečnosti by se více překrývaly a byly by z hlediska obtížnosti méně specifické.

Vzhledem k tomu, že testovaní mohou mít velmi odlišné studijní výkony, není možné pro výběr položek vystačit s jejich klasickými popisy (obtížnost, citlivost, ...) a je třeba použít odhady vlastností položek založené na teorii odpovědi na položku (IRT).

Princip adaptivního testování spočívá v tom, že znalost studenta se na počátku odhadne, poté se postupně kladením dalších testových otázek (položek) zpřesňuje až do chvíle, kdy se podaří dosáhnout určité (předem stanovené) přesnosti odhadu úrovně znalosti.

Jak bylo naznačeno výše, předpokladem pro nasazení adaptivního testování je dostupnost dostatečně velké banky kalibrovaných položek (několik tisíc položek [74] se známými psychometrickými vlastnostmi). Například Breithauptová et al. [75] odhaduje, že na relativně malý závěrečný test se čtyřiceti položkami je třeba mít banku se dvěma tisíci kalibrovaných položek. Adaptivní testování tedy umožňuje na jednu stranu ušetřit čas (a náklady) při samotném testování, ale vyžaduje velkou počáteční investici na pořízení a naplnění položkové banky.

U položek je přitom zapotřebí nejen to, aby byly známé jejich psychometrické vlastnosti, ale také, aby jejich celková obtížnost byla „jednorozměrná“, tedy aby položky netestovaly (byť skrytě) více oblastí. Pokud například budeme testovat v chemii znalost poločasu rozpadu na početním příkladu, pak studenti s lepší znalostí matematiky budou otázku vnímat jako snazší, než jejich kolegové hůře vybavení touto dovedností. K podobnému jevu může vést použití cizojazyčných nebo latinských výrazů. Použitím úloh, jejichž obtížnost je složená z více „nezávislých vektorů“, autor bezděčně ztrácí kontrolu nad obtížností, neboť pro různě disponované jedince budou mít rozdílnou celkovou obtížnost. Použití takovýchto „vícerozměrných“ položek v testu může být problematické.

Adaptivní testování umožňuje nezatěžovat studenta nadbytečným množstvím otázek, které by pro něj byly buď příliš jednoduché, nebo příliš obtížné. Zodpovídá jen otázky, které jsou pro něj přiměřené, stráví testováním kratší dobu a snižuje se tak jeho celková testová frustrace.

Tab. 8.4 Výhody a nevýhody adaptivního testování

Výhody adaptivního testování	Nevýhody adaptivního testování
▪ zkrácení testu ▪ snížení testové frustrace pro většinu respondentů ▪ bezprostřední zpětná vazba ▪ snížení nákladů ▪ možnost jednoduchého a automatizovaného odstranění nefunkčních položek z položkové banky	▪ nutnost rozsáhlé položkové banky ▪ nutnost kalibrace položek ▪ nemožnost přímo porovnat dva testované ▪ nemožnost vrátit se k předchozí položce ▪ nutnost formulovat položky, aby byly „jednorozměrné“

Další informace o adaptivním testování najde zájemce v knize Testování v psychologii [40], ve studii Filípkové a Byčkovského [73] nebo v článku [76].

Nástroje pro adaptivní testování

Pro adaptivní testování existuje řada softwarových nástrojů, např. komerční FastTEST, který je dostupný jak ve verzi pro PC, tak jako webová aplikace. Z nekomerčních řešení stojí za zmínku např. otevřená testovací platforma Concerto (http://www.psychometrics.cam.ac.uk/page/338/concerto-testing-platform.htm) [77] s modulem pro adaptivní testování catR, který byl vytvořen v prostředí R [78].

Mnoho zajímavých informací, přehled používaného software a příklady významných aplikací adaptivního testování najde zájemce na stránkách **Mezinárodní asociace pro počítačové adaptivní testování IACAT (http://iacat.org/).**

V domácím kontextu byl v Psychologickém ústavu AV ČR v Brně (2005) vytvořen vlastní původní **software pro adaptivní testování CATO** (http://www.psu.cas.cz/index.php?option=com_content&view=article&id=59&Itemid=70) [79].

Na Masarykově univerzitě (v Centru jazykového vzdělávání) byl v rámci evropského projektu vytvořen systém **Adaptivní test COMPACT**, který slouží každý semestr tisícům studentů jako prerekvizita k výuce akademické angličtiny. Po dohodě s odpovědným pracovištěm mohou software využívat učitelé i jiných fakult v rámci MU a plnit jej testovými položkami své odbornosti. Zaškolení pedagogů je snadné, k praktickému využívání není třeba žádných zvláštních znalostí. Základní informace poskytuje manuál v publikaci [80].

Realizace testů

Moderní testování zhodnocuje současné technologie a obsahuje prvky inženýrského přístupu. Nahlíží se na něj jako na **integrovaný systém systémů** [81]. Myslí se tím, že na systém pro tvorbu položek může navazovat systém pro jejich pilotování, poté systém pro předkládání testů a tak dále.

Podle míry elektronizace testů rozlišujeme:

- **počítačové testování** (*computer based assessment*, CBA), v němž se celý cyklus hodnocení odehrává v elektronické podobě,
- **počítačem podporované testování** (*computer assisted assessment*, CAA), které zahrnuje i kroky prováděné „ručně“, tedy např. vytvoření a zpracování papírové formy testů [82].

Pro počítačem podporované testování budeme v tomto textu používat názornější, i když ne zcela přesný termín *papírové testování*.

Zabezpečení testů

Dobré zabezpečení testů a testových otázek tvoří základní podmínku jejich věrohodného použití. Při neformálním průzkumu provedeném v roce 2007 mezi 30 000 americkými vysokoškoláky [83] se ukázalo, že 60,8 % jich během studia podvádělo. Tentýž průzkum ukázal, že 16,5 % z nich to nepociťuje jako etické provinění.

Poněkud 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, zda významnou roli při formování těchto postojů nehraje dnešní kultura, zaměřená na úspěch a nehledící na druhé. V sázce je velké množství stipendií, ocenění, stáží a dalších pobídek. Je proto možné, že ti, kdo dosahují úspěchu nepoctivě, odůvodňují své chování právě odměňováním za „úspěch“.

Mají-li být výsledky testování věrohodným obrazem znalostí studentů, musejí být vyloučeny vlivy, které by výsledky mohly zkreslit.

Preventivní a metodická opatření

Součástí standardizace jsou preventivní a metodická opatření, která směřují mimo jiné k zajištění rovných podmínek pro všechny testované.

Před testem

V období před testem musí být zajištěn rovný přístup kandidátů ke znění testových otázek. Jsou dvě možnosti, jak riziko nerovného přístupu k otázkám snížit. Jednou je omezit exponování testu, tj. omezit příležitosti, při kterých může dojít k zaznamenání a přenesení testových položek do populace budoucích účastníků testu. Při každém použití testu toto riziko hrozí, i když ne vždy je naplněno. U testů s významným dopadem pro studenty je vhodné zveřejnit jen modelové otázky a položky použité v předchozích kolech testu považovat za pravděpodobně vynesené. Přístup k ostré verzi testu je třeba omezit na co nejužší okruh osob i v rámci pedagogického sboru, aby se snížilo riziko úniku položek ať již nedopatřením, či prolomením etických a profesionálních standardů.

Jiným způsobem regulace zmíněného rizika je navýšení počtu připravených otázek nad míru naučitelnou zpaměti. Je-li v položkové bance několik tisíc položek, je ještě možné se odpovědi mechanicky naučit; pokud je však položek více, je již jednodušší se naučit látku samotnou. Regulace rizik se hojně diskutuje i v odborné literatuře [84].

Během testu

V době testu je třeba zajistit dozor, který vyloučí používání nepovolených pomůcek, opisování a napovídání. Některé školy v zahraničí zavedly režimová opatření týkající se odchodů studentů na toaletu v průběhu testování, neboť se ukázalo, že odchod zvyšuje riziko použití nedovolených pomůcek a napovídání. Při větším množství studentů je nezbytné zajistit i objektivní identifikaci účastníků testu, aby test nemohl složit někdo jiný.

Po testu

Vyplněné testy je třeba chránit před dodatečnými zásahy a zajistit jejich objektivní vyhodnocení. Jednou z cest ochrany je oddělení identity testovaného od zpracovávaného testu a jejich opětovné spojení až po kompletním zpracování a vyhodnocení testů.

Ani sebelepší preventivní opatření však nemohou podvádění zcela zamezit. Proto vznikly metody detekce podvádění na základě rozboru výsledků testu [85]. Krátce jsme se o nich zmínili již v kapitole 8 věnované analýze testů; některým se budeme věnovat v následujícím textu.

Následná detekce incidentů

Statistické vyhodnocení výsledků testů může pomoci identifikovat degradované položky, jejichž přínos pro objektivní a spravedlivé hodnocení byl narušen vynesením, popřípadě identifikovat účastníky testu, jejichž výsledky vykazují podezřelé znaky. Řada prací se zabývá přímo metodikou statistické detekce prozrazených otázek [86], [87], [88]. Přehled problematiky vyzrazených otázek ve vztahu k adaptivnímu testování podává souborná práce mapující aktivity na tomto poli v letech 1983 až 2005 [89] .

Příkladem řešení problémů s testováním, které již proběhlo, byl zásah firmy Caveon specializované na bezpečnost testování. V lednu 2007 tato společnost řešila problém americké Federace státních

komisí pro fyzikální terapii (FSBPT). Při detašovaných zkouškách fyzioterapeutů v Manile měla část účastníků testů k dispozici otázky, které zachytili účastníci předchozích testů; tyto otázky pak byly hromadně distribuované. 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 potvrdit podezřelé výsledky testů by znamenalo ohrozit integritu celého testování a dobré jméno FSBPT.

V této situaci dostala společnost Caveon kompletní testová data ze všech testovacích míst za poslední dva roky k forenzní analýze. Důvodem byl předpoklad, že by zvýhodnění účastníci měli na otázky odpovídat neobvyklým způsobem. K identifikaci odlišně vyplněných testů použila společnost tři nezávislé statistické ukazatele, jejichž kombinací bylo možné dosáhnout pravděpodobnosti správné detekce s rizikem chyby menším než 1:1 000 000. Z prověřovaných 23 000 testů tak byla odhalena dvacítka, která měla všechny tři sledované ukazatele odchylné od normálu. Na základě toho byly zmíněné testy prohlášeny za neplatné a ostatní uznány jako platné [90].

Programy pro počítačové testování

Tip: Vytvořte elektronickou verzi svého testu během tří minut!

Testy je možné připravit, distribuovat i vyhodnotit „ručně“. Výhodnější však může být použití některého z řady připravených počítačových programů, které poskytují podporu v jednom, několika, nebo výjimečně i všech krocích testového cyklu.

Testové moduly v systémech pro podporu výuky

Systémy pro podporu výuky (*learning management system*, LMS) se používají na většině vysokých škol. Nejrozšířenější z nich je Moodle (http://moodle.org/), o němž pojednává samostatný odstavec. Pro formativní testování lze použít i jednodušší, leč komerční Adobe Connect (http://www.adobe.com/products/adobeconnect.html), jehož licenci některé fakulty rovněž mají.

Dříve byl v této kategorii velmi populární systém pro podporu výuky WebCT [92]. Jako jeden z prvních webových nástrojů pro řízení výuky se objevil v roce 1996. Vyráběla a šířila jej společnost WebCT, Inc. a v „nejlepších letech“ jej používalo více než 10 milionů studentů v 80 zemích. V roce 2005 byla společnost pohlcena svým největším konkurentem Blackboard Inc. Program byl velmi komplexní – dokonce až do té míry, že bylo poměrně obtížné jej přizpůsobovat novým požadavkům. Postupně tak získal reputaci až příliš složitého produktu, závislého na Javě a množství nově otevíraných (pop-up) oken. V Čechách se WebCT testovalo na ČVUT a od roku 2001 se používá na Univerzitě Hradec Králové [93].

Komplexní testové programy

Z řady testových programů nabízejí nejvíc možností a pokročilých funkcí specializované produkty, které podporují celý proces přípravy a distribuce testů. Patří mezi ně např. Rogo (pojednané dále), Questionmark (https://www.questionmark.com/us/Pages/default.aspx) nebo komplexní webová aplikace FastTest Web (http://www.fasttestweb.com/).

Proprietární programy pro podporu testování

Řada programů řeší jen některou část testové agendy. Pro základní on-line testování se hodí QuizStar (http://quizstar.4teachers.org) nebo Quia Web (http://www.quia.com/web/), pro přípravu papírových testů EasyTestMaker (http://www.easytestmaker.com).

Celkově lze říci, že pro formativní testování existuje řada programů, které pomáhají testy vytvářet a doručovat. To však neplatí pro sumativní hodnocení, zvláště pak při zkouškách, které rozhodují o dalším postupu studenta. V takovém případě potřebujeme produkty, které spadají do zcela jiné kategorie, neboť je nutné zajistit odpovídající bezpečnost, spolehlivost, výkon a důvěryhodnost.

Pokud navíc hledáme nekomerční program, zůstane jen několik možností. Dva takové testové programy představíme v následujícím bloku.

Moodle

Nejznámějším a světově asi nejrozšířenějším programem, v němž lze mimo jiné i elektronicky testovat, je **Moodle**. Jedná se o komplexní systém výuky přes internet, který obsahuje i pokročilý testovací modul. Moodle je volně šiřitelný software s otevřeným kódem. Vznikl v roce 2002 a průběžně se aktualizuje. Se svým otevřeným kódem, volným šířením a komunitní podporou představuje častou volbu vysokých škol a univerzit. K dispozici je i na všech lékařských fakultách v České republice.

Výhody

Na neustálém vývoji Moodle spolupracuje rozsáhlá a aktivní komunita. Služby, které přesahují možnosti jednotlivých správců, poskytují 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. Pro tento systém existuje více než 600 ověřených rozšiřujících modulů a díky otevřenosti kódu lze vytvářet další.

Pomocí rozšiřujících modulů může program nabývat nečekaných schopností. Např. doplněk Drag and drop upload pro verze Moodle 1.9–2.x umožňuje přetahovat myší soubory z počítače do

programu. Do verze Moodle 2.3 toto rozšíření již není třeba přidávat, neboť funkce drag and drop je už zabudovaná v základu programu.

Nevýhody

Jistou nevýhodou při používání Moodle je ne zcela uživatelsky přátelské ovládání a nezbytnost zaškolení vyučujících. O existenci řady doplňků běžní uživatelé ani nevědí. Systém přináší možnost tvorby obsahu výukového webu bez znalosti jazyka HTML, ale v praxi je alespoň základní znalost HTML kódování pro vyučujícího/tvůrce studijního materiálu velkou výhodou. Příprava testů v tomto prostředí je komplikována nízkou ergonomií a nepřehledností rozhraní pro zadávání úloh. Souvisí to s velkým množstvím dalších funkcí, které Moodle jako komplexní LMS zvládá. Je nutně velmi obtížné uživatelské rozhraní optimalizovat pro všechny funkce současně. Systém Moodle nemá nástroje pro podporu týmové práce při přípravě testových úloh. Jisté drobné potíže může působit i nejednotnost používaných verzí programu, zvláště proto, že aktualizace je většinou třeba provádět postupně po jednotlivých krocích a některá dříve používaná rozšíření nemusí na nových verzích Moodle pracovat.

Z pohledu potřebné počítačové infrastruktury je třeba mít na zřeteli, že při velkém počtu současně testovaných studentů se server může začít zahlcovat požadavky [94]. Řešení je ve vhodném dimenzování infrastruktury, např. rozložením zátěže na více serverů [95].

Podpora zabezpečení testů v Moodle

Předností Moodle je vysoká úroveň zabezpečení testů. Systém umí omezit používání dalších aplikací, umí otevřít přístup k testu jen na základě hesla a případně i vymezených IP adres. Další bezpečnostní prvky přidává volitelný modul **Moodle inspektor**, který umožňuje kontrolovat některé nežádoucí komunikační aktivity studentů při testu.

Test lze v Moodle zabezpečit v několika úrovních

- heslem pro přístup do kurzu s testem,
- heslem k vlastnímu testu,
- stanovením počtu možných pokusů,
- stanovením časového odstupu mezi jednotlivými pokusy,
- přesným stanovením času pro otevření a uzavření testu,
- časovým limitem pro běh vlastního testu,
- zabezpečením pomocí JavaScriptu – vyskakovací okno přes celou obrazovku,
- omezením na IP adresy konkrétních počítačů.

Tab. 9.1 Výhody a nevýhody systému Moodle

Výhody Moodle	Nevýhody Moodle
▪ Nízké náklady na pořízení ▪ Celosvětové rozšíření ▪ Dostatečná dokumentace a návody ▪ Vysoká úroveň zabezpečení ▪ Velký výběr typů testových úloh ▪ Česká lokalizace	▪ Vysoká pracnost přípravy testu ▪ Mnoho dalších funkcí mimo testování ▪ Úroveň intuitivity neodpovídá dnešním standardům ▪ Vyžaduje proškolení nebo prostudování manuálů ▪ Nepodporuje plně týmovou spolupráci učitelů

Podrobný návod na tvorbu testu v Moodle lze najít na webu (http://docs.moodle.org/archive/cs/Přidání/Úprava_testu (http://docs.moodle.org/archive/cs/P%C5%99id%C3%A1n%C3%AD/%C3%BAprava_testu)). Testové úlohy lze připravit i mimo prostředí Moodle, nebo je možné je převzít ze sdílené položkové banky. Moodle podporuje standardy interoperability QTI a kolem 12 různých konkrétních importních formátů.

Rogo

Rogo je, stejně jako Moodle, volně šiřitelný program, je však zaměřený na elektronické testování ve všech jeho fázích. Rogo vzniklo v roce 2003 na lékařské fakultě University of Nottingham pod názvem **TouchStone** („prubířský kámen“). Po úspěchu na domovské fakultě bylo přijato jako klíčový systém pro celou univerzitu, převedeno na software s otevřeným zdrojovým kódem a při té příležitosti i přejmenováno, aby nedocházelo k záměnám s jinými systémy. *Rogō* v latině znamená "ptám se" [96].

Licence

Z pohledu licence je webová aplikace pro on-line testování Rogo je svobodně šiřitelný *open source* program, uvolněný pod GPL verze 3.0. Je tedy možné kód měnit, rozšiřovat jej a přispívat tak k projektu.

Rozšíření

Zásluhou britské společnosti JISC (http://www.jisc.ac.uk/whatwedo/programmes/elearning/assessmentandfeedback/rogo.aspx) získala finanční podporu otevřená komunita, která se zabývá vývojem Rogo. Tato komunita se také zasloužila o rozšíření systému na dalších pět britských univerzit (University of Oxford, University of Bedfordshire, University of East Anglia, De Montfort University a University of the West of Scotland), které u sebe systém implementují, testují a hodnotí.

Tvorbu testů v Rogo je možné si vyzkoušet na demo verzi University of Nottingham. Více informací naleznete na hlavní stránce projektu (http://rogo-oss.nottingham.ac.uk/).

Stav implementace v ČR

V ČR je systém Rogo nainstalován (na serveru 1. LF UK) na adrese https://www.rogo.cz/. 1. LF připravila český překlad prostředí a pracuje kontinuálně i na lokalizaci nápovědy. Studenti jsou do systému importování automaticky ze SIS a mohou se autentifikovat svým CAS účtem.

Dokumentace

- Správa projektu je otevřená a používá systém pro řízení projektů JIRA (https://rogo-eassessment.atlassian.net/).
- Dokumentace je k dispozici ZDE (https://rogo-eassessment-docs.atlassian.net/wiki/spaces/ROGO/pages/491548/Functional+Specification/)

Výhody

Na rozdíl od jiných programů Rogo pokrývá a podporuje všechny kroky 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 Rogo jsou naopak oponenti vyzváni k připojení do systému, takže testové úlohy systém vůbec neopustí. Komentáře a připomínky se opět vkládají přímo do Rogo a autoři úloh na ně mohou ihned reagovat. Poté, co test proběhl, je možné zobrazit popisné charakteristiky, histogram celkových skóre všech studentů, obtížnosti a diskriminační indexy položek apod. 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. [97]

Systém Rogo vznikl na lékařské fakultě a je pro výuku medicíny velmi dobře přizpůsoben. Mimo jiné dovoluje použít v úlohách interaktivních obrázků, na nichž student myší vyznačí hledaný objekt. Systém pak vyhodnotí, zda hledanou strukturu označil s požadovanou přesností.

Rogo podporuje autentizaci pomocí adresářové služby LDAP, kterou používá celá Univerzita Karlova v Praze. Pomocí webové služby jsou importovaná metadata popisující, jaké předměty mají studenti zapsané.

Tab. 9.2 Výhody systému Rogo

- Nízké náklady na pořízení
- Podpora celého procesu testování
- Podpora týmové spolupráce
- Vysoká úroveň zabezpečení
- Velký výběr typů testových úloh včetně obrazových

Nevýhody

Do vývoje testovacího systému Rogo je zapojena řádově menší komunita, než do vývoje Moodle. Je to dáno menším rozšířením, užším zaměřením i poměrně krátkou dobou od uvolnění do režimu open source. Podíváme-li se např. na blog nottinghamské univerzity, zjistíme, že počet příspěvků vztahujících se k Rogo, je asi desetkrát menší než k Moodle. Na druhou stranu pro řešení problémů a nových úkolů vývoje se používá tzv. lístkový systém, díky čemuž požadavky na další vývoj může přidat kterýkoli uživatel.

Tip: Jak Rogo počítá diskriminační schopnost položky?

inQsit

Za vtipným názvem odkazujícím se ke středověkým inkvizitorům se skrývá webový testovací systém vyvinutý na Ball State University a používaný na dvacítce dalších univerzit a škol v USA. Systém inQsit má bohatou dokumentaci a vyznačuje se mimořádně intuitivním prostředím pro zadávání otázek. Je možné jej vyzkoušet ve třicetidenní bezplatné verzi, nebo pořídit neomezenou licenci pro školství za 500 USD s ročními náklady 250 USD. Více informací je na adrese http://inqsit.bsu.edu/.

Questionmark

Questionmark je mimořádně propracovaný a uživatelsky přátelský komerční systém pro tvorbu, doručování a analýzu testů. Systém v několika modulech nabízí veškeré funkcionality, které si lze v testování a hodnocení výsledků studia přát. Pro tvorbu testových úloh jsou k dispozici tři moduly:

- Modul **Authoring Manager** je aplikace pro PC, která umožňuje tvorbu a organizaci položek a testů, jejich ukládání na lokálním PC nebo do vzdálené databáze. Jsou přitom k dispozici nástroje usnadňující práci i začátečníkům.
- Modul **Browser-Based Authoring** podporuje přípravu testových úloh z webového prohlížeče. Prostředí ovšem není tak příjemné jako u *Authoring Manager*, např. není k dispozici vizuální editor.
- Modul **Questionmark Live** je nástroj pro online přípravu položek a testů a jejich export do testovacích nástrojů této firmy. Licenční podmínky tohoto modulu neomezují počet autorů.

Další dva moduly pak zahrnují předchozí a pokrývají celý proces, včetně doručování a vyhodnocení testů.

- **Questionmark Perception** podporuje mimo jiné i tisk a skenování testů, distribuci testů prostřednictvím

prohlížečů, mobilních zařízení či USB disků. V testu je možné používat náhodné řazení otázek nebo adaptivní větvení. Výsledky testu je možné zpracovat do přehledné zprávy a otázky lze vyhodnotit položkovou analýzou.
- **Questionmark OnDemand** nabízí obdobné možnosti v režimu SaaS (software jako služba), kdy se program spouští ze serverů poskytovatele.

Výhodou všech modulů je podpora testování ve více jazykových mutacích (avšak čeština a slovenština dosud nejsou implementovány) a podpora používání videa, zvukových záznamů a grafických prvků v testových úlohách. Cena licence pro 1000 studentů se pohybuje v řádu stovek tisíc Kč. Více na stránkách http://www.questionmark.com.

ExaMe

Příkladem na míru vytvořeného českého testovacího systému je **ExaME** (http://www.exame.cz/), který od roku 1998 vyvíjí a používá EuroMISE Centrum UK a AV ČR [98]. Systém pracuje s položkami typu MTF. Počet nabízených odpovědí není omezen, alespoň jedna z nich však musí být správná a alespoň jedna nesprávná. Tvorba testů v systému ExaME funguje ve třech vrstvách. *Znalostní báze* obsahují položky s nabízenými odpověďmi a vyznačením jejich správnosti. V *ohodnocené bázi* jsou navíc i údaje o důležitosti a obtížnosti jednotlivých položek a nabízených odpovědí. Obtížnost a důležitost položek se mohou pro jednotlivé kurzy nebo školy lišit, z jedné znalostní báze tak lze připravit více ohodnocených bází. Samotný *test* se vytváří výběrem položek z ohodnocené báze.

ExaME nabízí dva druhy testů. *Pevné testy* jsou sumativní testy vytvářené učitelem na míru danému kurzu. Obvykle se pevný test administruje v počítačové učebně a přístup k němu je omezen časově i rozsahem IP adres, z nichž je dostupný. Studenti jsou při vyplňování testu přihlášeni ke svým účtům a položky dostávají v náhodném pořadí. *Automatické testy* se naproti tomu generují online dle studentem požadované délky a obtížnosti. Tyto formativní testy mohou studenti využívat např. při samostudiu v rámci e-learningových kurzů. Offline verze formativního testu byla spolu s příslušnou znalostní bází publikována také na CD jako příloha k učebnici [99].

Výsledek v testu (celkové skóre), jakož i položkové skóre se v ExaME vypočítává jako vážený průměr [100]. Systém umožňuje export výsledků ve formátu csv a následná položková analýza je připravena ve volně šiřitelném statistickém prostředí R.

Programy pro analýzu testů

Nabídka softwarových nástrojů pro analýzu výsledků testu je podstatně skromnější. Chceme-li získat informace o vlastnostech položek, můžeme použít některý z obecných statistických programů, jako je např. volně šiřitelné statistické prostředí R (http://www.r-project.org/). Jiných možností je jen málo. Jednou z nich je komerční software Iteman (http://www.assess.com/xcart/product.php?productid=417), zaměřený na položkovou analýzu a analýzu výsledků testu podle klasické testové teorie.

Iteman

Program Iteman je zajímavý komerční software pro analýzu položek a testů pomocí klasické teorie testů (CTT). Za cenu kolem 500 USD získáte nástroj, který vytváří rozsáhlé zprávy o kvalitě testových položek, o testu jako celku a o jeho psychometrických vlastnostech. Popis jedné ze starších verzí přináší Byčkovský [56] . 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). Poskytuje profesionální zprávy shrnující výsledky analýzy, včetně vložených tabulek a grafů. Program má uživatelsky velmi přátelské rozhraní. Oba programy jsou volně ke stažení ve verzi omezené na 50 položek a 50 testovaných. Plná verze pro akademické použití stojí 490 USD.

TiaPlus

Nizozemská společnost CITO, zabývající se vývojem psychometrických modelů a programů, nabízí pro akademické a nekomerční užití zdarma své softwarové produkty, mezi nimiž je program TiaPlus pro testovou a položkovou analýzu založenou na klasické teorii testů (CTT). Program nabízí široké spektrum možností ve způsobech bodování položek, poradí si s chybějícími položkami a podporuje práci se smíšenými typy formátů položek. TiaPlus také umožňuje analýzu neférovosti položek DIF a vytváří číselné i grafické výsledky analýz pro reportování. Více informací o produktu TiaPlus na stránkách společnosti CITO: http://www.cito.com/, nebo přímo na stránkách programu TiaPlus: http://tiaplus.cito.nl/.

Papírové testování

V anglické odborné terminologii se rozlišují dva pojmy: čistě *počítačové testování* a *počítačem podporované testování*. Ve druhém případě může sběr odpovědí probíhat i pomocí papírových dotazníků.

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 začátek s počítačovým zpracováváním testů a s používáním souvisejících metodik. Vhodně zvolené programy a technologie nám mohou výrazně ulehčit práci.

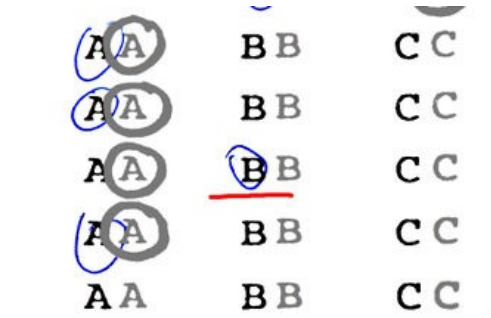
Vytváření papírových testů

Podoba odpovědních listů pro papírové testování musí odpovídat předpokládanému způsobu vyhodnocování. Pro ruční vyhodnocování stačí volně vytištěný seznam nabídnutých odpovědí. Pro automatizované vyhodnocení je třeba navrhnout formuláře tak, aby byly snadno strojově čitelné a vyhovovaly požadavkům na optické rozpoznávání značek (OMR, *optical mark recognition*). Příklad odpovědního formuláře je v příloze. Další příklady strojově zpracovatelných dotazníkových listů lze vyhledat na internetu pod termíny „bubble answer sheet“ nebo „OMR answer sheet“.

Testy lze generovat a tisknout přímo i z programů podporujících testování, jako je Moodle, který má pro tvorbu strojově čitelných formulářů rozšíření *Quiz OMR*. Další možností je specializovaný testový program Rogo, který rovněž podporuje tisk strojově čitelných formulářů, včetně vytváření několika verzí testu s různě seřazenými otázkami. V praxi se pro přípravu testových dotazníků užívá i řada proprietárních systémů, například pro zpracování přijímacích testů [101], [102]. I tyto programy obvykle nabízejí možnost promíchávat položky a vytvořit několik verzí stejného testu, což umožňuje lépe využít plochu testovací místnosti.

Vyhodnocování papírových testů

Ruční vyhodnocení testů pomocí průsvitky



Obr. 9.1 **Princip kontroly správnosti odpovědí pomocí průsvitky.**
Na snímku je odpovědní dotazník vytištěný černě, studentovy odpovědi jsou zakroužkované modře a fólie pro kontrolu správnosti je vytištěná šedě. Chybná odpověď byla podtržena červeně.

Tradičně se „kroužkovací“ papírové testy vyhodnocovaly pomocí průhledné fólie – průsvitky, na které byly vyznačené správné odpovědi. Průsvitka se přiložila na testový list se zakroužkovanými odpověďmi studenta a učitel mohl rychle a poměrně přesně vyhodnotit vyplněné testy. Nejčastějším zdrojem nepřesností bývaly chyby učitele při sčítání správných odpovědí. Výhodou bylo naopak snadné vyřešení položek, v nichž student svou odpověď dodatečně opravil.

Skenování a strojové rozpoznávání

Máme-li po ruce příslušnou techniku, nebo je-li potřeba vyhodnotit testy průkazně a bezrozporně (např. když má výsledek velký vliv na přijetí uchazeče nebo na další studium respondenta), je vhodné testy skenovat a strojově opticky vyhodnotit.

Formulář pro strojové rozpoznávání odpovědí je třeba navrhnout předem. Můžeme jej vytvořit ručně nebo použít některý specializovaný program (např. [103]). Pro čtení datových polí formulářů je možné pořídit některý z řady různých programů, např. Remark Office OMR [104], SPSS Data Collection Paper [105], ABBYY FlexiCapture [106], nebo Kodak Capture Pro Software [107].

Své služby na tomto poli nabízejí i četné firmy specializované na sběr dat [108]. Zajímavé řešení nabízí společnost Epson. Při nákupu jejích skenerů je možné jako bonus získat program Epson Document Capture Pro [109], který čtení dat z dotazníků rovněž podporuje. Digitalizace formulářů je také oblíbeným tématem seminárních prací z informatiky na Vysoké škole ekonomické v Praze, takže i v nich lze najít inspiraci (např. [110], [111]).

Je velmi důležité, jak konkrétní program pro vytěžování dat z formulářů dokáže řešit nestandardní situace. Mezi nejčastější patří nejasné vyznačení odpovědí, posunutí formuláře, zdeformování jeho digitálního obrazu, zmačkání odpovědního listu v podavači skeneru či vtažení dvou listů současně. Při velkém počtu formulářů, např. při přijímacím řízení, může být řešení takových problémů kritické. Před pořízením konkrétního programu je proto vhodné si opatřit reference od dosavadních uživatelů.

Příklady testování pomocí skenovaných formulářů v praxi

Tématu vytěžování papírových dotazníků věnuje pozornost řada škol. Uveďme zde několik příkladů.

Česká zemědělská univerzita v Praze zpracovává v každém dni přijímacího řízení 2500 testů o 20–50 otázkách. Kvůli problémům při skenování změnila technologii z původního software Tests Checker 2.30 na program Remark Office OMR, s nímž je řešení incidentů při rozpoznávání snazší a efektivnější [112].

Fakulta informatiky **Masarykovy univerzity v Brně** se dlouhodobě zabývá vývojem systému pro vytváření, zpracovávání a vyhodnocování písemných testů [113], [114]. Vynaložné úsilí se zúročilo v systému Skenování písemek, který je dnes součástí informačního systému univerzity [115]. Je již vyzkoušený, používá se například na lékařské fakultě MU pro testování předmětu Lékařská biologie. Několik set studentů dostává testy se zhruba stovkou otázek. Testy se následně skenují, vyhodnocují a výsledky se přenášejí do informačního systému školy [116].

Náklady elektronického testování

Příprava, provedení a vyhodnocení testů jsou velmi 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 standardizovaných 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 zkušených, neboť vysoké pořizovací náklady budou vyváženy nízkými provozními náklady.

Jednou z nejdražších položek při hodnocení studentů pomocí testů jsou *náklady na testovou úlohu*. Vytvoření dobrých otázek totiž vyžaduje tým expertů v příslušném oboru, proškolených navíc v metodice tvorby testů. Další výdaje přináší oponování otázek a pilotní testování s dostatečně velkou skupinou studentů.

Náklady na přípravu položek pro důležité testy se odhadují na 1000 USD za úlohu; obecně se dá říci, že náklady na položku neklesají pod 300 USD. [44]

Celkové náklady na vývoj jedné kvalitní položky do adaptivního přijímacího testu spočítal např. Rudner [117]. 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 [75] (asi 2000 položek v bance), dojdeme k astronomické částce 3–4 miliony USD za test [74] .

Celkové náklady se mohou snížit, použijeme-li již hotové otázky. Mohou to být otázky, které jsme připravili v minulých kolech testování. Nebo můžeme testové položky koupit. K dostání jsou sady otázek z různých oborů medicíny v obvykle používaných testových formátech (SBA a MCQ). Nakladatelství Amazon nabízí např. 500 otázek z chirurgické anatomie a akutní péče ve formátu SBA za 22 GPB [118]. Položky je ovšem nutné odborně přeložit a použít přitom aktuální vyučovanou terminologii, což vede i v tomto případě k nezanedbatelným výdajům, náklady na kalibraci zůstávají stejné.

Snížení nákladů na pořízení a kalibraci položek by mohlo přinést sdílení otázek mezi pedagogy. Této problematice, včetně úspěšných příkladů řešení ze zahraničí, se věnujeme v samostatné kapitole 9.6.2 Sdílené banky testových úloh. Efektivitou vynakládaných prostředků na hodnocení studentů medicíny se v obecnější rovině zabývá kapitola Cost-effective assessment v knize [119].

Automatické generování položek a testů

S tím, jak přibývá počítačem podporovaného testování, a zvláště pak s rozvojem adaptivního testování, nabývají na významu metody, kterými se automatizuje tvorba testových úloh. 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 reviduje a upravuje. 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 [120].

Automatické vytváření položek (*Automatic Items Generation*, **AIG**) představuje jiný přístup k tvorbě úloh. Doplňuje tradiční postup o speciálně naprogramované algoritmy, které vytvářejí klony původních otázek. Cílem AIG je vytvořit velké množství vysoce kvalitních položek s malým vkladem lidské práce před samotným testem [121].

AIG se v principu skládá ze dvou kroků. V prvním vytvoří tým specialistů úlohu a zvýrazní ty její parametry či části, které se dají měnit. Ve druhém kroku se tyto části (často strojově, např. pomocí slovníků synonym) nahrazují či mění tak, aby vznikaly nové položky [122].

Výhodou automatické tvorby položek je, že předem známe psychometrické vlastnosti odvozených úloh, pokud jsou známé psychometrické parametry původní zdrojové otázky. Nevýhodou naopak je, že odvozené položky nemůžeme považovat za vzájemně nezávislé, neboť navzdory vnější odlišnosti testují stejnou znalost, a nemá tedy smysl je použít současně v jednom testu.

Banky testových úloh

Nejstarší formou položkových bank byly patrně lístkové katalogy testových otázek, které si vytvářeli učitelé sami [123]. Pro sdílení položek je výhodnější použít počítačové **banky testových úloh** (**BTÚ**, *item bank*). Jejich úkolem není sdílet jen samotný text úlohy, ale i klíčová slova, informace o původu položek, jejich psychometrických vlastnostech a další metadata.

Banku testových úloh můžeme chápat i jako ucelený informační systém, který zahrnuje jak „úložiště“ testových položek, tak i veškeré procesy od jejich vývoje až po sestavování testů. Takto pojímá banku testových úloh například CERMAT [124], který vypracoval komplexní model banky testových úloh včetně návrhu řešení HW i SW, mapy procesů, analýzy rizik atd. V našem textu se však budeme držet užšího vymezení ve smyslu **úložiště testových úloh** (též **položková banka**).

Jak jsme již uvedli, banky testových úloh obsahují nejen samotný text úlohy, ale i rozsáhlé informace o vývoji úlohy a psychometrické vlastnosti položek. Tato metadata položky většinou zahrnují následující údaje [125]:

1. autor položky,
2. datum vložení,
3. oponent,
4. status položky (např. nová, pilotovaná, aktivní, vyřazená),
5. hraniční skóre položky dle Angoffovy metody,
6. číslo správné odpovědi,
7. formát položky,
8. parametry položky podle klasické teorie,
9. parametry položky podle IRT,
10. deskriptor MeSH,
11. téma výuky,
12. uživatelem definovaná pole.

Vzhledem k tomu, že položková banka je v podstatě jednoduchá databáze, může být uložena v téměř libovolném databázovém systému, nebo dokonce v prostředí tabulkového procesoru. Nicméně existuje řada komerčních řešení určených přímo pro vedení bank testových úloh [126]. Obsahují často nástroje pro hodnocení položek, umožňují zobrazit úlohu v podobě, jak ji uvidí testovaný, podporují tisk testů a podobně [127].

Formáty výměny testových položek

Testové položky mohou vznikat v mnoha různých systémech a v nejrůznějších systémech je lze také používat. Je proto mimořádně důležité, aby je bylo možné přenášet mezi platformami. Proto postupně vznikla řada formátů, které export a import umožňují. Jednoduché proprietární formáty podporují často přenos jen mezi několika určitými programy, nebo

podporují jen několik málo formátů testových úloh. Na druhou stranu bývají 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.

QTI

QTI (*Question & Test Interoperability*) je název standardu pro výměnu testových úloh, který vytvořilo IMS Global Learning Consortium. Konsorcium IMS vyvíjí specifikace pro interoperabilitu výukových materiálů a zabývá se i standardy výměny dat mezi e-learningovými nástroji LTI (*learning tools interoperability*). Standardy LTI umožňují přenos dat, např. známek z testovacího programu do virtuálního výukového prostředí. K vytvoření standardu výměny otázek QTI vedla potřeba zabránit zmaření práce, která byla vložená do přípravy úloh, při změně technologie testování.

Poslední všeobecně přijímanou verzí QTI je vydání 1.2.1. Novější verze 2.0 zatím nebyla implementovaná všemi výrobci testovacích programů a verze 2.1, která by měla řešit problémy předchozích verzí, dosud nebyla uvolněna ve stabilní verzi.

Proprietární formáty výměny

Aiken

Aiken je velmi jednoduchý formát umožňující výměnu položek s vícenásobnou odpovědí. Jeho výhodou je syntaxe, která se blíží přirozenému jazyku:

> Kmen otázky ... :
> A) distraktor 1
> B) správná odpověď
> C) distraktor 2
> D) distraktor 3
> ANSWER: B

GIFT

Proprietární formát Moodle pro import otázek do testů **GIFT** je složitější, ale podporuje daleko širší spektrum typů testových položek: multiple choice, true/false, short answer, matching a numerical, jakož i otázky na doplňování chybějících slov. Importovaný textový soubor může obsahovat kombinaci různých typů otázek, komentáře, názvy otázek, hodnocení odpovědí pomocí vah a další parametry. V českém jazyce je možné psát otázky např. v textovém editoru Word, pokud použijeme kódování Windows-1250 a výsledek uložíme jako prostý text (tj. soubor.txt).

Sdílené banky testových úloh

Ukázalo se, že sdílet testové otázky pro hodnocení některých etap studia medicíny je velmi přínosné. Důvody ke spolupráci jsou přinejmenším tři.

> 1. Sdílení testových úloh a spolupráce při jejich tvorbě vede ke **snižování nákladů**.
> 2. Spolupráce a sídlení výsledků vedou ke zvyšování kvality testů a přispívají tak ke **zvyšování úrovně zdravotní péče**.
> 3. Standardizace testování a sdílení položek zlepšují **mezinárodní porovnatelnost výsledků výuky**, což je důležité při hledání financí a získávání akreditací.

Ze zahraniční známe několik sdružení, která provozují položkové banky a podporují sdílení kalibrovaných položek. Jsou to především

- globální sdružení **IDEAL**, které vzniklo rozšířením spolupráce na sdílení otázek mezi jednou kanadskou a jednou čínskou lékařskou školou,
- **Medical Assessment Alliance**, která působí v německé jazykové oblasti,
- **Medical Schools Council Assessment Alliance**, jež sdružuje všechny lékařské školy ve Velké Británii.

Ze spolupráce plynou nejen výhody, ale objevují se i problémy spojené s přenosem otázek z jedné školy na druhou. Způsobují je rozdíly v učebních osnovách, kulturní rozdíly a podobně. Ukazuje se však, že otázky z položkových bank lze pro vlastní použití přizpůsobit s vynaložením relativně malého množství práce. [128]

IDEAL Consortium

IDEAL Consortium (http://www.idealmed.org/) (*International Database for Enhanced Assessments and Learning*) je dobrovolné sdružení 27 vysokých škol z 11 zemí z celého světa [129], na nichž se medicína přednáší v angličtině. Nejvíce jich je z Austrálie a Kanady, významný podíl tvoří země středního východu. Jde přitom v mnoha případech o významné univerzity, které se v žebříčcích řadí mezi nejlepších 150 na světě.

Cíle sdružení jsou formulovány ve třech bodech:

- vytvořit a sdílet velký počet kvalitních testových otázek pro účely lékařského vzdělávání v mezinárodním měřítku,
- podporovat komunikaci mezi lékařskými fakultami v oblasti standardů kvality testování,
- podporovat výzkum a rozvoj mezinárodních norem v posuzování lékařské způsobilosti.

Konsorcium buduje dvě oddělené databáze otázek. Databáze pro sumativní testování obsahuje 13 tisíc otázek a přístup do ní je striktně omezen. Druhá databáze s otázkami pro formativní hodnocení obsahuje 6 tisíc položek určených pro sebevzdělávání a přístup do ní je možný po přihlášení. Výše uvedené informace o počtech otázek pocházejí z webových stránek konsorcia (http://www.idealmed.org/). V publikaci z roku 2012 je celkový počet otázek v databázi odhadnut na 32 tisíc.[130].

Otázky jsou opatřeny klasifikátory (metadaty), které v nich umožňují efektivně vyhledávat. Jedním z klasifikátorů je řízený slovník biomedicínských deskriptorů MeSH (Medical Subject Headings). Dalšími jsou předdefinované kvalitativní charakteristiky, jako jsou disciplína, nebo úkoly lékaře a nakonec výkonové (psychometrické) charakteristiky jako jsou úroveň obtížnosti, diskriminační schopnost, počet použití apod. Kvalita otázek se zdá velmi dobrá a i přes kulturní rozdíly odrážející místo vzniku je lze s malým vkladem práce přizpůsobit pro místní podmínky [128].

Konsorciem byla vytvořena sada programů, které slouží:

- Pro skladování, vyhledávání a získávání položek (testových otázek).
- Pro administraci hodnocení, analýzu studentských odpovědí a poskytování zpětné vazby.
- Pro správu a sdílení položek uvnitř i mezi institucemi

Vývojové náklady byly menší než 1 mil. USD a byly hrazeny z hongkongských grantů pro rozvoj výuky.

Podporované typy testových položek

- SBA (single best answer) – otázky s jedinou nejlepší odpovědí
- EMQ (extended-matching questions) – rozšířené přiřazovací otázky
- SAQ (short-answer questions) – otázky s krátkou tvořenou odpovědí
- MRQ (multiple response questions) – otázky s více správnými odpověďmi (~ MTF)
- MEQ (modified assay question) – modifikovaný esej
- OSCE/OSPE (objective structured clinical/practical examination) - objektivně strukturované klinické/praktické zkoušení

Součástí otázek mohu být i vložené objekty – audio, video, grafy a tabulky. Otázky s sebou nesou metadata zahrnující mimo jiné informace o původním autorovi, o zdrojovém materiálu, o zpětné vazbě od studentů a podobně. Banky otázek s omezeným přístupem (pro sumativní hodnocení) i volným přístupem (pro formativní testování) jsou každoročně aktualizovány. Software sdružení IDEAL podporuje práci jak na lokálním PC, tak online na vzdáleném serveru. Programy umožňují nejen zadávání otázek a klasifikaci položek, ale též jejich položkovou analýzu.

Kvalita obsahu je podpořena i pravidly, podle nichž jsou členové sdružení přijímáni. Musí jít o (lokálně) akreditované vysoké školy, které mají zavedenou kontrolu kvality hodnocení studentů. Předpokladem je ochota přispět nejlepšími testovými otázkami (minimálně 150 položek ročně) a podmínkou přijetí je i záruka některého stávajícího člena sdružení.

Sdružení poskytuje svým členům podporu formou školení školitelů (v Hongkongu), vydává manuály a učí i pravidlům pro správnou tvorbu otázek.

Více informací na stránkách sdružení: http://www.idealmed.org/.

Medical Assessment Alliance

V německy mluvících zemích se spoluprací při tvorbě a sdílení testových otázek zabývá **Medical Assessment Alliance** (https://www.ims-m.de/joomla/index.php/de/das-imsm-othermenu-48) (MAA), která provozuje svůj *Item Management System* (IMS). Vznikla v roce 2006 jako výsledek spolupráce lékařských fakult univerzity v Heidelbergu, v Berlíně a v Mnichově. Skupina se od té doby rozrostla na 31 fakult v Německu a ve Švýcarsku. Na přípravě otázek spolupracuje více než 2 800 uživatelů v 750 pracovních skupinách [130].

Pro zajištění podpory při spolupráci a výměně otázek byla založena nezisková organizace, jejíž náklady se dělí mezi všechny členy. Problémy se řeší na společných jednáních a rozhodnutí se přijímají na základě hlasování. Hlavní cíle aliance jsou

- spolupráce ve všech oblastech testování studentů,
- zajištění kvality a dostupnosti otázek pro testování,
- vypracovávání a zavádění standardů,
- podpora inovativních forem testování,
- vývoj v oblasti testování.

Zmíněný systém IMS (*Item Management System*) je koncipován jako řešení „vše v jednom“, podporující všechny fáze přípravy otázek. Testování i vyhodnocování výsledků odpovídá směrnici pro zkoušení na lékařských vysokých školách v Německu podle mezinárodních standardů. [131]

Pracovní cyklus v IMS se skládá ze šesti modulů:

- **Tvorba obsahu zkoušky** a její sestavování
- **Předběžná recenze** jako nástroj zabezpečování jakosti před zkouškou
- **Zkoušení** – provádění zkoušky v písemné, ústní nebo praktické formě
- **Vyhodnocení zkoušky** podle ustanovení příslušných studijních předpisů
- **Následná recenze** jako záruka kvality po zkoušce
- **Mezifakultní zásoba položek** pro sběr, sdílení a výměnu všech získaných údajů

V databázi IMS je aktuálně uloženo více než 90 000 otázek. Od roku 2007 proběhlo přibližně 4 600 testování studentů.

Medical Schools Council Assessment Alliance

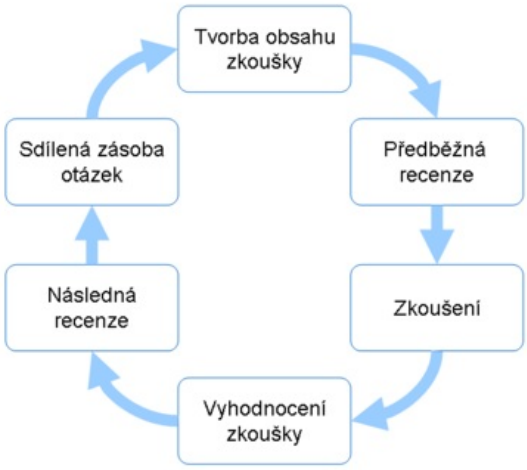
Medical Schools Council Assessment Alliance (http://www.medschools.ac.uk/MSC-AA/Pages/default.aspx) (MSC-AA) je organizace sdružující 31 lékařských vysokých škol ve Velké Británii. Jejím cílem je zlepšit současnou praxi hodnocení výsledků pregraduální výuky. Spolupráce byla zahájena v srpnu 2010 na základě rozhodnutí General Medical Council, že lékařské školy mohou sdílet testové položky, aby se dosáhlo srovnatelného hodnocení studentů. Jde o alternativu k porovnávání výsledků výuky pomocí národních licenčních zkoušek.

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 otázek ve formátech MCQ a EMQ a mělo dosáhnout vyšší kvality testových otázek. Sdružení se postupně rozrůstalo o další školy a po roce 2009 se přeměnilo na MSC-AA.

Konečným cílem MSC-AA je zvětšit důvěru veřejnosti, zaměstnavatelů a regulačních orgánů v kvalitu absolventů lékařských vysokých škol, a to

- vývojem vysoce kvalitních testových položek pro pregraduální studenty,
- sdílením zkušeností a tvorbou přidané hodnoty prostřednictvím spolupráce, spíše než konkurencí,
- prokázáním rovnocennosti norem uplatňovaných lékařskými fakultami.

Díky MSC-AA se do tvorby testových položek a přípravy testů zapojuje více pedagogů, což vede k lepší kvalitě hodnocení výsledků výuky. Partnerské školy mají přístup k bance vysoce kvalitních testových položek v různých formátech, s dobrou reliabilitou a validitou. 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 srovnatelnost těchto zkoušek.



Obr. 9.2 Pracovní cyklus IMS

Sdílení otázek v portálu TESTY

I v českém prostředí a lékařském kontextu se objevují první pokusy o sdílení testových otázek. Záměrem portálu TESTY (http://testy.lf3.cuni.cz/moodle/) bylo vytvořit zabezpečený webový prostor pro sdílení testových otázek lékařských fakult [132].

Portál vznikl jako samostatná instance Moodle. V pilotním provozu byla naplněna složka biofyziky, ve které bylo ke konci roku 2012 asi 3000 otázek ve formátu MTF od autorů z pěti lékařských fakult. Otázky prošly anonymní oponenturou ostatními autory a byly roztříděny do osmi kategorií podle tematického zaměření. V ostatních oborech se podařilo shromáždit významný počet otázek ve složce patologie a prvních 500 otázek v dosud nezveřejněné složce biochemie. Otázky byly importovány ve formátu Aiken. Návod k obsluze obsahuje dvě instruktážní videa k založení testu a tvorbě testů z databáze otázek. Probíhá diskuse o následujících otázkách:

Má se celá databáze otázek nebo její část otevřít studentům?
Má smysl statisticky ověřovat obtížnost otázek stanovenou oponentem?
Má se studentům umožnit samotestování?

Projekt ukázal, že sdílení otázek je možné, prakticky proveditelné a je o ně na fakultách zájem. Současně projekt ukázal na meze zvoleného přístupu:

- Řada pedagogů se sdílením svých otázek váhá, z části i proto, že nejsou jasně stanovena pravidla, co se s otázkami stane, a obávají se exponování otázek.
- Podporované formáty otázek jsou zatím omezeny na MCQ (respektive MTF).
- Ukládání otázek v Moodle je výhodou pro uživatele tohoto LMS systému, ale může být omezením pro ostatní.
- Projekt v současné podobě nepracuje s kalibrovanými položkami a nejsou vytvořeny mechanismy pro sběr a ukládání psychometrických charakteristik položek.

Závěr

Motto: *Pokud máš jen dvě mince, kup za jednu chléb a za druhou květinu. Tou první si kupuješ život, tou druhou dáváš tomu životu smysl.* (Konfucius, 500 př.n.l.)

Cyklus přípravy testů zpravidla končí vyvozením závěrů. Získané zkušenosti se mohou odrazit v úpravě testových otázek, které položková analýza identifikovala jako problematické či zcela nevyhovující. Může se také promítnout do závěru, že testovou agendu by neměl připravovat jediný člověk, či do poznání, že spolupráce s kvalifikovanými statistiky práci urychlí a zkvalitní.

V průběhu posledních desetiletí se v oboru testování objevily nové formáty testových položek, které lépe testují osvojení hlubších znalostí (SBA), nové programy podporující testování v jeho jednotlivých krocích i jako celek (Rogo), i vyšší požadavky na kvalitu testových otázek, které postupně vedly ke spolupráci při jejich tvorbě a nakonec i k jejich sdílení v rozsáhlých položkových bankách, na nichž spolupracují pedagogové a odborníci z desítek škol a univerzit.

Cílem tohoto průvodce bylo seznámit čtenáře s rámcem, principy testování, klíčovými pojmy a praktickými kroky standardizace tak, aby mohl při výběru metod a postupů testování výsledků výuky činit kvalifikovaná rozhodnutí.

Přílohy

Pohřebiště formátů otázek s mnohočetným výběrem odpovědi (MCQ)

Početnou rodinu otázek s mnohočetným výběrem odpovědi (MCQ) můžeme rozdělit do dvou skupin položek. Do první patří ty, které od zkoušeného vyžadují vyznačit všechny vhodné odpovědi (ano/ne), do druhé pak položky, které od zkoušeného vyžadují, aby z nabídnutých odpovědí vybral jednu, nebo jiný předem stanovený počet odpovědí, které nejlépe odpovídají na otázku. Do každé skupiny patří řada možných formátů; nejvýznamnější jsou následující podtypy MCQ [13]

Tab. 10.1 Nejvýznamnější podtypy MCQ (položky typu ano/ne)

Položky typu „ano/ne“ – zkoušený má označit všechny možnosti, které jsou pravdivé:
typ C – položky s odpověďmi typu (A/B/obě/žádná)
typ K – komplexní nebo mnohonásobné položky typu ano/ne
typ X – jednoduché položky typu ano/ne – v našem textu označované **MTF**

Tab. 10.2 Nejvýznamnější podtypy MCQ (položky s jedinou nejlepší odpovědí)

Položky „s jedinou nejlepší odpovědí“ – zkoušený má označit jen jednu možnost, která nejlépe odpovídá zadání:
typ A – otázky s jedinou nejlepší odpovědí ze tří nebo více možností, v našem textu označované jako **SBA**
typ B – přiřazovací otázky se čtyřmi nebo pěti možnostmi v sadě 2–5 položek
typ R – rozšířené přiřazovací otázky – v našem textu označované jako **EMQ**

Písmena použitá pro označení jednotlivých variant MCQ nemají žádný hlubší význam, byla k typům otázek přiřazována postupně, tak jak se objevovaly.

Jak již tento neúplný **přehled formátů MCQ** napovídá, v průběhu let bylo vyvinuto a v praxi vyzkoušeno velké množství (více než dvacet) podtypů otázek s mnohočetným výběrem odpovědí. Postupně se ukázalo, že některé podtypy jsou si natolik podobné, že je není vhodné rozlišovat.

> Je příznačné, že kapitola věnovaná přehledu všech podtypů MCQ v manuálu amerického National Board of Medical Examiners (NBME) se jmenuje **Hřbitov formátů položek dle NBME** [13]. V osmdesátých letech 20. století popisovala doporučení NBME již jen sedm typů otázek s mnohočetným výběrem odpovědí (označovaných jako typy A, B, C, G, K, X a M).
>
> Další revize již počítala pouze se čtyřmi podtypy (A, B, C a K). V devadesátých letech 20. století pak byly vyvinuty rozšířené přiřazovací otázky (R-typ, EMQ). Poslední doporučení NBME z r. 1998 pracují jen se dvěma podtypy otázek s mnohočetným výběrem odpovědí: s otázkami s jedinou nejlepší odpovědí (A-typ, SBA) a s rozšířenými přiřazovacími otázkami (R-typ, EMQ).
>
> Použití ostatních podtypů v medicíně považují autoři těchto doporučení za vysloveně nevhodné.

Ukázky vybraných starších typů otázek s mnohočetným výběrem odpovědí (MCQ)

Typ B – přiřazovací otázky

Přiřazovací otázky typu B se v současné době nahrazují rozšířenými přiřazovacími otázkami (EMQ).

> *POKYNY: Ke každé očíslované položce přiřaďte jeden nadpis označený písmenem, který s ní nejtěsněji souvisí. Ke každé očíslované položce patří právě jeden nadpis. Každý nadpis lze použít jednou, vícekrát nebo vůbec.*
>
> *A. Koarktace aorty*
> *B. Ductus arteriosus patens*
> *C. Fallotova tetralogie*
> *D. Aortální cévní prstenec*
> *E. Atrézie trikuspidální chlopně*
>
> *1. Zlepší se systemo-pulmonární arteriální anastomózou*
> *2. Nejčastější příčina vrozené cyanotické srdeční vady*
> *3. Chirurgicky se koriguje resekcí a anastomózou end-to-end*
> *4. Možná příčina dysfagie u dětí*
> *5. Hypertenze na pažích a hypotenze na dolních končetinách.*

Typ C – otázky typu A/B/obě/žádná

> *POKYNY: Ke každé číslované položce přiřaďte jednu možnost označenou písmenem. Každá možnost označená písmenem může být použita jednou, vícekrát či vůbec.*
>
> *A. Malárie způsobená Plasmodium vivax*
> *B. Malárie způsobená Plasmodium falciparum*
> *C. Platí A i B*
> *D. Neplatí A ani B*

1. Při záchvatu je léčbou volby kombinace primachinu a chlorochinu
2. Při pobytu v endemické oblasti se k potlačení záchvatů užívá chlorochin 1× týdně
3. Trvale se vyléčí chlorochinem
4. Profylaxe infekce se provádí chlorochinem 1× týdně

Typ K – komplexní položky typu ano/ne

Vyberte pravdivá tvrzení

A – platí 1, 2 a 3
B – platí 1 a 3
C – platí 2 a 4
D – platí 4
E – platí vše

Dítě s akutní exacerbací revmatické horečky má obvykle

1. zvýšenou sedimentaci
2. prodloužený interval PR
3. zvýšený titr antistreptolysinu O
4. podkožní uzlíky

Podrobnosti o typech testových úloh, pokyny pro jejich vytváření a hodnocení

Esej

Esej může být zadán dvěma základními způsoby:

1. Testovaný dostává otázku nebo sadu otázek, na něž má odpovědět ve stanoveném čase. Dodržují se pravidla obvyklá u zkoušek, není povoleno používat žádné zdroje informací.
2. Zadává se téma eseje. Testovaný píše esej ve vymezeném čase, ale smí používat různé zdroje informací. V některých případech se témata eseje dokonce sdělují testovaným předem.

V prvním případě je výsledek více závislý na znalostech zkoušeného a na jeho paměti. V každém případě esej dobře testuje schopnost komplexního zpracování odpovědi na obtížnou otázku (někdy dokonce na otázku, jejíž řešení není dosud známé, nebo kterou ani není možné za stávajícího stavu poznání zodpovědět). Předpokladem úspěchu testovaného jsou ovšem také jeho vyjadřovací schopnosti, znalost jazyka a další. Zkoušení pomocí eseje je racionální, pokud chceme testovat i tyto schopnosti. V případě, že máme hodnotit pouze přesně definované okruhy znalostí a dovedností, které nesouvisejí se zběhlostí v používání jazykových prostředků, je vhodnější sáhnout po jiném formátu otázek.

Pokyny pro zadávání eseje

1. Rozhodněte, zda testovaní budou moci při psaní eseje používat nějaké zdroje informací.

   Záleží na tom, zda vyžadujeme, aby si student přesně pamatoval větší množství dat z oblasti, kterou má v eseji zpracovat, nebo zda chceme spíše testovat, jaký má ve zkoušené oblasti přehled a jak dokáže informace zpracovávat.

2. Vyberte problém, který je možné uspokojivě pojednat v čase, který bude pro psaní eseje k dispozici, popřípadě jednoznačně vymezte určitou podotázku, kterou se mají zkoušení zabývat.
3. Zadání formulujte jednoznačně. V zadání se vyhněte "chytákům", záporům, zkratkám apod.
4. Popište, jak strukturovanou odpověď očekáváte, popřípadě z jakého pohledu má být problém zpracován.
5. Pojmenujte myšlenkové pochody, které chcete v odpovědi hodnotit. Namísto slov *popište*, *vysvětlete* používejte raději *porovnejte*, *doporučte*, *odhadněte další průběh*, *interpretujte výsledky*, *navrhněte další postup* apod.
6. Nezadávejte kontroverzní témata, u nichž lze považovat za přijatelné i značně rozdílné názory.
7. Napište modelovou odpověď nebo seznam součástí, které má mít správná odpověď.
8. Zadání vyzkoušejte na malé skupině lidí, o kterých lze předpokládat, že znají správnou odpověď. Zkontrolujte, že v odpovědích této skupiny se objevily všechny položky, které mají být součástí správné odpovědi. Každá položka se musí objevit aspoň v některé odpovědi z této skupiny.
9. Nechejte zadání zkontrolovat třem recenzentům.

Pokyny pro hodnocení eseje

1. Každý esej by měl být ohodnocen dvěma hodnotiteli, kteří pracují nezávisle na sobě. Pokud to není možné, měl by jeden hodnotitel ohodnotit stejnou otázku všem testovaným osobám.
2. Všichni hodnotitelé by se měli seznámit s odpověďmi všech testovaných na otázku, kterou bodují.
3. Je vhodné poskytnout hodnotitelům modelové odpovědi pro jednotlivé známky. Modelové odpovědi by měly odpovídat hranicím mezi známkami (tedy pokud se např. otázka známkuje od A do E, měli by hodnotitelé mít k dispozici modelovou odpověď, která odpovídá hodnocení mezi A a B, další pro B/C atd.).
4. Pokud není možné vypracovat modelové odpovědi pro hodnocení, vypracujte osnovu hodnocení.
5. Odpovědi by měly být hodnocené anonymně.

Otázky s krátkou odpovědí (SAQ)

Při vytváření otázek s krátkou odpovědí pro písemné testy dbejte těchto doporučení: [14]

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

  Dobrá otázka s krátkou odpovědí testuje znalost konkrétních fakt nebo schopnost analyzovat a klinicky 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 a jak bude otázka 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 tučného písma nebo kurzívy („Které antibiotikum je v této situaci ***ne***vhodné?“).

- **Dbejte, aby odpovědi nenapověděla např. velikost vynechaného místa pro její vepsání.**
- **Připravte pokyny pro hodnotitele.**

  Uveďte všechny možné správné odpovědi. Odpověď, která kromě správného řešení obsahuje i další nesprávnou odpověď, se obvykle hodnotí jako nesprávná.

> *K čemu slouží stanovení dusitanů v moči?*
>
> *Odpovězte několika slovy (1 bod)*
>
> *Pokyny pro hodnotitele:*
>
> *Za správnou odpověď se považuje kterákoliv z následujících:*
>
> *k průkazu infekce močových cest, k průkazu bakterií*
> *hodnocení: 1 bod*
>
> *Nesprávná odpověď, žádná odpověď, správná odpověď a současně další nesprávná odpověď se hodnotí 0 body.*
>
> *Samozřejmě se předpokládá, že otázka bude bodována kvalifikovaným hodnotitelem, který na základě těchto pokynů jako správnou uzná např. i odpověď „k průkazu bakteriurie“, „je pozitivní při močové infekci“ apod.*

Některé testové programy umožňují automatické vyhodnocování otázek s krátkou odpovědí. Je účelné použít tyto funkce, pokud je požadovaná odpověď opravdu velmi krátká a jednoznačná. Autor otázky musí definovat všechna možná znění správné odpovědi. Nesmí se zapomenout na možná synonyma či pravopisné varianty odpovědi (např. termíny zakončené na „-óza“/„-osa“). Většina programů umožňuje nastavit, zda se mají rozlišovat velká a malá písmena, jaká je dovolená tolerance při vkládání číselných údajů či jaké jsou přípustné pravopisné tvary konkrétního slova.

Otázky s jednou nejlepší odpovědí

Zásady pro tvorbu otázek s jedinou nejlepší odpovědí:

- **Každá otázka by měla být zaměřena na důležitý klinický problém.**

  Zaměřte se na problémy ze života. Neztrácejte čas testováním triviálních ani příliš složitých otázek. Nepoužívejte „chytáky“. Zkoušejte znalost, nikoliv pozornost.

- **Každá otázka by měla testovat využití znalostí, nikoliv znalost izolovaného faktu.**

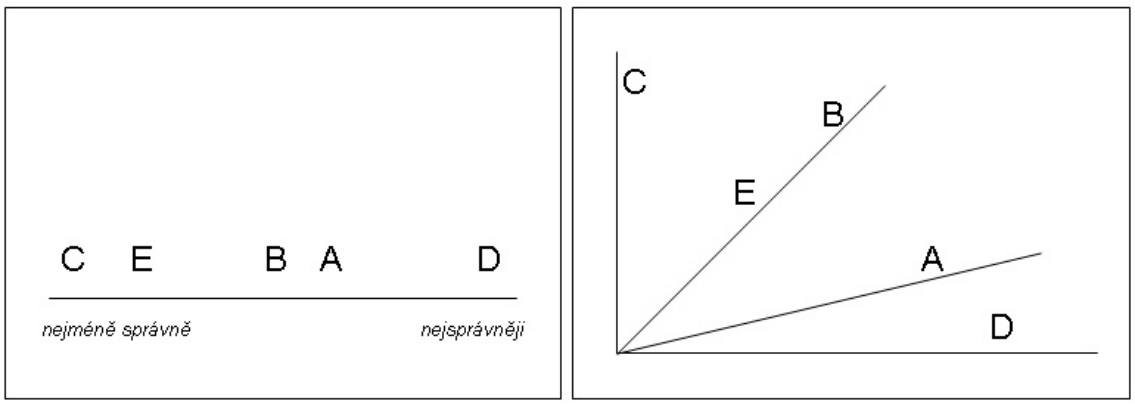
  Sama otázka může být poměrně dlouhá, ale odpovědi by měly být krátké.
  Klinická otázka by měla být uvozena medailonkem, který popisuje pacientovy obtíže a anamnézu (včetně délky trvání příznaků), výsledky vyšetření, prvotní léčbu. Není třeba za každou cenu použít všechny tyto položky, údaje by však měly být v uvedeném pořadí.
  Otázky z teoretických předmětů mohou být uvozeny velmi krátkým klinickým medailonkem nebo „laboratorním medailónkem“.

- **Otázka musí být formulována jasně. Musí být možné na ni správně 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, otázku přepište.

- **Všechny nabídnuté odpovědi (včetně nesprávných) musejí být homogenní.**

  Všechny nabídnuté možnosti musejí spadat do stejné kategorie. Lze to jednoduše ověřit: nabídnuté odpovědi lze seřadit na pomyslné přímce od nejméně správné po nejsprávnější:



Správně:
nabídnuté možnosti jsou homogenní

Špatně:
nabídnuté možnosti spadají do různých kategorií

Obr. 10.1a,b Na příkladu vlevo (a) 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 (b) 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

Distraktory (nesprávné odpovědi) musí působit pravděpodobně. Možnosti mají být seřazeny náhodně nebo podle abecedy. Pokud možnosti obsahují číselný údaj, měly by být seřazené podle něj.
Všechny možnosti by měly být psané podobným stylem, měly by mít stejnou větnou stavbu a podobnou délku.

- **Nepoužívejte otázky typu „Které z následujících tvrzení je správné?“, nebo „Všechna z následujících tvrzení jsou správná s výjimkou“.**

  Stejně tak je nevhodná většina otázek, ve kterých se vyskytnou slova „vždy“, „většinou“, „zřídka“, „výjimečně“, „nikdy“ apod.

💡 **Tip: Zkontrolujte, že každá otázka splňuje pět výše uvedených zásad.**

Přehled forem testů

Didaktický test se od „normálního zkoušení“ liší především tím, že je navrhován a interpretován podle předem stanovených pravidel. Stručnou definici uvádí P. Byčkovský: Didaktický test je nástroj systematického zjišťování výsledků výuky.

Typologie didaktických testů

V následujícím textu uvádíme přehled druhů didaktických testů z různých pohledů. Slouží spíše pro představu o šíři tématiky a terminologie, než pro praktickou tvorbu testů samotných.

Tab. 10.3 **Druhy didaktických testů, upraveno volně podle Byčkovského** [133]

Klasifikační hledisko	Druhy testů		
Měřená charakteristika výkonu	Testy rychlosti	Testy úrovně	
Úroveň přípravy testu	Standardizované testy	Kvazistandardizované testy	Nestandardizované testy
Povaha činnosti testovaného	Kognitivní testy	Psychomotorické testy	
Znalosti/schopnosti zjišťované			

testem	Testy výsledků výuky	Testy studijních předpokladů	
Interpretace výkonu	Testy rozlišující (relativního výkonu)	Testy ověřující (absolutního výkonu)	
Časové zařazení do výuky	Testy vstupní	Testy průběžné (formativní)	Testy výstupní (sumativní)
Tematický rozsah	Testy monotematické	Testy polytématické (souhrnné)	
Míra objektivity skórování	Testy objektivně skórovatelné	Testy kvaziobjektivně skórovatelné	Testy subjektivně skórovatelné

Testování handicapovaných studentů a studentů se speciálními vzdělávacími potřebami

Studentům se speciálními potřebami, např. s poruchou zraku, dyslexií či dysgrafií, se upravují podmínky pro psaní testu tak, aby nebyli znevýhodněni oproti ostatním.

Vzhledem k tomu, že pro konkrétního studenta je třeba upravit podmínky řady zkoušek, které postupně absolvuje během studia, je výhodné, aby vysoká škola měla pracoviště, které se této problematice věnuje a dokáže zkoušejícím vydat příslušná doporučení.

Na lékařských fakultách se v praxi setkáváme jen s některými typy handicapů a míra postižení bývá mírná. V dalším textu shrneme proto doporučení věnovaná jen omezenému výběru poruch. Prakticky zpracované další informace čtenář nalezne například v materiálech ke státním maturitám [134], [135] nebo v publikacích organizace Association for Higher Education Access and Disability (http://www.ahead.ie), zejména [136].

Ve velké části případů je vhodné upravit podmínky zkoušky způsobem uvedeným v tabulce 10.4:

Tab. 10.4 Doporučené úpravy podmínek testování pro handicapované studenty

	Dyslexie	**Dysgrafie**	**Poruchy zraku**
Časový limit	Prodloužení časového limitu o 25 %	Prodloužení časového limitu o 25 %, pokud test obsahuje otevřené otázky (pro testy složené jen z výběrových otázek se časový limit neprodlužuje)	Prodloužení časového limitu o 75 %
Formální úpravy zadání	▪ Větší bezpatkový font (např. Arial 14 nebo 16 bodů) ▪ Světle modrý, světle zelený nebo růžový papír ▪ Výrazné oddělení otázek		▪ Větší bezpatkový font (např. Arial 14 až 26 bodů) ▪ Řádkování 1,5 nebo 2 ▪ Zarovnání k levému okraji
Způsob odpovídání na test	▪ Nepoužívat formuláře pro strojové zpracování ▪ Umožnit zapisování odpovědi přímo do zadání namísto do odpovědního archu ▪ Delší odpovědi umožnit psát na počítači nebo zaznamenat zvukově	▪ Nepoužívat formuláře pro strojové zpracování ▪ Umožnit zapisování odpovědi přímo do zadání namísto do odpovědního archu ▪ Delší odpovědi umožnit psát na počítači nebo zaznamenat zvukově	▪ Nepoužívat formuláře pro strojové zpracování ▪ Umožnit zapisování odpovědi přímo do zadání namísto do odpovědního archu ▪ Delší odpovědi umožnit psát na počítači vybaveném software pro zrakově postižené nebo zaznamenat zvukově

Dále může být vhodné, aby student používal při zkoušce vhodné kompenzační pomůcky (např. elektronickou lupu). V každém případě je na jedné straně třeba minimalizovat znevýhodnění studenta se specifickými potřebami oproti intaktní populaci, na druhé straně však musí být pomůcky a úpravy podmínek takové, aby nevznikla neoprávněná výhoda oproti ostatním.

Mnohdy je vhodné, aby byl student se specifickými potřebami zkoušen odděleně od ostatních (např. v jiné místnosti). Je to nutné k tomu, aby nebyl vyrušován např. odcházejícími studenty, pokud mu byl prodloužen časový limit, nebo aby naopak nerušil ostatní, pokud např. část odpovědí nahrává jako zvukový záznam.

Rozhodnutí, že studentovi budou upraveny podmínky pro konání zkoušky, musí být podloženo odpovídajícím odborným posudkem o jeho speciálních potřebách. Posudek by kromě charakteristiky postižení a určení jeho tíže měl obsahovat i popis, jaké konkrétní činnosti jsou postižením ovlivněny, a obecné doporučení, jak při zkoušce omezit jeho dopad.

Stručný přehled nejvýznamnějších forem praktického zkoušení

Pro hodnocení komplexních dovedností v medicíně se používají zkoušky, jejichž součástí je sledování výkonu kandidáta při praktických činnostech. Pro závěrečné zkoušení se hodí především **objektivní strukturované klinické zkoušení** (*Objective Structured Clinical Examination*, **OSCE**). Jde o formát zkoušky, který je poměrně náročný na přípravu i

provedení, umožňuje však objektivní a vzájemně srovnatelné hodnocení studentů. Pro méně významné zkoušky se pak hodí další formáty praktického zkoušení.

Cílem níže uvedeného přehledu je umožnit čtenáři, aby si udělal hrubou představu o hlavních formátech praktického zkoušení, bez nároku na přesnost a úplnost. Zaměření ani rozsah této publikace neumožňují věnovat se této problematice podrobněji, další informace však čtenář nalezne v početné literatuře (např. [137], [138], [139]).

Objektivní strukturované klinické zkoušení (OSCE)

Objektivní strukturované klinické zkoušení (*Objective Structured Clinical Examination*, **OSCE**) je dnes **zlatým standardem** pro zkoušení klinických dovedností, jako jsou komunikace s pacientem, odběr anamnézy, fyzikální vyšetření, provádění některých zákroků, preskripce, hodnocení rentgenových snímků, čtení EKG a mnoha dalších.

Zkouška má několik částí, tzv. **stanic**, kterými student postupně prochází. Absolvování každé stanice trvá obvykle 5–10 minut. Na každé stanici student řeší určitou situaci nebo má splnit nějaký úkol. Je při tom hodnocen jedním či dvěma examinátory. Na rozdíl od „tradičního“ praktického zkoušení examinátor studenta neprovází po celou dobu zkoušky – examinátoři jsou pevně přiřazeni ke stanicím a studenti mezi nimi rotují. Tím se zvyšuje objektivita zkoušky.

Dalším typickým prvkem je zapojení pacientů. Může jít o reálné pacienty, v poslední době se ale stále více upřednostňuje využití tzv. **simulovaných pacientů** – vyškolených a přesně instruovaných herců, kteří dokáží opakovaně reprodukovat modelovou situaci všem studentům, kteří postupně na stanici přicházejí.

Výkon studentů hodnotí examinátor podle předem připravené hodnotící osnovy. Součástí zkoušky ale mohou být i stanice, na nichž student odpovídá písemně a hodnocení probíhá stejně jako u psaných testů (např. psaní receptů, interpretace laboratorních výsledků, čtení EKG křivky, hodnocení rentgenových snímků).

Základní vlastností OSCE je vysoká míra objektivity a strukturovanost zkoušky. Pro jednotlivé stanice se připravují podrobné materiály, díky nimž je zajištěno, že všichni studenti zkoušení v jeden den plní stejně zadané úkoly, jsou známkováni dle stejných kritérií a obecně absolvují zkoušku za velmi srovnatelných podmínek. Zkouší-li se více studentů v několika dnech, stanice se sice liší, ale i v tomto případě se klade velký důraz na standardizaci celé zkoušky a vzájemnou srovnatelnost výsledků.

Zkoušení při práci

Pro další formáty praktického zkoušení se používá anglický termín *Workplace Based Assessment* (**WPBA**). Zkoušený vykonává určitou činnost v reálném klinickém prostředí a je při tom sledován zkoušejícím. Doporučuje se především pro formativní zkoušení; v tom případě je důležitou součástí zkoušky zpětná vazba, kterou vyučující studentovi vhodnou formou poskytne. WPBA se ovšem používá i pro některé typy sumativních zkoušek.

V současné době se používají čtyři formy WPBA.

mini-CEX

Mini-Clinical Evaluation Exercise (**mini-CEX**) je způsob zkoušení vyvinutý původně pro postgraduální vzdělávání ve vnitřním lékařství. Vychází z podstatně rozsáhlejšího zkušebního formátu, *Clinical Evaluation Exercise* (CEX), při němž měl student vyšetřit pacienta a zpracovat jeho případ. Byl při tom sledován jedním zkoušejícím; celá zkouška CEX trvala kolem dvou hodin.

Základní myšlenkou mini-CEX je, že se CEX rozdělí do mnoha krátkých aktů. Každý mni-CEX trvá kolem 15 až 25 minut včetně doby věnované poskytnutí zpětné vazby studentovi učitelem. V průběhu mini-CEX plní student konkrétní, jasně definovaný úkol a je při tom sledován zkoušejícím. Zkouška může probíhat v reálném provozu klinického pracoviště nebo ambulance. V klasickém provedení poté zkoušející hodnotí

- získání anamnézy
- fyzikální vyšetření
- klinické rozhodování a syntézu získaných údajů
- celkový přístup a celkový dojem z výkonu zkoušeného.

Mini-CEX jsou určené především pro formativní zkoušení. Výhodné jsou zejména tehdy, pokud student v průběhu kurzu absolvuje větší počet těchto zkoušek. Významným prvkem je zpětná vazba, kterou zkoušející pokaždé poskytne studentovi bezprostředně po každém zkoušení i s doporučením, jak dále pokračovat ve studiu. Použití mini-CEX při významných zkouškách se považuje za nevhodné.

Přímé sledování procedurálních dovedností (DOPS)

Přímé sledování procedurálních dovedností (*Direct Observation of Procedural Skills*, DOPS) se vyvinulo z mini-CEX a slouží k praktickému zkoušení zákroků a úkonů prováděných na pacientech. V podstatě jde o alternativu k logbookům, kterými se v průběhu studia sleduje, zda se student seznámil s konkrétními úkony. DOPS k čistě kvantitativnímu hodnocení počtu provedených úkonů přidává i kvalitativní hodnocení jejich provedení.

Jak už bylo uvedeno, DOPS principiálně vychází z mini-CEX, takže základní představu o průběhu této zkoušky dává předchozí odstavec. Konkrétní zkušební akty se samozřejmě liší podle zkoušeného úkonu, stejně jako hodnocené dovednosti. Dovednosti, které mají obecnou povahu (např. zachování asepse, komunikace s pacientem) mívají v hodnocení často větší váhu, než úkony specifické pro konkrétní výkon.

Diskuse nad případem (CBD)

Diskuse nad případem (*Case Based Discussion*, CBD) slouží především k hodnocení klinického rozhodování a aplikace medicínských vědomostí. Zkouška probíhá jako diskuse nad chorobopisem nebo dokumentací pacienta. Zkoušený navrhuje diagnosticko-terapeutický postup, zkoušející se přitom ptá na důvody jednotlivých rozhodnutí. Problémem CBD je menší objektivita zkoušky.

Zpětná vazba z několika zdrojů (MSF)

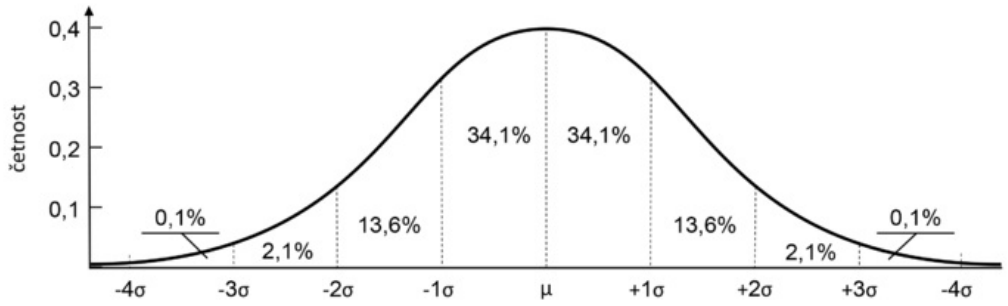
Posledním z častěji používaných formátů praktického zkoušení je zpětná vazba z několika zdrojů (*Multisource Feedback*, MSF, též *360-degree evaluation*). Tento způsob hodnocení se používá v řízení lidských zdrojů k poskytnutí zpětné vazby okolí v porovnání se sebehodnocením pracovníka.

Existuje několik provedení MSF. V zásadě jde vždy o hodnocení založené na dotazníkovém šetření. Kolegové studenta, jeho učitelé, lékaři z oddělení a další zdravotnický, ale i administrativní personál obdrží dotazníky, které se ptají na hodnocení studenta např. v průběhu stáže. Kromě toho student hodnotí i sám sebe. Dotazníky jsou obvykle anonymní a sbírají se buď na papíře, nebo elektronicky. Na základě dotazníkového šetření pak učitel zpracovává celkové hodnocení a poskytuje studentovi zpětnou vazbu, která má sloužit jako podklad pro jeho rozvoj v průběhu dalšího studia.

Statistické nástroje

Normální rozdělení

Stoupáte-li po starých schodech, můžete si všimnout, že schodišťové stupně jsou nejvíc vyšlapané veprostřed. Profil vzniklý opotřebením souvisí s rozdělením náhodných jevů – v tomto případě s náhodným místem, kam na schodišťový stupeň jednotliví lidé stoupnou. Ve většině případů spočine náhodná noha náhodného chodce někde u středu a s klesající pravděpodobností na místě vzdálenějším směrem k okrajům. Po čase se působením mnoha náhodných otěrů schodu vizualizuje křivka hustoty normálního rozdělení – Gaussova křivka.



Obr. 10.2 **Normální rozdělení (Gaussova křivka)**.

Graf znázorňuje hustotu normálního rozdělení se střední hodnotou rovnou μ a směrodatnou odchylkou rovnou σ. Hodnota funkce říká, v jakých oblastech je výsledek náhodného pokusu více pravděpodobný a v jakých méně. Výsledky poblíž střední hodnoty μ jsou pravděpodobnější než odlehlé. Většina výsledků (~ 95,5 %) se vyskytuje v rozmezí dvou směrodatných odchylek od středu tj. od –2σ do 2σ.

Příkladem normálního rozdělení může být například rozdělení vědomostí a dovedností v populaci. Pokud se potřebujeme o normalitě ujistit, tak kromě vizuální kontroly (zda je rozdělení "podobné" normálnímu) můžeme použít některý test normality.

Korelační koeficienty

Pearsonův korelační koeficient měří sílu lineární závislosti mezi dvěma veličinami. Pomůže nám například vyčíslit, jak silná je vazba mezi výsledky ve dvou různých testech, nebo mezi výsledkem v testu a průměrnou známkou na vysvědčení.

> 💡 **Tip: Korelace neznamená kauzalitu**

Předpokládáme, že máme u n jedinců dvojice hodnot $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. Pearsonův korelační koeficient je pak dán vztahem

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}},$$

kde $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ je aritmetický průměr prvních měření a $\bar{Y}$ je aritmetický průměr druhých měření.

Ekvivalentně lze Pearsonův korelační koeficient vyjádřit pomocí součinů *z*-skórů

$$r = \frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i - \bar{X}}{s_X}\right)\left(\frac{Y_i - \bar{Y}}{s_Y}\right)$$

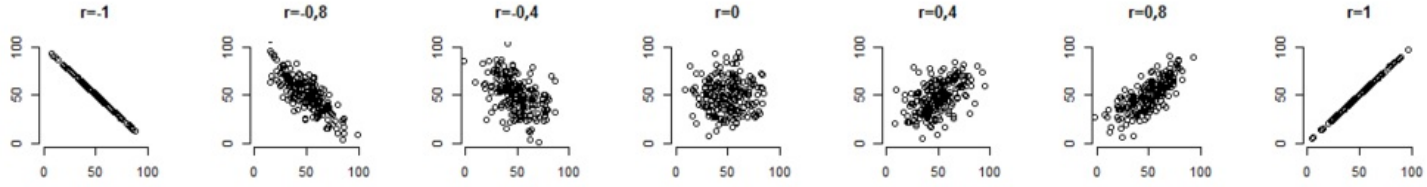
kde $s_X = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}$ je směrodatná odchylka (standard deviation, SD) prvních měření, s_Y směrodatná odchylka druhých měření. *z*-skór $\left(\frac{X_i - \bar{X}}{s_X}\right)$ říká, jak daleko je hodnota X_i od průměru $\bar{X}$, přičemž za jednotku vzdálenosti bereme směrodatnou odchylku s_X.

V případě, kdy jeden ze znaků je binární (např. sledujeme-li závislost binární položky Y a celkového počtu bodů X, tzv. index RIT), se korelační koeficient redukuje na tzv. *bodově-biseriální korelační koeficient*

$$r_{bis} = \frac{\bar{X}_1 - \bar{X}_0}{s}\frac{n_0 n_1}{n(n-1)},$$

kde $\bar{X}_1$ je průměr celkového počtu bodů těch studentů, kteří mají hodnotu sledované položky rovnou 1, $\bar{X}_0$ je průměr celkového počtu bodů těch studentů, kteří mají hodnotu sledované položky rovnou 0, s^2 je směrodatná odchylka celkových bodových zisků, n_1 je počet jedniček a n_0 je počet nul.

Korelační koeficient nabývá pouze hodnot z intervalu od -1 do 1. Svých extrémních hodnot (tedy 1 a -1) nabývá pouze pokud všechny body (X_i, Y_i) leží na jedné přímce. Korelační koeficient je roven 1, pokud je mezi veličinami vztah přímé úměry (tedy čím větší je hodnota jedné veličiny, tím větší je hodnota i druhé veličiny). Pokud je mezi veličinami vztah nepřímé úměry, je korelační koeficient roven -1.



Obr. 10.4 Korelační koeficient nabývající hodnot z intervalu od -1 do 1

Jsou-li veličiny nezávislé, je korelace mezi nimi nulová. Nicméně Pearsonův korelační koeficient je pouze odhad populačního korelačního koeficientu, a při každém výběru *n*-tice studentů vyjde hodnota odhadu nepatrně jiná. Proto i pro nezávislé náhodné veličiny zpravidla vyjde Pearsonův korelační koeficient nenulový. Nulovost pak můžeme testovat pomocí statistického testu (viz např. [141]).

Pearsonův korelační koeficient se pro popis závislosti dvou veličin nehodí vždy. Jeho použití je optimální, pokud jsou veličiny normálně rozdělené (tj. řídí se Gaussovým rozdělením). Není vhodné ho používat, pokud data obsahují odlehlé hodnoty (tedy hodnoty příliš vzdálené od ostatních), neboť ty mohou silně ovlivnit hodnotu Pearsonova koeficientu. Výsledný odhad může být zkreslený také pokud jsou data zešikmená (to nastává např. pokud je test příliš snadný nebo příliš obtížný). Řešením v takovém případě může být použití Spearmanova korelačního koeficientu. Počítá se tak, že do vzorce pro Pearsonův korelační koeficient místo skutečných hodnot $X_1, \ldots X_n$ a $Y_1, \ldots, Y_n$ vložíme jejich pořadí $R_1, \ldots, R_n$ a $Q_1, \ldots, Q_n$. Vzorec Spearmanova korelačního koeficientu lze potom upravit do známějšího vztahu

$$r_S = 1 - \frac{6}{n(n^2-1)}\sum_{i=1}^{n}(R_i - Q_i)^2.$$

Cronbachovo alfa

Cronbachovo alfa je mírou vnitřní konzistence položek, používá se často k odhadu reliability celého testu. Cronbachovo alfa [60], [142], [143] je dáno vztahem:

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k}\mathrm{var}(Y_j)}{\mathrm{var}(Y)}\right),$$

kde k je počet položek testu, $\mathrm{var}(Y_j)$ je rozptyl hodocení *j*-té položky a $\mathrm{var}(Y)$ je rozptyl celkových skóre v testu.

Pro položky typu ano/ne se vzorec Cronbachova alfa zjednoduší na tvar zvaný Kuder-Richardsonova formule 20 dle článku [144]

$$\alpha = \frac{k}{k-1}\left(1 - \frac{\sum_{j=1}^{k} p_j q_j}{\mathrm{var}(Y)}\right),$$

kde p_j je pravděpodobnost správné odpovědi na *j*-tou položku, $q_j = 1 - p_j$ je pravděpodobnost nesprávné odpovědi.

Hodnota rovna jedné nastává, pokud jsou položky svázány lineárně. Malé hodnoty naopak vypovídají o nízké vnitřní konzistenci položek, nebo nízké spolehlivosti testu.

Indexy shody hodnotitelů

Cohenovo kappa [145] měří shodu mezi dvěma hodnotiteli, kteří hodnotí stejnou skupinu *n* studentů. Shodu by bylo možné vyčíslit jednoduše v procentech. Cohenovo kappa je univerzálnější v tom smyslu, že bere v úvahu také pravděpodobnost náhodné shody.

Cohenovo kappa κ je dáno vztahem:

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)},$$

kde Pr(*a*) je relativní shoda mezi hodnotiteli a Pr(*e*) je odhad pravděpodobnosti náhodné shody. Pokud se hodnotitelé shodli v hodnocení všech jedinců, pak κ = 1. Pokud je celkové procento shody rovno pravděpodobnosti očekávané shody při náhodném rozhodování, je κ = 0. Pokud je dokonce procento shody menší, je κ záporné. Interpretace hodnot κ je obvykle následující: Hodnotu κ > 0,75 považujeme za výbornou shodu, κ mezi 0,40 a 0,75 za dobrou shodu a κ < 0,40 považujeme za špatnou shodu.

- Příklad

Představme si, že testujeme dvěma testy 5 studentů a obdržíme tyto výsledky:

Tab. 10.5 Výsledky studentů ve dvou testech

	skóre v prvním testu	skóre v druhém testu
1. student	18	48
2. student	45	75
3. student	33	63
4. student	48	78
5. student	51	81

Pearsonův korelační koeficient je roven 1, mezi dvěma výsledky je přímo lineární závislost (druhý je vždy právě o 30 bodů vyšší než první). Budeme-li však rozhodovat u složení zkoušky na základě 50 bodové hranice, první test úspěšně absolvuje 1 student z 5, zatímco v druhém to budou 4 studenti z 5. Testy by rozhodly o ne/úspěchu shodně jen u prvního a posledního studenta, tedy ve 40 % případů. Jaká je pravděpodobnost náhodné shody testů? První test rozhoduje o úspěchu v 20 % případů a druhý v 80 % případů, náhodné shodné rozhodnutí o úspěchu by tedy nastalo v 20 % · 80 % = 16 % případů, podobně náhodné shodné rozhodnutí o neúspěchu by nastalo v 80 % · 20 % = 16 % případů, celkem tedy náhodná shoda nastává v 32 %. Cohenovo kappa je proto

$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} = \frac{0{,}40 - 0{,}32}{1 - 0{,}32} = 0{,}12,$$

což svědčí o velmi malé shodě mezi dvěma testy.

- Příklad

Představme si situaci, kdy o udělení atestace rozhodují nezávisle dvě komise. Každý atestovaný je hodnocen dvakrát a každá komise rozhodne o udělení atestace buď kladně "Ano", nebo záporně "Ne". Výsledky jsou jako v tabulce 10.6, přičemž řádky odpovídají komisi A a sloupce komisi B:

Tab. 10.6

		B	B	
		Ano	Ne	Celkem
A	Ano	20	5	25
A	Ne	10	15	25
	Celkem	30	20	50

Z celkem 50 atestovaných tedy bylo 20 atestovaných hodnoceno kladně oběma komisemi a 15 atestovaných hodnoceno záporně. Procento shody je tedy $\Pr(a) = (20 + 15)/50 = 0{,}70.$

Abychom odhadli pravděpodobnost náhodné shody $\Pr(e)$, všimněme si nejdříve, že:

- komise A hodnotila kladně 25 atestovaných a záporně také 25 atestovaných. Kladně tedy hodnotila v 50 % případů.
- komise B hodnotila kladně 30 atestovaných a záporně také 20 atestovaných. Kladně tedy hodnotila v 60 % případů.

Proto pokud by komise rozhodovaly náhodně, pravděpodobnost, že obě komise řeknou "Ano" je $0{,}50 \cdot 0{,}60 = 0{,}30$ a pravděpodobnost, že obě řeknou "Ne" je $0{,}50 \cdot 0{,}40 = 0{,}20$. Celková pravděpodobnost náhodné shody je tedy $\Pr(e) = 0{,}30 + 0{,}20 = 0{,}50$.

Po doplnění do vzorce pro Cohenovo kappa dostáváme:

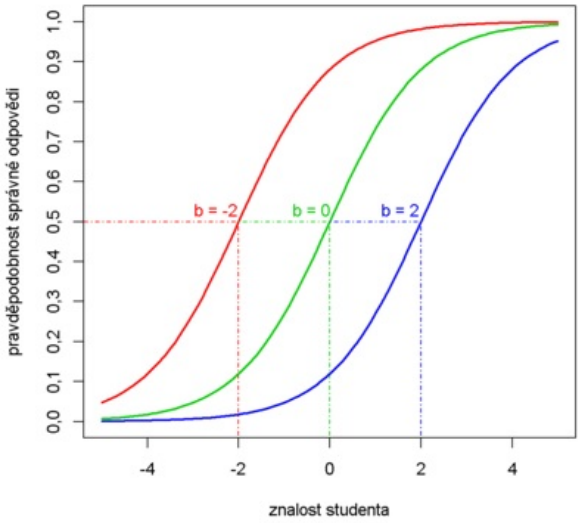
$$\kappa = \frac{\Pr(a) - \Pr(e)}{1 - \Pr(e)} = \frac{0{,}70 - 0{,}50}{1 - 0{,}50} = 0{,}40.$$

Další koeficienty shody

Cohenovo kappa lze rozšířit na případ více kategorií. V případě, kdy nás zajímá více hodnotitelů, je namístě použít Fleissovo kappa [146] [147].

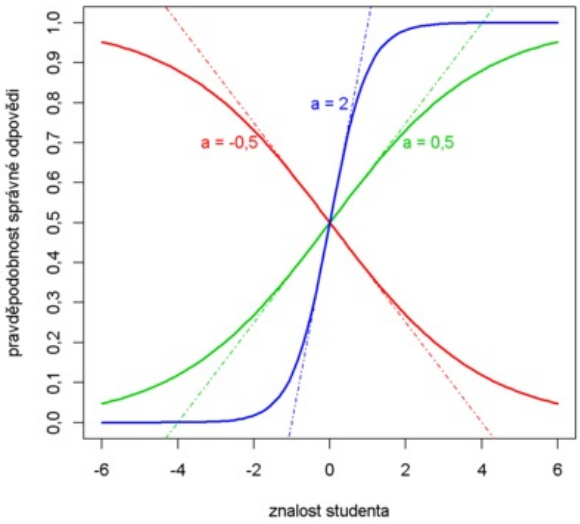
Základní IRT modely

Nejjednodušším IRT modelem je jednoparametrický logistický model. Říká se mu také **Raschův model**, podle dánského matematika George Rasche, který o něm pojednal ve své knize již v roce 1960 [148]. Pravděpodobnost správné odpovědi na položku je v Raschově modelu definovaná jako funkce jedné proměnné – schopnost studenta, a jediného parametru – parametru obtížnosti b. Parametr obtížnosti lze popsat jako úroveň schopnosti, při které student zodpoví položku správně právě s poloviční pravděpodobností. Na grafu 10.5 jsou vyobrazeny charakteristické křivky pro tři různé hodnoty parametru obtížnosti položky:



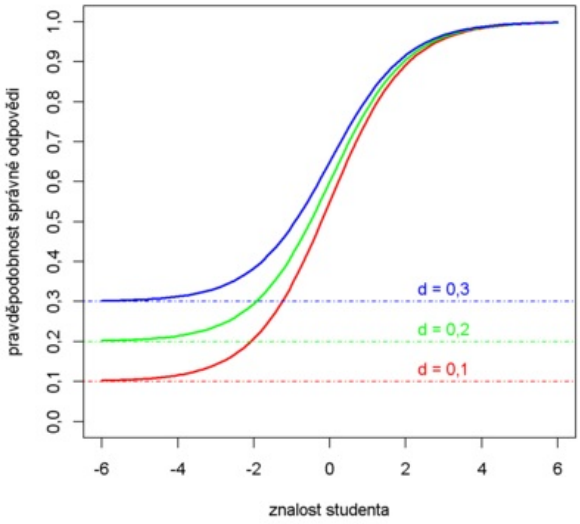
Obr. 10.5 **Raschův model**
Charakteristické křivky tří položek se stejnou diskriminační schopností (sklonem), ale různou obtížností.
——— Charakteristická křivka těžké položky. Poloviční pravděpodobnost správné odpovědi má student s vysokou celkovou znalostí charakterizovanou hodnotou $b = 2$
——— Charakteristická křivka středně obtížné položky. Poloviční pravděpodobnost správné odpovědi má student s průměrnou znalostí
——— Charakteristická křivka snadné položky. Poloviční pravděpodobnost správné odpovědi vykazuje už student s nízkou znalostí charakterizovanou hodnotou $b = -2$

Dvouparametrický logistický model přidává k parametru obtížnosti ještě parametr citlivosti položky a. Ten popisuje sklon charakteristické funkce položky v bodě obtížnosti b. Odhad parametru citlivosti je blízký nule, pokud položka špatně rozlišuje mezi lepšími a slabšími studenty. V případě, kdy slabší studenti odpovídají na položku lépe než lepší studenti, je citlivost položky záporná.



Obr. 10.6 **Charakteristické křivky tří položek se stejnou obtížností, ale různými diskriminačními schopnostmi**
——— Charakteristická křivka položky s menší diskriminační schopností
——— Charakteristická křivka položky s větší diskriminační schopností
——— Charakteristická křivka typická pro distraktor, nebo špatně napsanou položku. Její diskriminační schopnost je záporná. Čím horší student, tím spíš položku označí jako pravdivou.

Dvouparametrický logistický model je vhodný v případě, kdy se nedá očekávat, že odpovědi na položky jsou snadno uhádnutelné. To platí např. pro osobnostní dotazníky, kde žádná odpověď není nesprávná. V případě položek s vícenásobnou odpovědí, kdy právě jedna z m položek je správná, lze ale předpokládat, že i zcela neznalí studenti správnou odpověď alespoň s pravděpodobností $1/m$ uhodnou. V takovém případě má opodstatnění **tříparametrický logistický model**, v němž třetí parametr d vyjadřuje pravděpodobnost toho, že i zcela neznalý student odpoví na položku správně. Na obrázku 10.7 vidíme charakteristické křivky tří položek lišících se v parametru uhádnutelnosti.



Obr. 10.7 ——— nejsnáze uhádnutelná položka ze tří zobrazených – zcela neznalý student na ni odpoví správně s pravděpodobností 0,3
——— na tuto položku zcela neznalý student odpoví správně s pravděpodobností 0,2
——— nejhůře uhádnutelná položka ze tří zobrazených – zcela neznalý student na ni zodpoví správně s pravděpodobností 0,1

Příklady realizace

Automatická tvorba testů na 3. LF UK

V rámci integrovaného předmětu *Buněčné základy medicíny* vytvořil v roce 2011 tým učitelů 3. LF UK systém pro automatizovanou tvorbu testů z předem připravených souborů tvrzení.

Použité testy mají formální strukturu MTF testů, tedy jde o soubor tvrzení, u nichž studenti určují, zda jsou pravdivá či nikoli. Z hlediska teorie testování se nejedná o ideální formu, avšak tato volba byla výsledkem kompromisu mezi vhodností formátu a složitostí přípravy testů. Omezení MTF formátu jsou částečně kompenzována druhou částí testu, která se skládá z úloh typu *short answer*. Konkrétní MTF testy jsou náhodně losovány z textových souborů tvrzení, z nichž každé je označeno číslem oboru, kurzu (podjednotka intergrovaného předmětu), přednášky a písmenem označujícím obtížnost daného tvrzení.

Jednotlivé úlohy v MTF testu jsou složeny vždy ze čtyř tvrzení, které spojuje tématická příbuznost; v případě tohoto předmětu jde vždy o čtyři tvrzení k tématu jedné přednášky. Kmen každé úlohy je naprosto generický, např. *Která z následujících tvrzení jsou správná?* Aby byla zaručena srovnatelná obtížnost testů v různých termínech, jsou jednotlivá tvrzení označena relativní obtížností (A, B nebo C), kterou určí přednášející daného tématu po případné dohodě s dalšími učiteli daného tématu či oboru. Jedna z možností nastavení škály obtížnosti je následující. Otázku C musí znát každý student, jde o tzv. bazál. Obtížnost A je pro výborné studenty, kteří aspirují na nejlepší známku, často se jedná o tvrzení týkající se detailů či vyžadující studium z dodatečných materiálů. Tvrzení s obtížností B jsou potom někde mezi těmito dvěma extrémy – znalost alespoň poloviny z nich (spolu s tvrzeními C) je minimálním předpokladem úspěšného složení zkoušky. V popisovaném případě obsahuje každá úloha jedno tvrzení C, dvě B a jedno A. Tyto úrovně obtížnosti mohou být zhruba namapovány na Angoffovy pravděpodobnosti takto: C=1,0; B=0,75 a 0,5<A<0,75. Označení obtížnosti se používá pouze pro vytvoření testu, studenti je nevidí.

Takovéto rozdělení obtížností za předpokladu jejich realistického přiřazení navíc umožňuje nastavit *a priori* minimální bodový zisk nutný pro složení zkoušky. Student, který zná všechna tvrzení obtížnosti C a polovinu tvrzení B tak získá 50 % bodů a vzhledem k pravděpodobnosti úspěšného natipování má slušnou pravděpodobnost dosáhnout 75 %, což je v daném případě hodnota minimálního počtu bodů pro úspěšné zvládnutí zkoušky. Vzhledem k celkovému počtu 396 tvrzení v tomto testu je pravděpodobnost získání 75 % při znalosti 50 % tvrzení asi 0.53 a prudce stoupá s dodatečnými tvrzeními, které daný student zná.

Jelikož tvrzení C jsou učiteli-experty považována za esenciální, při vyhodnocování testu je jejich neznalost navíc výrazně trestána anulováním celého bodového zisku z dané otázky (tedy souboru čtyř tvrzení). Tak se výrazně snižuje pravděpodobnost, že uspěje student, který sice zná tzv. špeky, ale nemá patřičné základní pochopení či vědomosti.

Testy jsou generovány pomocí skriptu v Perlu podle předem daných pravidel určujících počet úloh z daných oborů a témat. Vygenerované testy jsou pak importovány do programu, který vytvoří varianty a po napsání testu umožní naskenovat odpovědní formuláře a test automaticky opravit.

Zkratky v textech o testování

Abecední seznam zkratek z oboru 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 diskuze.

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/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 zalosti (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

EMQ
: EMQ nebo též EMI (Extended Matching Question/Item) jsou nově zaváděné „rozšířené přiřazovací otázky“. Důvodem pro jejich zavedení je, že MTF svou formou svádějí k biflování, nikoli k pochopení. Charakteristickým znakem EMQ je rozšířená nabídka možností (alespoň osm), z nichž testovaný vybírá (přiřazuje) odpověď. Stejná sada možností se zpravidla používá pro několik scénářů (zadání, otázek). Otázka má být položena tak, aby student byl schopen zformulovat správnou odpověď, aniž by předtím viděl nabízené varianty.

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.

GPA
: Grade Point Average – studijní průměry. Studijní průměry ze střední školy bývají jako dobrý prediktor úspěšnosti studia zařazovány mezi výběrová kritéria pro přijetí na lékařské 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. Toto pořadí poskytuje asi 45 % středních škol. 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.

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

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 1991–2 byl MCAT znovu revidován a restrukturalizován a jeho nová podoba nese výše uvedené označení.

MCQs
Multiple Choice Question – otázka se mnohočetným výběrem odpovědi, nejobecnější formát, který 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) se používá např. v postgraduálním vzdělávání praktického lékařství. Pomocí tohoto typu úloh lze posoudit analytické uvažování, interpretaci nálezů a klinické 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.

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 360-degree 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 formátem 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).

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 5-10 minutová 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 R_{IR}) 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 R_{IT}) označuje korelační koeficient mezi úspěšností v dané testové úloze a celkovým počtem bodů v testu. Koeficient RIT se používá podobně jako index RIR.

SAQ
Otázky s krátkou odpovědí (Short-Answer Questions) tvoří otázka s nabídnutou jednoslovnou (nebo velmi krátkou) odpovědí. Správných odpovědí 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 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 lékařských fakultách. 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 diskuzi 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.

Literatura

1. MILLER, G. E. The assessment of clinical skills/competence/performance. *Academical Medicine.* 1990, vol. 65, no. 9 Suppl, s. 63-7, ISSN 1040-2446. PMID: 2400509 (https://www.ncbi.nlm.nih.gov/pubmed/2400509).
2. ANDERSON, Lorin W, David R KRATHWOHL a Peter W AIRASIAN, et al. *A taxonomy for learning, teaching, and assessing : A revision of Bloom's taxonomy of educational objectives.* 2. vydání. New York : Pearson, 2000. 336 s. ISBN 978-0801319037.
3. BOURSICOT, Katharine. *Principles and theory of good assessment practice: Interactive lecture* [přednáška k předmětu St George's Fundamentals of Assessment Course, obor Medical & Healthcare Education, Population Health Sciences and Education St George's University]. London. 2012-02-23.
4. BULL, Joanna a Myles DANSON. *Computer-assisted Assessment (CAA)* [online] . 1. vydání. Heslington York : Learning and Teaching Support Network, 2004. 24 s. Assessment Series; sv. 14. Dostupné také z <http://www.heacademy.ac.uk/assets/documents/assessment/LTSNassess14_computer_assisted_assessment.pdf>. ISBN 1-904190-53-7.
5. DENNICK, Reg, Simon WILKINSON a Nigel PURCELL. Online eAssessment: AMEE guide no. 39. *Medical Teacher* [online]. 2009, vol. 31, no. 3, s. 192-206, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/19811115>. ISSN 0142-159X (print), 1466-187X (online). PMID: 19811115 (https://www.ncbi.nlm.nih.gov/pubmed/19811115).
6. Educational testing service. *How ETS Develops Test Questions.* [online]. [cit. 2013-04-12]. <http://www.ets.org/s/understanding_testing/flash/how_ets_creates_test_questions.html>.
7. SCHUWIRTH, Lambert WT a Cees PM VAN DER VLEUTEN. General overview of the theories used in assessment: AMEE Guide No.57. *Med Teach* [online]. 2011, roč. -, vol. 3, no. 10, s. 783-797, dostupné také z <https://informahealthcare.com/doi/abs/10.3109/0142159X.2011.611022>. sv. ISBN 978-1-903934-97-5. ISSN 0142-159X. PMID: 21942477 (https://www.ncbi.nlm.nih.gov/pubmed/21942477).DOI: 10.3109/0142159X.2011.611022 (http://dx.doi.org/10.3109%2F0142159X.2011.611022).
8. VAN DER VLEUTEN, C P, G R NORMAN a E DE GRAAFF. Pitfalls in the pursuit of objectivity: issues of reliability. *Med Educ* [online]. 1991, vol. 25, no. 2, s. 110-8, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/2023552>. ISSN 0308-0110.
9. TSAI, Fu-Ju a Hoi K SUEN. A brief report on a comparison of six scoring methods for ultiple true-false items. *Educational and Psychological Measurement.* 1993, roč. 53, s. 399-404, ISSN Print: 0013-1644 Online: 1552-3888. DOI: 10.1177/0013164493053002008 (http://dx.doi.org/10.1177%2F0013164493053002008).
10. BURTON, Richard F a David J MILLER. Statistical Modelling of Multiple-choice and True/False Tests: ways of considering, and of reducing, the uncertainties attributable to guessing. *Assessment and Evaluation in Higher Education.* 1999, vol. 24, no. 4, s. 399-411, ISSN 0260-2938 (Print), 1469-297X (Online). DOI: 10.1080/0260293990240404 (http://dx.doi.org/10.1080%2F0260293990240404).
11. CHANDRATILAKE, Madawa, Margery DAVIS a Gominda PONNAMPERUMA. Assessment of medical knowledge: the pros and cons of using true/false multiple choice questions. *Natl Med J India* [online]. 2011 Jul-Aug, vol. 24, no. 4, s. 225-8, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/22208143>. ISSN 0970-258X.
12. ANDERSON, John. Multiple choice questions revisited. *Med Teach* [online]. 2004, vol. 26, no. 2, s. 110-3, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/15203517>. ISSN 0142-159X.
13. CASE, Susan M a David B SWANSON. *Constructing written test questions for the basic and clinical sciences* [online] . 3. vydání. Philadelfia : National Board of Medical Examiners, 2002. 181 s. Dostupné také z <http://www.nbme.org/publications/item-writing-manual-download.html>.
14. JOLLY, Brian. *Written examinations* [online] . In SAWANWICK, Tim. *Understanding medical education: Theory and practice*. 1. vydání. Oxford : Wiley-Blackwell, 2010. 464 s. s. 208-230. Dostupné také z <http://dx.doi.org/10.1002/9781444320282.ch15>. doi: 10.1002/9781444320282.ch15. ISBN 978-1-4051-9680-2
15. DOWNING, Steven M. Guessing on selected-response examinations. *Med Educ* [online]. 2003, vol. 37, no. 8, s. 670-1, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/12895242>. ISSN 0308-0110.
16. BRENNAN, Liam. *Single Best Answer MCQs* [online]. The Royal College of Anaesthetist, ©2010. Poslední revize 2010, [cit. 2012-01-03]. <http://www.rcoa.ac.uk/docs/sba-questions.pdf>.
17. GAY, Lorraine R. The comparative effects of multiple-choice versus short-answer tests on retention. *Journal of educational measurement* [online]. 1980, vol. 17, no. 1, s. 45-50, dostupné také z <http://onlinelibrary.wiley.com/doi/10.1111/j.1745-3984.1980.tb00813.x/abstract>. ISSN 1745-3984. DOI: 10.1111/j.1745-3984.1980.tb00813.x (http://dx.doi.org/10.1111%2Fj.1745-3984.1980.tb00813.x).
18. CULBERTSON, John C. An Essay Review: The Bell Curve: Class Structure and the Future of America. *Education Policy Analysis Archives* [online]. 1995, vol. 3, no. 2, s. 1-12, dostupné také z <http://epaa.asu.edu/ojs/article/view/645/767>. ISSN 1068-2341.
19. Educational testing service. *ETS Guidelines for Fairness Review of Assessments* [online] . Educational Test ing Service, 2009. Dostupné také z <http://www.ets.org/Media/About_ETS/pdf/overview.pdf>.
20. ALDERSON, J, Caroline CLAPHAM a Dianne WALL. Language test construction and evaluation. New York, NY, USA: Cambridge University Press, 1995, 310 p. ISBN 0-521-47255-5.
21. KOMENDA, Martin a Andrea POKORNÁ. *Benefity a úskalí elektronického testování* [online]. Brno : Masarykova univerzita, 2011, dostupné také z <http://www.mefanet.cz/index.php?pg=publikace-reporty--souborne-prace&prid=31>.
22. TAVAKOL, Mohsen a Reg DENNICK. *Post Examination Analysis of Objective Tests.* 1. vydání. AMEE, 2011. AMEE guide; sv. 54. ISBN 978-1-903934-91-3.
23. AYERS, William. *To teach : The journey of a teacher.* 2. vydání. New York : Teachers College Press, 2001. 151 s. s. 116. ISBN 08-077-3985-5.
24. DAVIDSON, Cathy N. *Now you see it: How the Brain Science of Attention Will Transform the Way We Live, Work and Learn : [object Object].* 1. vydání. Viking Adult. 2011. 342 s. ISBN 9780670022823.
25. URBÁNEK, Tomáš, Denisa DENGLEROVÁ a Jan ŠIRŮČEK. *Psychometrika : Měření v psychologii.* 1. vydání. Praha : Portál, 2011. 320 s. ISBN 978-807-3678-364.
26. BYČKOVSKÝ, Petr. *Základy měření výsledků výuky.* Praha : Výzkumný ústav inženýrského studia při ČVUT, 1982,
27. BECKER, Jasper, Petra ANDĚLOVÁ a Petr BLÁHA. *Čína na přelomu století.* 1. vydání. Praha : BB art, 2002. 393 s. s. 16. Úvod online http://www.rotrekl.cz/becker.htm. ISBN 80-725-7729-8.
28. Wikipedia, Die freie Enzyklopädie. *Chinesische Beamtenprüfung während der Qing-Dynastie* [online]. Poslední revize 2012-11-5, [cit. 2012-12-16]. <https://de.wikipedia.org/w/index.php?title=Chinesische_Beamtenpr%C3%BCfung&oldid=110161210>.
29. ENCYCLOPÆDIA BRITANNICA. ENCYCLOPÆDIA BRITANNICA ONLINE ACADEMIC EDITION,. *Chinese civil service* [online]. Encyclopædia Britannica Inc, ©2012. [cit. 2012-12-12]. <https://www.britannica.com/EBchecked/topic/112424/Chinese-civil-service>.
30. KOUKOLÍK, Tomáš. *Historie úřednických zkoušek v císařské Číně a osmičlenný esej jako literární žánr vyžadovaný u*

30. KOUKOLÍK, Tomáš. *Historie úřednických zkoušek v císařské Číně a usmíchlemný esej jako literární žánr vyžadovaný u zkoušek doby Ming a Qing.* Praha : Univerzita Karlova v Praze, Filozofická fakulta, Ústav dálného východu, 1997, Diplomová práce (Mgr.). Vedoucí diplomové práce David Sehnal
31. HUDDLESTON, Mark W a William W BOYER. *The higher civil service in the United States: quest for reform : [object Object].* 1. vydání. Pittsburgh : University of Pittsburgh Press, 1996. 229 s. ISBN 08-229-5574-1.
32. , Michael Kazin; associate EDITORS a Rebecca EDWARDS, et al. *The Princeton Encyclopedia of American Political History.* 1. vydání. Princeton : Princeton University Press, 2009. 992 s. sv. 2. ISBN 0-691-12971-1.
33. DOLEJŠ, Martin, Michal MIOVSKÝ a Vladimír ŘEHAN. *Testová příručka ke škále osobnostních rysů představujících riziko z hlediska užívání návykových látek : (SURPS - substance use risk profile scale).* 1. vydání. Praha : Klinika adiktologie, 1. lékařská fakulta Univerzity Karlovy v Praze a Všeobecná fakultní nemocnice v Praze ve vydavatelství Togga, 2012. 84 s. ISBN 978-80-87258-81-1.
34. BAUMGARTNEROVÁ, Gabriela a Andrea KAPUSTOVÁ. *Metodický materiál pro hodnotitele písemných prací z českého jazyka a literatury* [online] . Centrum pro zjišťování výsledků vzdělávání, 2013. 37 s. Dostupné také z <http://www.novamaturita.cz/index.php?id_document=1404036293&at=1>.
35. Čeština pro cizince. *Pokyny k organizaci zkoušky z českého jazyka pro trvalý pobyt v ČR*. 2010. Dostupné také z URL <http://cestina-pro-cizince.cz/uploads/Dokumenty/Pokyny%20k%20organizaci.pdf>.
36. Šachový svaz České republiky. *Pokyny k testu trenérů 4. třídy* [online]. chess.cz (Šachový svaz České republiky), [cit. 2013-04-16]. <http://www.chess.cz/www/informace/komise/tmk/dokumenty/test-trener4.html>.
37. BAUMGARTNEROVÁ, Gabriela a Andrea KAPUSTOVÁ. *Jak na to: Písemná práce z českého jazyka* [online] . Centrum pro zjišťování výsledků vzdělávání, 2012. Dostupné také z <http://www.novamaturita.cz/index.php?id_document=1404035884&at=1>. sv. ISBN 978-80-87337-13-4.
38. KOLEN, Michael J, Robert L BRENNAN a Michael J KOLEN. Test equating, scaling, and linking: methods and practices. 2nd ed. New York: Springer, c2004, xxvi, 548 p. ISBN 0-387-40086-9.
39. DAVIER, Alina A. *Statistical models for test equating, scaling, and linking*. New York: Springer, c2011, xix, 367 p. ISBN 978-0-387-98138-3-.
40. JELÍNEK, Martin a Petr KVĚTON. *Testování v psychologii : Teorie odpovědi na položku a počítačové adaptivní testování.* 1. vydání. Praha : Grada, 2011. 160 s. ISBN 978-802-4735-153.
41. Han, K. T. (2009). IRTEQ: Windows application that implements IRT scaling and equating [computer program]. Applied Psychological Measurement, 33(6), 491-493.
42. CHRÁSKA, Miroslav. *Metody pedagogického výzkumu.* 1. vydání. Praha : Grada Publishing a.s., 2007. 265 s. ISBN 80-247-1369-1.
43. BOURSICOT, Katharine. *Introduction to standard setting.* London : St. George's University, 2011.
44. DOWNING, Steven M a Thomas M HALADYNA. *Handbook of test development.* 1. vydání. Mahwah : Lawrence Erlbaum Associates, 2006. 778 s. ISBN 9780805852653.
45. HAMBELTON, Ronald K a Barbara S PLAKE. Using an extended Angoff procedure to set standards on complex performance assessments. *Applied measurement in education.* 1995, roč. 8, vol. 8, no. 1, s. 41-55, ISSN 0895-7347 (Print), 1532-4818 (Online). DOI: 10.1207/s15324818ame0801_4 (http://dx.doi.org/10.1207%2Fs15324818ame0801_4).
46. CANTOR, Jeffrey A. 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.* 1989, roč. 49, vol. 49, no. 3, s. 709-721, ISSN (Print) 0013-1644; (Online) ISSN: 1552-3888.
47. VIOLATO, Claudio, Anthony MARINI a Curtis LEE. A validity study of expert judgement procedures for setting cutoff scores on high stakes credentialing examinations using cluster analysis. *Evaluation and the Health Professions* [online]. 2003, roč. 26, vol. 26, no. 1, s. 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 (https://www.ncbi.nlm.nih.gov/pubmed/22973420).DOI: 10.1177/0163278702250082 (http://dx.doi.org/10.1177%2F0163278702250082).
48. AZIZ, Saman. *A Modified Ebel Standard Setting Method for a Medical School Clinical Skills Assessment.* Chicago : University of Illinois, 2005. 162 s.
49. BUTTERWICK, D. J, D.M PASKEVICH a A.L VALLEVAND, et al. Development of content-valid technical skill assessment instruments for athletic taping skills. *Journal of Allied Health.* 2006, roč. 35, vol. 35, no. 3, s. 149-157, ISSN (Print) 0090-7421, (Online) 1945-404X. PMID: 17036669 (https://www.ncbi.nlm.nih.gov/pubmed/17036669).
50. VIOLATO, Claudio, Lanree SALAMI a Sylvia MUIZNIEKS. Certification Examinations for Massage Therapists: A Psychometric Analysis. *Journal of Manipulative Physiological Therapeutics* [online]. 2002, roč. 25, vol. 25, no. 2, s. 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 (http://dx.doi.org/10.1067%2Fmmt.2002.121413).
51. LAFAVE, M, L KATZ a D.J BUTTERWICK. Development of a content-valid standardized orthopedic assessment tool (SOAT). *Advances in health sciences education : theory and practice.* 2008, roč. 13, vol. 13, no. 4, s. 397-406, ISSN (Print) 1382-4996, (Online) 1573-1677. PMID: 17203268 (https://www.ncbi.nlm.nih.gov/pubmed/17203268).
52. JAMES, Richard. *A comparison of norm-referencing and criterion-referencing methods for determining student grades in higher education* [online]. Centre for the study of higher education, ©2002. [cit. 2013-04-16]. <http://www.cshe.unimelb.edu.au/assessinglearning/docs/AssessingLearning.pdf>.
53. The University of North Carolina at Chappel Hill. *Grading systems* [online] . Center for Faculty Excellence, 2012. 6 s. Dostupné také z <http://cfe.unc.edu/pdfs/FYC10.pdf>.
54. MEZERA, Antonín. *Školní měření a evaluace výsledků vzdělávání ve škole : Studijní materiál pro interní potřebu učitelů základních a středních škol* [online]. [cit. 2012-12-18]. <http://www.pppppraha7a8.cz/files/zaklady%20skolniho%20mereni.pdf>.
55. MCLACHLAN, John C a Susan C WHITEN. Marks, scores and grades: scaling and aggregating student assessment outcomes. *Medical Education* [online]. 2000, roč. 34, vol. 34, no. 10, s. 788-797, dostupné také z <http://doi.wiley.com/10.1046/j.1365-2923.2000.00664.x>. ISSN 0308-0110. DOI: 10.1046/j.1365-2923.2000.00664.x (http://dx.doi.org/10.1046%2Fj.1365-2923.2000.00664.x).
56. BYČKOVSKÝ, Petr a Marie MARKOVÁ. *Využití software ITEMAN k položkové analýze a analýze výsledků testů* [online] . In -. *11. konference ČAPV – Sociální a kulturní souvislosti výchovy a vzdělávání. Sborník referátů [CD-ROM]*. 1. vydání. Brno : Masarykova Univerzita, 2003. Dostupné také z <http://www.ped.muni.cz/capv11/5sekce/5_CAPV_Byckovsky.pdf>.
57. AERA,, APA a NCME. *Standardy pro pedagogické a psychologické testování.* 1. vydání. Praha : Testcentrum, 2001. 320 s. ISBN ISBN 80-86471-07-1..
58. SPEARMAN, Charles. Correlation calculated from faulty data. *British Journal of Psychology.* 1910, roč. 7, vol. 3, no. 3, s. 271–295, ISSN 2044-8295, DOI: 10.1111/j.2044-8295.1910.tb00206.x (http://dx.doi.org/10.1111%2Fj.2044

-8295.1910.tb00206.x).
59. BROWN, William. Some experimental results in the correlation of mental abilities. *British Journal of Psychology.* 1910, roč. 7, vol. 3, no. 3, s. 296–322, ISSN 2044-8295. DOI: 10.1111/j.2044-8295.1910.tb00207.x (http://dx.doi.org/10.1111%2Fj.2044-8295.1910.tb00207.x).
60. CRONBACH, Lee J. Coefficient alpha and the internal structure of tests. *Psychometrika* [online]. 1951, roč. -, vol. 16, no. 3, s. 297–334, dostupné také z <http://psych.colorado.edu/~carey/Courses/PSYC5112/Readings/alpha_Cronbach.pdf>. ISSN (print) 0033-3123 (online) 1860-0980.
61. TAVAKOL, Mohsen a Reg DENNICK. Making sense of Cronbach's alpha. *International Journal of Medical Education* [online]. 2011, roč. -, vol. -, no. 2, s. 53-55, dostupné také z <http://www.ijme.net/archive/2/cronbachs-alpha/>. ISSN 2042-6372. DOI: 10.5116/ijme.4dfb.8dfd (http://dx.doi.org/10.5116%2Fijme.4dfb.8dfd).
62. SIJTSMA, Klaas. On the use, the misuse, and the very limited usefulness of Cronbach's alpha. *Psychometrika* [online]. 2009, roč. -, vol. 74, no. 1, s. 107-120, dostupné také z <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2792363/>. ISSN (print) 0033-3123 (online) 1860-0980. PMID: 20037639 (https://www.ncbi.nlm.nih.gov/pubmed/20037639).DOI: 10.1007/s11336-008-9101-0 (http://dx.doi.org/10.1007%2Fs11336-008-9101-0).
63. ZINBARG, Richard E, William REVELLE a Iftah YOVEL, et al. Cronbach's alpha, Revelle's beta, and McDonald's omega: Their relations with each other and two alternative conceptualizations of reliability. *Psychometrika* [online]. 2005, roč. -, vol. 70, no. 1, s. 123-133, dostupné také z <http://personality-project.org/revelle/publications/zinbarg.revelle.pmet.05.pdf>. ISSN (print) 0033-3123 (online) 1860-0980. DOI: 10.1007/s11336-003-0974-7 (http://dx.doi.org/10.1007%2Fs11336-003-0974-7).
64. CRONBACH, Lee J, Nageswari RAJARATNAM a Goldine C GLESER. Theory of generalizability: A liberation of reliability theory. *The British Journal of Statistical Psychology.* 1963, roč. 17, vol. 16, no. 2, s. 137-163, ISSN 2044-8317. DOI: 10.1111/j.2044-8317.1963.tb00206.x (http://dx.doi.org/10.1111%2Fj.2044-8317.1963.tb00206.x).
65. SHAVELSON, Richard J a Noreen M WEBB. *Generalizability Theory: A Primer : [object Object].* 1. vydání. Newbury Park, Calif : Sage Publications, 1991. 152 s. ISBN 978-0803937451.
66. MARTINKOVÁ, Patrícia a Karel ZVÁRA. Reliability in the Rasch model. *Kybernetika* [online]. 2007, roč. -, vol. 43, no. 3, s. 315-326, dostupné také z <http://dml.cz/bitstream/handle/10338.dmlcz/135776/Kybernetika_43-2007-3_4.pdf>. ISSN 0023-5954.
67. MARTINKOVÁ, Patrícia a Karel ZVÁRA. Reliability of Composite Dichotomous Measurements. *European Journal for Biomedical Informatics* [online]. 2010, roč. -, vol. 6, no. 2, s. 14-23, dostupné také z <http://www.ejbi.org/en/ejbi/article/77-en-reliability-of-composite-dichotomous-measurements.html>. ISSN 1801-5603.
68. ZVÁRA, Karel. Měření reliability aneb Bacha na Cronbacha. *Informační bulletin České statistické společnost* [online]. 2002, roč. -, vol. 13, no. 2, s. 13-20, dostupné také z <http://www.statspol.cz/bulletiny/ib-02-2.pdf>. ISSN 1210-8022.
69. SCHINDLER, Radek. Rukověť autora testových úloh. Vyd. 1. Praha: centrum pro zjišťování výsledků vzdělávání, 2006, 86 s. ISBN 80-239-7111-5., on-line:http://www.cermat.cz/rukovet-autora-testovych-uloh-1404034186.html
70. ŠTUKA, Čestmír, Patrícia MARTINKOVÁ a Karel ZVÁRA, et al. The prediction and probability for successful completion in medical study based on tests and pre-admission grades. *The New Educational Review* [online]. 2012, roč. -, vol. 28, no. 2, s. 138-152, dostupné také z <http://www.educationalrev.us.edu.pl/vol/tner_2_2012.pdf>. ISSN 1732-6729.
71. ZVÁRA, Karel. *Regrese.* 1. vydání. Praha : MATFYZPRESS, vydavatelství Matematicko-fyzikální fakulty Univerzity Karlovy v Praze, 2008. 254 s. ISBN 978-80-7378-041-8.
72. BYČKOVSKÝ, Petr a Karel ZVÁRA. *Konstrukce a analýza testů pro přijímací řízení.* 1. vydání. Praha : Univerzita Karlova v Praze, Pedagogická fakulta, 2007. 79 s. ISBN 978-80-7290-331-3.
73. FILÍPKOVÁ, Zuzana a Petr BYČKOVSKÝ. *Studie proveditelnosti počítačem adaptovaného testování v prostředí českých škol* [online] . Systémový projekt Kvalita I: CERMAT, 2008. 26 s. Dostupné také z <http://www.esf-kvalita1.cz/Vystupy_projektu/1A1U2_Osobni%20portfolio%20zaka/cat/Studie_CAT_2008.pdf>. sv. [cit. 2012-11-08].
74. GIERL, Mark J, Hollis LAI a Simon R TURNER. Using automatic item generation to create multiple-choice test items. *Medical Education* [online]. 2012, roč. -, vol. 46, no. 8, s. 757-765, dostupné také z <http://onlinelibrary.wiley.com/doi/10.1111/j.1365-2923.2012.04289.x/full>. sv. DOI:. ISSN 1365-2923. PMID: 22803753 (https://www.ncbi.nlm.nih.gov/pubmed/22803753).DOI: 10.1111/j.1365-2923.2012.04289.x (http://dx.doi.org/10.1111%2Fj.1365-2923.2012.04289.x).
75. BREITHAUPT, Krista, Adelaide A ARIEL a Donovan R HARE. *Assembling an inventory of multistage adaptive testing systems* [online] . In van der Linden, Wim J.; Glas, Cees A.W. *Elements of Adaptive Testing*. 1. vydání. New York : Springer, 2010. s. 247-266. Dostupné také z <http://link.springer.com/chapter/10.1007%2F978-0-387-85461-8_13#page-2>. DOI:10.1007/978-0-387-85461-8_13. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8
76. JELÍNEK, Martin, Petr KVĚTON a Denisa DENGLEROVÁ. Adaptivní testování - základní pojmy a principy. *Československá psychologie.* 2006, roč. 50, vol. -, no. 2, s. 163-173, ISSN (Print) 0009-062X; (Online) 1804-643.
77. MAGIS, David. *A small overview of available computer software to support computerized adaptive testing*. Příspěvek na konferenci 15th biennal conference of the European Association for Research on Learning and Instruction. Mnichov. 27- - 31. srpen 2013. Dostupné také z <http://orbi.ulg.ac.be/handle/2268/145056>.
78. MAGIS, David a Gilles RAICHE. Random Generation of Response Patterns under Computerized Adaptive Testing with the R Package catR. *Journal of Statistical Software* [online]. 2012, roč. -, vol. 48, no. 8, s. 1-31, dostupné také z <http://www.jstatsoft.org/v48/i08/paper>. ISSN 1548-7660.
79. KVĚTON, Petr, Martin JELÍNEK a Denisa DENGLEROVÁ, et al. Software pro adaptivní testování: CAT v praxi. *Československá psychologie.* 2008, roč. 52, vol. -, no. 2, s. 145-154,
80. VÁŇOVÁ, Tamara, Jiří PROCHÁZKA a Denisa DENGLEROVÁ. *Adaptivní test COMPACT.* 1. vydání. Brno : Masarykova univerzita, 2012. 99 s. ISBN 978-80-210-5742-5.
81. EDITOR, Brennan, Robert L. *Educational measurement.* 4. vydání. Westport, Ct : Praeger Publishers, c2006. sponsored jointly by National Council on Measurement in Education and American Council on Education. ISBN 02-759-8125-8.
82. HIGGINS, Colin A a Brett BLIGH. Formative computer based assessment in diagram based domains. *ACM SIGCSE Bulletin.* 2006, roč. -, vol. 38, no. 3, s. 98-102, ISSN 0097-8418. DOI: 10.1145/1140123.1140152 (http://dx.doi.org/10.1145%2F1140123.1140152).

83. Online Education Database (OEDb). *8 Astonishing Stats on Academic Cheating* [online]. ©2010. Poslední revize 2010-12-19, [cit. 2012-11-25]. <http://oedb.org/library/features/8-astonishing-stats-on-academic-cheating>.
84. PITONIAK, Mary J. *item generation methodology in theory and practice.* Amherst : University of Massachusetts, School of Education, 2002. Center for Educational Assessment Research Report No. 468;
85. SHU, Zhan. *Detecting test cheating using a deterministic, gated item response theory mode* [online]. Greensboro : The University of North Carolina, 2010, dostupné také z <http://libres.uncg.edu/ir/uncg/listing.aspx?id=4823>. Vedoucí disertační práce Dr. Richard Luecht
86. YANG, Yongwei. *Exposed Items Detection in Personnel Selection Assessment: An Exploration of New Item Statistic* [online]. Annual meeting of the National Council of Measurement in Education in Chicago, ©2007. [cit. 2013-04-17]. <http://www.measuredprogress.org/documents/10157/19213/ExposedItemDetection.pdf>.
87. VEERKAMP, Wim.J.J a Cees.A.W GLAS. Detection of known items in adaptive testing with a statistical quality control method. *Journal of Educational and Behavioral Statistics.* -, roč. 25, vol. 4, s. 373-389, ISSN (Print) 1076-9986, (Online) 1935-1054.
88. HAN, Ning a Ronald HAMBELTON. *Detecting exposed test items in computer-based testing* [online]. Paper Presented at the Annual Meeting of the National Council on Measurement in Education, [cit. 2013-04-17]. <http://www.psych.umn.edu/psylabs/catcentral/pdf%20files/ha04-01.pdf>.
89. GEORGIADOU, Elissavet G, Evangelos TRIANTAFILLOU a Anastasios A ECONOMIDES. A review of item exposure control strategies for computerized adaptive testing developed from 1983 to 2005. *Journal of Technology, Learning, and Assessment* [online]. 2007, roč. -, vol. 5, no. 8, s. -, dostupné také z <http://escholarship.bc.edu/ojs/index.php/jtla/article/view/1647>. ISSN 1540-2525.
90. FSBPT. *Forensic Analysis Conducted to Investigate Effect of Trafficking in Recalled Test Items Leads to Invalidation of 20 Candidate Test Scores* [online]. The Federation of State Boards of Physical Therapy, ©2012. [cit. 2012-12-13]. <https://www.fsbpt.org/forfaculty/yourquestions/index.asp#InvalidatedNPTEScores>.
91. POLLACK, Robert. *Self-grading multiple-choice Tests With Google Docs* [online]. Rpollack.net, Poslední revize 2008-09-17, [cit. 2012-12-12]. <http://rpollack.net/2008/09/self-grading-multiple-choice-tests-with-google-docs/>.
92. POULOVÁ, Petra a Hana ŠRÁMKOVÁ. *Vytváření a management testů a projektů* [online]. Fakulta informatiky a managementu Univerzity Hradec Králové, ©2008. Poslední revize 2008-01-14, [cit. 2012-12-03]. <http://fim.uhk.cz/oliva/tvorba_vedeni/REKAP-www/M2.pdf>.
93. ŠAFRÁNKOVÁ, Hana a Petra POULOVÁ. *WebCT na Univerzitě Hradec Králové* [online]. Alternativní metody výuky 2008: 6. ročník konference, ©2008. [cit. 2013-04-17]. <http://everest.natur.cuni.cz/konference/2008/prispevek/safrankova.pdf>.
94. KORVINY, Petr, Roman FOLTYN a Robert KEMPNÝ. *LMS Moodle na více serverech* [online] . 1. vydání. 2009. 443 s. s. 239-244. Dostupné také z <http://korviny.cz/clanky_pdf/smm2009-korviny_foltyn_kempny-clanek.pdf>. Proceedings of International Conferences: ICT Bridges, Sunflower 2009, Silesian Moodle Moot 2009. ISBN 978-80-248-2117-7
95. KORVINY, Petr a Roman FOLTYN. *LMS Moodle v clusteru.* In EUNIS-CZ. *Open Source na vysokých školách: sborník příspěvků ke konferenci : Špindlerův Mlýn 23.-25.9.2012*. 1. vydání. Západočeská univerzita, 2012. 71 s. ISBN 8026101499, 9788026101499
96. WILSON, Scott. *Rogō: an open source solution for high-stakes assessment* [online]. OSS Watch team blog: open source software advisory service, ©2012. [cit. 2012-12-09]. <http://osswatch.jiscinvolve.org/wp/2012/09/13/rogo-an-open-source-solution-for-high-stakes-assessment/>.
97. BAYLEM, N.J, S WILKINSON a R DENNICK. Would the MRCS Written Papers Benefit from Computerisation? The University of Nottingham Experience. *Bulletin of The Royal College of Surgeons of England* [online]. 2011, roč. -, vol. 93, no. 1, s. 1-5, dostupné také z <http://docserver.ingentaconnect.com/deliver/connect/rcse/14736357/v93n1/s24.pdf?expires=1366271393&id=73830846&titleid=6331&accname=1.LF+UK+-+Ustav+vedeckych+informaci&checksum=D5E75025CEC1A5B4BD83B83C9590CF74>. ISSN 14736357. DOI: 10.1308/147363511X546545 (http://dx.doi.org/10.1308%2F147363511X546545).
98. ZVÁROVÁ, J a K ZVÁRA. *Evaluation of Knowledge using ExaMe program on the Internet.* In Iakovidis, I., Maglavera, S., Trakatellis, A. *User Acceptance of Health Telematics Applications*. 1. vydání. Amsterdam : IOS Press, 2000. 197 s. sv. 72. s. 145-151. DOI: 10.3233/978-1-60750-916-5-145. ISBN 978-90-5199-415-5
99. DOSTÁLOVÁ, Taťjana a Michaela SEYDLOVÁ, et al. *Dentistry and oral diseases : for medical students.* 1. vydání. Praha : Grada, 2010. 208 s. ISBN 978-80-247-3005-9.
100. MARTINKOVÁ, Patrícia, Karel Jr ZVÁRA a Jana ZVÁROVÁ, et al. The new features of the ExaMe evaluation system and reliability of its fixed tests. *Methods of information in medicine.* 2006, roč. 45, vol. -, no. 3, s. 310-315, ISSN 0026-1270.
101. ŠTUKA, Čestmír. *Úspěšnost studia z pohledu moderních metod analýzy dat.* Praha : 1. LF UK v Praze, 2012, Vedoucí disertační práce prof. MUDr. Štěpán Svačina, DrSc
102. HRAD, Miroslav. *vytváření, zpracování a vyhodnocení písemných testů.* Brno : Masarykova univerzita v Brně, Fakulta informatiky, 2001, Diplomová práce
103. GRAVIC, Inc. *Remark Test Grading Edition* [online]. [cit. 2013-02-18]. <http://www.remarktestgrading.com/>.
104. GRAVIC, Inc. *Remark Office OMR Software* [online]. [cit. 2013-02-18]. <http://www.gravic.com/remark/officeomr/>.
105. IBM. *SPSS Data Collection Paper* [online]. [cit. 2013-02-18]. <http://www-142.ibm.com/software/products/us/en/spss-paper/>.
106. ABBYY. *ABBYY FlexiCapture* [online]. [cit. 2013-02-18]. <http://www.abbyy.com/data_capture_software/>.
107. KODAK. *Kodak: Document Imaging* [online]. [cit. 2013-02-18]. <http://graphics.kodak.com/DocImaging/GB/en/Products/Software/Document_Scanning_Software/Capture_Pro_Software/
108. SCANSERVICE. *Scanservice: End of Entropy* [online]. [cit. 2013-02-18]. <http://scanservice.cz/>.
109. EPSON. *Capture Pro: Free versatile document scaninng software from Epson* [online]. [cit. 2013-02-18]. <http://assets.epson-europe.com/gb/en/document_capture_pro/index.html />.
110. HORÁKOVÁ, Monika. *Produkty pro digitalizaci formulářů dostupné na českém trhu* [online]. Praha : VŠE, 2008/9, dostupné také z <sys.kx.cz/4it333/digitalizace_formularu.doc>. Seminární práce
111. REJNKOVÁ, Petra. *Produkty pro digitalizaci formulářů dostupné na českém trhu* [online]. Praha : VŠE, 2009, dostupné také z <sys.kx.cz/4it333/2tym_4tema_Rejnkova.doc>. Seminární práce
112. ACREA CR. *Vyhodnocování testů a přijímacích řízení na ČZU : Česká zemědělská univerzita zvýšila efektivitu při zpracování testů pomocí Remark Office* [online]. ©2012. [cit. 2012-12-04]. <http://www.acrea.cz/upload/download/casestudy/czu.pdf>.
113. HRAD, Miroslav. *Automatizace vytváření, zpracování a vyhodnocení písemných testů.* Brno : Masarykova univerzita v Brně, Fakulta informatiky, 2001

v Brne, Fakulta informatiky, 2001,
114. HRAD, Miroslav a Petr SOJKA. Automatizace sazby a skenování formulářů. *Zpravodaj Československého sdružení uživatelů TEXu* [online]. 2002, roč. 12, vol. -, no. 3-4, s. 123-139, dostupné také z <http://www.fi.muni.cz/usr/sojka/papers/sltpap02.pdf>. ISSN (Print) 1211-6661, (Online) 1213-8185.
115. Fakulta Informatiky Masarykovy Univerzity. Skenování písemek. *Elportál* [online]. [cit. 2013-02-19], roč. -, vol. -, s. -, dostupné také z <http://is.muni.cz/elportal/tipy/zkouseni_skenovanim.pl>. ISSN 1802-128X.
116. DVOŘÁK, Petr, et al. Zkoušení skenovatelnými odpovědními listy. *Elportál* [online]. 2006, [cit. 2013-02-19], roč. -, vol. -, s. -, dostupné také z <http://is.muni.cz/elportal/zkusenosti/zkouseni_skenovanim_biologie.pl>. ISSN 1802-128X.
117. RUDNER, Lawrence M.. *Implementing the graduate management admission test computerised adaptive test* [online] . In van der Linden, Wim J.; Glas, Cees A.W. *Elements of Adaptive Testing*. 1. vydání. New York : Springer, 2010. s. 151-165. Dostupné také z <http://link.springer.com/search?query=Implementing+the+graduate+management+admission+test+computerised+adaptive+test#page-1>. ISBN (Print) 978-0-387-85459-5, (Online) 978-0-387-85461-8
118. (EDITOR), Darren K Patten, David Layfield (EDITOR) a Shobit Arya (EDITOR), et al. *Single best answers in surgery.* 1. vydání. London : Hodder Arnold, 2009. 448 s. ISBN 03-409-7235-1.
119. (EDITOR), Kieran Walsh a Sir Donaldson (FOREWORD). *Cost effectiveness in medical education.* 1. vydání. Oxford : Radcliffe Pub, 2010. 154 s. ISBN 18-461-9410-5.
120. DRASGOW, Fritz, Richard M LUECHT a Randy E BENNETT. *Technology and testing.* In Brennan, Robert L. *Educational measurement*. 4. vydání. Praeger Publishers, 2006. 779 s. Washington, DC: American Council on Education. ISBN 0275981258, 9780275981259
121. ALVES, Cecilia B, Mark J GIERL a Hollis LAI. *Using automated item generation to promote principled test design and development* [online]. Paper presented at the annual meeting of the American Educational Research Association, Denver, CO, ©2010. [cit. 2013-04-18]. <http://www2.education.ualberta.ca/educ/psych/crame/files/AERA%202010%20Denver%20Task%20Model%20AIG.pdf>.
122. GIERL, Mark J a Thomas M HALADYNA. *Automatic item generation: Theory and practice.* 1. vydání. New York : Routledge, 2012. 256 s. ISBN 978-0-415-89750-1.
123. ANZALDUA, Ric M. *Item banks: Where, why, and how* [online] . Austin, TX : Education Resources Information Center on May 12, 2005, 2002. Dostupné také z <http://www.eric.ed.gov/ERICDocs/data/ericdocs2/content_storage_01/000000 0b/80/0d/c8/e2.pdf>. sv. 25th annual meeting of the Southwest Educational Research.
124. Cermat, Banka testavých úloh (BTÚ). *Systémový projekt KVALITA I* [online]. ©2008. [cit. 2012-11-08]. <http://www.esf-kvalita1.cz/Vystupy_projektu/vystupy.php>.
125. VALE, C.D. *Computerized item banking.* In Downing, Steven M, Haladyna, Thomas M. *Handbook of test development*. 1. vydání. New York : Routledge, 2006. ISBN 0805852654, 9780805852653
126. WEISS, David. *Item Banking, Test Development, and Test Delivery* [online] . In GEISINGER, Kurt F a Bruce A BRACKEN. *APA handbook of testing and assessment in psychology*. 1. vydání. Washington : American Psychological Association, 2013. Dostupné také z <http://www.assess.com/docs/Weiss_Handbook_Chapter.pdf>. ISBN 978-1-4338-1227-9
127. Wikipedia. *Item bank* [online]. Wikipedia, the free encyclopedia, ©2012. [cit. 2012-12-21]. <https://en.wikipedia.org/w/index.php?title=Item_bank&oldid=527670124>.
128. FREEMAN, Adrian, Anthony NICHOLLS a Chris RICKETTS, et al. Can we share questions? Performance of questions from different question banks in a single medical school. *Med Teach* [online]. 2010, vol. 32, no. 6, s. 464-6, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/20515373>. ISSN 0142-159X (print), 1466-187X.
129. IDEAL Consortium: Our Current Members. IDEAL CONSORTIUM. [online]. 2013 [cit. 2013-06-06]. Dostupné z: http://temporary.idealmed.org/wp/membership/our-current-members/
130. HOCHLEHNERT, Achim, Konstantin BRASS a Andreas MÖLTNER, et al. Good exams made easy: the item management system for multiple examination formats. *BMC Med Educ* [online]. 2012, vol. 12, s. 63, dostupné také z <http://www.biomedcentral.com/1472-6920/12/63>. ISSN 1472-6920. PMID: 22857655 (https://www.ncbi.nlm.nih.gov/pubmed/22857655).DOI: 10.1186/1472-6920-12-63 (http://dx.doi.org/10.1186%2F1472-6920-12-63).
131. Gesellschaft für Medizinische Ausbildung. Leitlinie für Fakultäts-interne Leistungsnachweise während des Medizinstudiums: Ein Positionspapier des GMA-Ausschusses Prüfungen und des Kompetenzzentrums Prüfungen Baden-Württemberg. *GMS Zeitschrift für Medizinische Ausbildung* [online]. 2008, roč. -, vol. 25, no. 1, s. -, dostupné také z <http://www.egms.de/static/en/journals/zma/2008-25/zma000558.shtml>. sv. Doc74. ISSN 1860-3572.
132. KVAŠŇÁK, Eugen a Mikuláš GANGUR. *Testing portal of medical biophysics questions.* Brno : Mefanet, 2012. sv. Příspěvek na konferenci Mefanet 2012.
133. CHRÁSKA, Miroslav. *Didaktické testy, příručka pro učitele a studenty učitelství.* 1. vydání. Brno : Paido, 1999. 91 s. ISBN 80-85931-68-0.
134. Cermat. *Maturita bez handicapu* [online]. ©2010. [cit. 2013-03-10]. <http://www.novamaturita.cz/maturita-bez-handicapu-1404033473.html>.
135. Cermat. *Maturita bez handicapu : Dokumenty ke stažení* [online]. ©2010. [cit. 2013-03-10]. <http://www.novamaturita.cz/dokumenty-ke-stazeni-1404034076.html>.
136. LOFTUS, Trudy. *Supporting students with dyslexia : Practical guidelines for institutions of further and higher education* [online] . 1. vydání. Dublin : AHEAD Educational Press, 2009. 55 s. Dostupné také z <http://www.ahead.ie/documents/Dyslexia_Handbook_09_Online.pdf>. ISBN 978-1-899951-20-8.
137. NORCINI, John a Vanessa BURCH. Workplace-based assessment as an educational tool: AMEE Guide No. 31. *Med Teach* [online]. 2007, vol. 29, no. 9, s. 855-71, dostupné také z <https://www.ncbi.nlm.nih.gov/pubmed/18158655>. ISSN 0142-159X (print), 1466-187X.
138. Academy of Medical Royal Colleges. . *Improving Assessment* [online] . 1. vydání. London. 2009. 48 s. Dostupné také z <http://www.aomrc.org.uk/publications/statements/doc_details/49-improving-assessment.html>.
139. SELBY, C, L OSMAN a M DAVIS, et al. Set up and run an objective structured clinical exam. *BMJ* [online]. 1995, vol. 310, no. 6988, s. 1187-90, dostupné také z <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC2549563/?tool=pubmed>. ISSN 0959-535X (print), 0959-8138.
140. SIES, Helmut. A new parameter for sex education. *Nature.* 1988, roč. -, vol. -, no. 332, s. 495, ISSN (Print) 0028-0836, (Online) 1476-4687. DOI: 10.1038/332495a0 (http://dx.doi.org/10.1038%2F332495a0).
141. ZVÁRA, Karel. *Základy statistiky v prostředí R.* 1. vydání. Praha : Karolinum, 2013. 249 s.
142. CRONBACH, Lee J. My Current Thoughts on Coefficient Alpha and Successor Procedures. *Educational and Psychological Measurement.* 2004, roč. 64, vol. -, no. 3, s. 391-418. ISSN (Print) 0013-1644, (Online) 1552-3888.

DOI: 10.1177/0013164404266386 (http://dx.doi.org/10.1177%2F0013164404266386).
143. NUNNALLY, Jum C. *Psychometric theory.* 2. vydání. McGraw-Hill, 1978. 701 s. ISBN 0070474656, 9780070474659.
144. KUDER, G.F a M.W RICHARDSON. The theory of the estimation of test reliability. *Psychometrika.* 1937, roč. -, vol. 2, no. 3, s. 151-160, ISSN (Print) 0033-3123, (Online) 1860-0980.
145. COHEN, Jacob. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement.* 1960, roč. -, vol. 20, no. 1, s. 37-46, ISSN (Print) 0013-1644, (Online) 1552-3888.
146. FLEISS, Joseph L. Measuring nominal scale agreement among many raters. *Psychological Bulletin.* 1971, roč. -, vol. 76, no. 5, s. 378-382, ISSN (Print) 0033-2909, (Online) 1939-1455.
147. FLEISS, Joseph L. *Statistical methods for rates and proportions.* 2. vydání. New York : Wiley, 1981. ISBN 9780471064282.
148. RASCH, George. *Probabilistic models for some intelligence and attainment tests.* 1. vydání. Copenhagen : Danish Institute for Educational Research, 1960. ISBN 000-000-00-0.

Index

Omr
Oponentura otázek
:: P
Papírové testování
Paralelní formy testu
Pearsonův korelační koeficient
Percentil
Percentilová škála
Pilotní testování
Položková banka
Počítačem podporované testování
Počítačové testování
Pretest
:: Q
Qti
Questionmark
:: R
Raschův model
Redakční revize
Relativní klasifikace
Relativní standardizace
Reliabilita
Revize férovosti
Rogo
:: S
Směrodatná odchylka
Spearmanova korelačního koeficientu
Standardizace testu
Subjektivní zpětná vazba
:: T
Teorie odpovědi na položku
Test-retest
Testový sešit
Touchstone
Týmové spolupráci
:: U
Upper-lower index
:: V
Validita
Validita inkrementální
Validita kriteriální
Validita obsahová
Validita položky
Validita predikční
Validita souběžná
Vyrovnávání obtížnosti testů
:: Z
Z-skór
:: Š
Šikmost

Doslov

Testování znalostí je jednou z cest k objektivnějšímu posuzování výsledků studia. Kvalita samotného testování je proto ve světě pečlivě sledována a studována. Rámec kvalitnímu testování znalostí zajišťují významné pedagogické a psychologické asociace American Educational Research Association (AERA (http://www.aera.net/)), American Psychological Association (APA (http://www.apa.org/)) a National Council on Measurement in Education (NCME (http://ncme.org/)), mimo jiné formou vydávání Standardů pro pedagogické a psychologické testování (http://www.apa.org/science/programs/testing/standards.aspx/).

Jedním z oborů, v němž se testování znalostí s úspěchem používá, je i moderní výuka medicíny. Velkou pozornost proto tomuto tématu věnují i Association for Medical Education in Europe (AMEE (http://www.amee.org/)), nebo v USA působící National Board of Medical Examiners (NBME (http://www.nbme.org)). Zmíněné asociace pořádají konference, vydávají odborné časopisy, knihy a také samostatná odborná doporučení, která odrážejí stav poznání v tomto oboru.

Tato publikace vznikla s cílem přispět k formování metodiky testování na českých a slovenských lékařských fakultách. Práce byla inspirována a zastřešena sdružením lékařských fakult MEFANET, které se na témata elektronické podpory výuky a spolupráci lékařských fakult zaměřuje.

Kniha je společným dílem autorů. Již sama spolupráce na tomto tématu byla pro autory přínosem a v této metodě vidí cestu jak téma dále kultivovat. Věříme, že tento text přispěje ke zvýšení zájmu o kvalitu didaktických testů na lékařských fakultách.