Tvorba a analýza didaktických testů

- povinně volitelný kurz z bloku didaktické aspekty vzdělávání pro studenty navazujícího magisterského studia
- **rozsah kurzu**: 1/1
- **výuka**: čtvrtek 16:15-17:45, výuka odpadá 12.12. a 19.12. (zahraniční cesta)
- **zakončení kurzu**: klasifikovaný zápočet
- poznámka: docházka na kurz není povinná, účast si nezapisuji doporučení: k porozumění látky kurzu je však pro většinu studentů velmi vhodné sledovat výklad, bude demonstrováno statistické vyhodnocení didaktického testu.

Požadavky ke klasifikovanému zápočtu
Článek s vlastním výzkumem

- **Návrhy témat pro Váš článek**
  - Vlastní didaktický test - návrh, sběr dat a jeho vyhodnocení (určení obtížnosti a diskriminační síly položek, rozložení skórů atd.)
  - Určení validity a reliability didaktického testu, který byl již zadán nějaké skupině žáků (např. nějaký učitel ho již použil pro testování žáků) a pro který máte tudíž sebraná data (jednotlivé položky v testu u každého žáka a jeho celkové skóre v testu)
  - Adaptivní testování - zpracování teorie, čerpat z anglicky psaných zdrojů (články, odborné publikace, internet - oficiální stránky)
  - Item response theory - zpracování teorie, čerpat z anglicky psaných zdrojů (články, odborné publikace, internet - oficiální stránky)

- další Vámi zvolená témata týkající se didaktických testů

- **Rozsah a formát článku**
  - bude možné odvezdat maximálně dva dokumenty = první dokument s vlastním článkem (formát PDF(preferovaný) či DOC, nikoli DOCX) a případně druhý dokument s datovým souborem (formát CSV či XLS)
  - vlastní článek - max 15 normostran, tj. max 27000 znaků (1 normostrana=1800 znaků) a to včetně literatury, tabulek a jejich popisů, popisů obrázků a poznámek pod čarou
  - struktura vlastního článku - záleží samozřejmě na obsahu, obecně se liší teoreticky a empiricky zaměřené články, vždy však je nutné uvést -
    a) název článku + autor;
    b) abstrakt + klíčová slova (alespoň v češtině, v angličtině vítáno, avšak není povinné), rozsah abstraktu - 1200 znaků, počet klíčových slov - max 7;
    c) úvod s přehledem literatury a vymezením cílů;
    d) pro empirické studie - popis výzkumného šetření a vzorku;
    e) prezentace výsledků (např. obtížnost položek a validita testu);
    f) závěr, shrnutí, doporučení, diskuze;
    g) seznam použité literatury
  - tabulky a grafy vkládejte za seznam použité literatury části nazvané "Příloha"
  - projděte si pedagogické časopisy, z nichž lépe pochopíte, jakou strukturu má článek mít

- **Deadline odevzdání práce**: půlnoc 1.1. 2013
  Způsob odevzdání: na webových stránkách ke kurzu (až tyto stránky zprovozním, dám vám emailem vědět) naleznete své jméno a vedle něj bude kolonka na nahrání(upload) Vašeho článku, tam Váš článek nahrajete, případnou přílohu (datový soubor, který

byl použit) bude možné nahrát též
své články odevzdávejte ve formátu PDF(preferovaný formát) či DOC (nikoli DOCX)
práce NEposílejte emailem, nahrávejte je na tuto stránku

Kontaktní údaje

- **dotazy vždy zasílejte elektronicky na**:
  h.vonkova@gmail.com
  znovu opakuji: články emailem neposílejte, nahrávejte je na stránky kurzu

- **konzultace**: čtvrtek 12:15-12:45 v R222

- **Informace ke kurzu** bude možné nalézt na internové adrese www.zla-ryba.cz/hanicka/testy

Výklad je založen na následující literatuře:

- Hopkins, K. D. *Educational and Psychological Measurement and Evaluation.* Needham Heights, MA : Allyn and Bacon,1998.

- Shults, K.S., Whitney, D.J., *Measurement Theory in Action.* Thousand Oaks, CA: Sage Publications, 2005.

- Chrástka, M. *Metody pedagogického výzkumu.* Praha : Grada, 2007.

- Gay, L.R., Mills, G.E., Airasian, P. *Educational Research. Competencies for Analysis and Application.* Upper Saddle River, NJ : Pearson Higher Education, 2008.

- Kolář, Z. , Šikulová, R. *Hodnocení žáků.* Praha: Grada, 2005.

1 Druhy didaktických testů

- testy rychlosti
- testy úrovně
- testy standardizované
- testy nestandardizované
- testy kognitivní a psychomotorické
- testy výsledků výuky a testy studijních předpokladů
- testy rozlišující (testy relativního výkonu)
- testy ověřující (testy absolutního výkonu)
- testy vstupní, průběžné a výstupní
- testy monotématické a polytématické
- testy objektivně skórovatelné
- testy subjektivně skórovatelné

2 Typy testových úloh

Následující materiál je kopií z publikace a je taktéž umístněn na webových stránkách k tomuto předmětu
CHRÁSTKA, M. *Metody pedagogického výzkumu.* Praha: Grada, 2007, s. 188-194.

- sedm naskenovanych obrazku chrastka-typy-uloh1.png, chrastka-typy-uloh2.png, chrastka-typy-uloh3.jpg, chrastka-typy-uloh4.png, chrastka-typy-uloh5.png, chrastka-typy-uloh6.png,chrastka-typy-uloh7.png

- poznámky k návrhům položek Test-items1.jpg

Cvičení

Ke každému z deseti uvedených typů úloh uveďte vlastní příklad. Diskutujte:

- Jaký typ úloh bylo pro Vás nejobtížnější sestavit?
- Je daný typ úlohy pro testovaný obsah vhodný? Nebylo by vhodné zvolit jiný typ úlohy? Pokud ano, jak byste danou úlohy reformulovali?
- Jakým způsobem byste jednotlivé úlohy vyhodnocovaly?

3 Postup konstrukce didaktického testu úrovně

- nezačínat navrhováním testových úloh
- začínat promyšlením účelu testu a dále stanovením obsahu testu - viz obrázek fig-4-1-illustration-of-topic-and-process.jpg
- pro úroveň osvojení poznatků je vhodné použít Bloomovu taxonomii výukových cílů (znalost, pochopení, aplikace, analýza, syntéza a hodnocení)
- stanovíme časový limit
- dále lze přistoupit k formulaci jednotlivých úloh, přičemž je nutné mít neustále na paměti, k jakému účelu úlohy slouží a na základě toho vybírat i vhodný typ testových úloh (otevřené, uzavřené atd.)
- test je vhodné nechat posoudit jiným hodnotitelem (posuzování obsahové validity)

- po sběru dat provedeme analýzu vlastností testových úloh a celého testu (výpočet obtížnosti a citlivosti položek, analýza nenormovaných odpovědí a reliability testu - viz další část)
- vyřadíme úlohy, které nejsou vhodné (např. záporná diskriminační síla)
- pokud má test úrovně, u nějž chceme mít obsahově homogenní úlohy, nízkou reliabilitu, pak výsledky žáků získaných pomocí tohoto testu nemůžeme považovat za spolehlivé a přesné
- následně provedeme standardizaci testu (podle počtu bodů z testu zařadíme žáka do určitého žebříčku)

3.1 Vlastnosti testových úloh - obtížnost, citlivost a analýza nenormovaných odpovědí

- zopakovat základní pojmy popisné statistiky - průměr, směrodatná odchylka, normální rozdělení a korelace na základě slajdů z Metodologie pedagogického výzkumu
- **Obtížnost položky - Hodnota obtížnosti položky**

$$Q = 100\frac{n_n}{N}$$

  - n_n je počet žáků, kteří NEodpověděli na položku správně
  - N celkový počet žáků

- **Obtížnost položky - Index obtížnosti položky**

$$Q = 100\frac{n_s}{N}$$

  - n_s je počet žáků, kteří odpověděli na položku správně
  - N celkový počet žáků

- **Citlivost položek - Koeficient ciltivosti ULI(upper-lower index)**

$$d = \frac{n_L - n_H}{0.5N}$$

  - n_L je počet žáků z "lepší poloviny", kteří odpověděli na položku správně
  - n_H je počet žáků z "horší poloviny", kteří odpověděli na položku správně
  - N celkový počet žáků

- Pro hodnoty obtížnost 30-70 se doporučuje, aby d bylo aspoň 0.25
  pro hodnoty obtížnost 20-30 a 70-80 se doporučuje, aby d bylo aspoň 0.15

- **Analýza nenormovaných odpovědí** = rozbor vynechaných nebo nesprávných odpovědí

- u otevřených úloh věnujeme pozornost těm, ve kterých vynechalo odpověď více než 30-40% žáků, u uzavřených úloh je to pak více než 20%

- u úloh uzavřených s výběrem odpovědi zkontrolujeme atraktivnost distraktorů - neatraktivní distraktor nahradíme jiným

- u uzavřených úloh rozdělíme nesprávné odpovědi do dvou kategorií - základní chyby (způsobené neznalostí učiva) a vedlejší chyby (způosbené náhodnými vlivy), odstraníme úlohy, kde převáží vedlejší chyby nad základními chybami

3.2 Reliabilita testu

- Didaktický test má dobrou reliabilitu, pokud poskytuje spolehlivé a přesné výsledky. Pokud bychom test neustále opakovali za stejných podmínek, měli bychom v případě testu s dobrou reliabilitou získat velmi podobné výsledky.

- Hodnota se pohybuje od 0 do 1
- test s dobrou reliabilitou má hodnotu alespoň 0.8
- vysoká VALIDITA $\Rightarrow$ vysoká RELIABILITA
- vysoká VALIDITA $\Leftarrow$ vysoká RELIABILITA
- **Kuder-Richardsonův vzorec pro výpočet reliability**
  pro položky skórované 0,1
  vhodný pro testy úrovně

$$r_{kr} = \frac{K}{K-1}\left(1 - \frac{\sum_k p_k q_k}{s^2}\right)$$

  - K počet úloh v testu
  - p_k podíl žáků, kteří řešili danou úlohu k správně
  - q_k podíl žáků, kteří řešili danou úlohu k chybně ($q_k = 1 - p_k$)
  - s^2 výběrový rozptyl pro celkové výsledky žáků v celém testu
- **Reliabilita vypočtená metodou půlení**
  skórování položek není omezeno
  vhodný jak pro testy úrovně, tak pro testy rychlosti

$$r_{sb} = \frac{2.r_b}{1 + r_b}$$

  - r_b korelační koeficient mezi výsledekem žáků v sudých a lichých úlohách

3.3 Standardizace testu

- počet bodů v testu neříká, zda je výkon žáka dobrý či slabý; jeden žák může získat v jednom testu relativně hodně bodů a v jiném relativně málo bodů

- u standardizovaných testů se výkon žáka provnává s výkonem jiných žáků z reprezentativního vzorku dané skupiny (v takovémto vzorku jsou zpravidla stovku žáků)
- standardizovat výsledky testu znamená vyjádřit je vzhledem k výsledkům standardizačního vzorku žáků
- **Percentilová škála** udává, kolik procent žáků dosáhlo horšího výsledku

$$PR = 100\frac{n_k - \frac{n_i}{2}}{N}$$

  - n_k kumulativní četnost daného výsledku
  - n_i četnost daného výsledku
  - N celkový počet žáků
- **z-škála** vychází z předpokladu normálního rozdělení
- vyjadřuje, jak daleko je výsledek od aritmetického průměru, jako jednotka vzdálenost je vzata směrodatná odchylka

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

  - X určitý testový výsledek
  - $\bar{X}$ aritmetický průměr všech výsledků
  - S směrodatná odchylka všech výsledků
- **Z-škála** vychází ze z-škály

$$Z = 100 + 10z$$

- **T-škála** vychází ze z-škály

$$T = 50 + 10z$$

Cvičení

Použij didtest_data_analyza.xls s daty o vysledcích 40 žáků z 10ti položkového testu. Proveď analýzu vlastností položek, vypočítej reliabilitu pomocí obou výše diskutovaných metod a proveď standardizaci testu (předpokládej, že se jedná o reprezentativní vzorek žáků, o jejichž výsledcích lze předpokládat, že jsou normálně rozdělené)

4 Validita a reliabilita testů - podrobnější diskuze

- při analýze didaktického testu jsme hovořili o obsahové validitě a reliabilitě měřené pomocí Kuder-Richardsonovy formule a metodou půlení, které se užívají především u učitelských testů
- podrobnější diskuze k různým typům validity a reliability lze nalézt na obrázcích Table-6-2-validity.jpg a Table-6-3-reliability.jpg

5 Modely srovnávání testů

tato část je převzata z webových stránek organizace Scio z internetové adresy
http://www.scio.cz/in/2vs/nsz/vysledek/metodika.asp

Srovnávací model náhodných skupin (Random Groups Design)

Tento model je využíván, pokud máme v jednom termínu dvě varianty stejného testu (např. testu OSP). Skupina testovaných je náhodně rozdělena na dvě poloviny, z nichž každá řeší jednu variantu testu. Obvyklá metoda rozdělení je tzv. "spiraling", kdy jsou obě varianty v jedné místnosti rozděleny střídavě. První testovaný píše variantu A, druhý variantu B, třetí variantu A atd. Při takovémto náhodném

rozdělení můžeme obě podskupiny považovat za rovnocenné (equivalent) a rozdíly ve statistických parametrech obou variant testu dosažených příslušnou podskupinou (průměrná úspěšnost, rozptyl skóre) přímo považujeme za rozdíly těchto dvou variant (bez vlivu úrovně testované skupiny). Tato metoda je použita pro potřeby NSZ.

Srovnávací model společných úloh pro neekvivalentní skupiny (Common -Item Nonequivalent Groups Design)

Tento model je užíván v případech, kdy dvě varianty testu řeší dvě různé (neekvivalentní) skupiny. Typickým příkladem jsou dva různé termíny jednoho testu, kdy ekvivalent skupin nejsme schopni nijak zaručit (např. hypotéza, že na první termíny se hlásí zodpovědnější uchazeči než na poslední. Dopad tohoto vlivu není možné předem odhadnout). Rozdíly v průměrné úspěšnosti a dalších statistických charakteristikách obou variant jsou ovlivněny nejen rozdílností variant, ale také rozdílností testovaných skupin. V tomto modelu varianta A a varianta B mají společnou podmnožinu úloh. Na těchto společných úlohách se porovnávají rozdílné úrovně obou testovaných skupin. A poté je možné provést srovnání obou variant očištěné od vlivu rozdílnosti skupin. Tato metoda je použita pro potřeby NSZ.

Další užívané srovnávací modely

Mezi další užívané srovnávací modely patří **Model jedné skupiny (Singel Group Design)**, kdy obě varianty testu jsou distribuovány stejné skupině testovaných, a **Vyvážený model jedné skupiny (Singel Group Design with Counterbalancing)**, kdy jsou obě varianty opět testovány na jedné skupině, ale polovina testovaných absolvuje nejprve variantu A a poté variantu B, zatímco druhá polovina řeší testy v opačném pořadí. Tento model eliminuje vliv zkušenosti s testem, který ovlivňuje úspěšnost druhého testu v pořadí. Oba tyto modely nejsou pro NSZ vhodné.

Metody srovnávání testů

Dvěma nejužívanějšími metodami srovnávání testů jsou metoda lineární a metoda ekvipercentilová. Lineární metoda je založena na srovnávání průměrné úspěšnosti a rozptylu skóre obou variant. Ekvipercentilová metoda je založena na porovnávání kumulativních distributivních křivek. Zjednodušeně řečeno, ekvipercentilová metoda srovnává účastníky, kteří v jednotlivých variantách dosáhli stejného percentilu (předstihli stejné množství ostatních účastníků dané varianty). Na rozdíl od lineární metody je ekvipercentilová metoda přesnější na celé škále skóre. Proto byla pro potřeby NSZ 2008/2009 zvolena ekvipercentilová metoda a v dalším textu je podrobně vysvětleno její konkrétní užití.

5.0.1 Ekvipercentilová metoda (Equipercentile Equating)

Ekvipercentilová metoda je založena na pojmu percentil skóre, který pro dané skóre uvádí, kolik procent z testovaných dosáhlo nižšího nebo stejného skóre (někdy se v definici uvažuje pouze nižší skóre, což je z faktického hlediska rovnocenné). Srovnání skóre z jedné varianty se skórem z druhé varianty pak dosáhneme tak, že ke každému skóre z první varianty přiřadíme skóre z druhé varianty, které má stejný percentil. Předpokladem ekvipercentilové metody je, že skupiny testovaných v obou variantách testu jsou rovnocenné, což platí například pro model náhodných skupin, kde se tato metoda hojně využívá. Přesná matematická definice je pak následující ...

5.0.2 Zřetězená ekvipercentilová metoda (Chained Equipercentile Equating)

Pro srovnávací model společných úloh pro neekvivalentní skupiny se užívá zřetězená ekvipercentilová metoda. Jak již bylo napsáno výše, tento model využívá společné množiny úloh, které se vyskytují ve variantě X i Y (označení X a Y užíváme, protože se obecně jedná o dvě varianty testu použité v jiných termínech). Srovnávání se pak skládá ze dvou ekvipercentilových srovnání na stejné skupině testovaných. Nejprve se skóre z varianty X ekvipercentilově srovná se

skórem na společných úlohách. Společné úlohy se zde uvažují jako samostatný test, který řešila stejná skupina lidí jako variantu X. Tytéž společné úlohy řešila také skupina lidí testovaných variantou Y. Opět můžeme skóre ze společných úloh (tentokrát řešených skupinou lidí z varianty Y) ekvipercentilově srovnat se skóre z varianty Y. Spojením (zřetězením) těchto dvou srovnávání dostaneme srovnání skóre varianty X se skórem varianty Y. Přesná matematická definice je pak následující: VZOREC matematické definice (PDF)

6 Příklady standardizovaných testů studijních předpokladů a psychologických testů

- obrázek fig-3-1-types-of-standard-scored-scales.png je převzat z publikace Hopkins, K. D. *Educational and Psychological Measurement and Evaluation.* Needham Heights, MA : Allyn and Bacon,1998, s. 53.
- ukázka GRE testu a TOEFL testu