MUNI FSS Katedra psychologie

Přednáška 6:
Férovost a zkreslení při testování

10. 11. 2020 | PSYn4790 | **Psychometrika: Měření v psychologii**

Katedra psychologie, Fakulta sociálních studií MU

Hynek Cígler | hynek.cigler@mail.muni.cz

Co si představíte
pod termínem **férovost**…
… v psychodiagnostice?
… a v psychometrice?

Férovost v psychometrice

2. kapitola českého překladu *Standardů pro pedagogické a psychologické testování* (AERA, 2001).

- Doporučuji vydání 2014 v Aj
  - A to i pro studium PSYn4020/PSYn5340.

„Férovost“ a s ní související téma multikulturního testování je jedním z důležitých témat současné psychometriky (zejména v rámci tzv. edukativního testování).

Klíčové pojmy

Přístupnost (accesibility).

- Měřený rys je *stejně dostupný* u všech potenciálních probandů.
- Příslušnost ke skupině probandů neovlivňuje výsledek v testu *po kontrole rysu*.
- Např. zrakové/sluchové znevýhodnění, znalost jazyka apod.

Univerzální design.

- Charakteristika testu, která zajišťuje přístupnost.
- Např. snaha o vyřazení položek se silnou kulturní specificitou.
- Nebo zvážení rozdílnost účelu testu napříč skupinami probandů.

4 základní významy férovosti (AERA, 2014)

Férovost při zacházení během testování.

- **Psychodiagnostika**.
- „Objektivita“, rovné zacházení... se všemi zacházím stejně.
- „Standardizace I“ dle Urbánka (2010).

Nepřítomnost testového zkreslení.

- **Psychometrika**.
- „Test bias“, „item-bias“.
- Test a položky měří u všech stejný rys.
- DIF, invariance atd.

Férovost jako přístupnost, otevřenost.

- **Psychometrika, psychodiagnostika, teorie.**
- „Accessibility“, „provability“.
- Vlastnost je u respondenta měřitelná.

Férovost jako interpretace individuálního skóre pro daný účel

- **Psychodiagnostika**.
- Zvážení jedinečnosti každého respondenta.
- Jaká individuální specifika mohou výkon ovlivnit.
- Akomodace, individuální úpravy testu.

AERA, APA, NCME (2014). *Standards for educational and psychological testing.* Washington: AERA. Strany 49–72.

Bias = zkreslení

Bias = systematické zkreslení testových výsledků

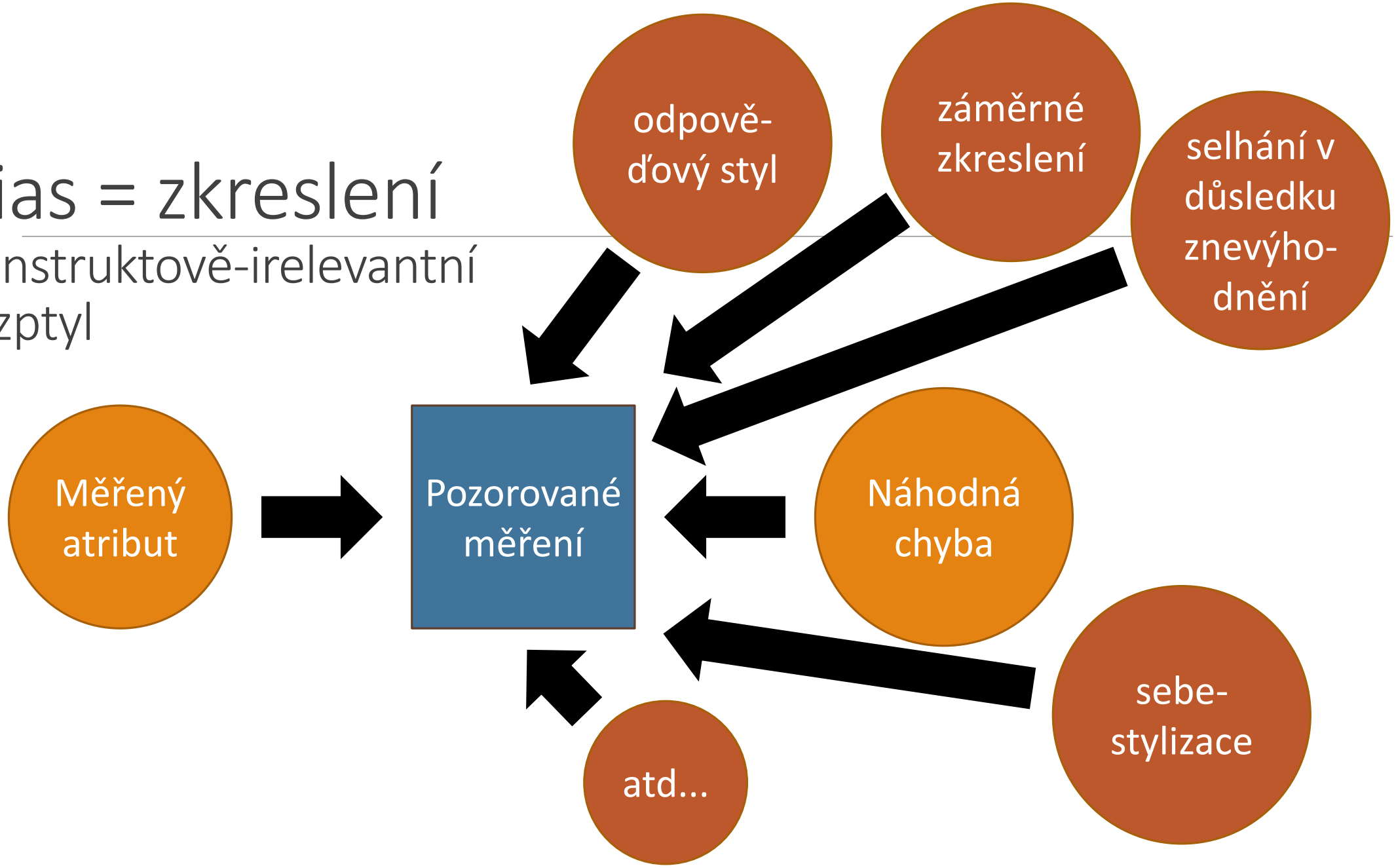
- Reliabilita: náhodné chyby měření, není tedy otázkou „bias“.
- Úvaha o zkreslení patří do validity, ale z praktických důvodů je vyčleňována.
- Slovem zkreslení označujeme nenáhodné, specifické chyby měření.
- *„Měří test jinak pro některé populace než pro jiné?“*
- *„Měří test jinak pro některé specifické osoby?“*
- *„Měří test obecně spravedlivě?“*

Může znamenat, že v různých populacích např.:

- Je test/položka příliš snadná/obtížná.
- Má test/položka jiný vztah k rysu.
- Test má jinou faktorovou strukturu.
- Test měří zcela či částečně něco jiného.
- ...
- *Výkon v testu je ovlivněn systematicky něčím, co nemá souvislost s tím, co chci měřit (**„konstruktově irelevantní rozptyl“**).*

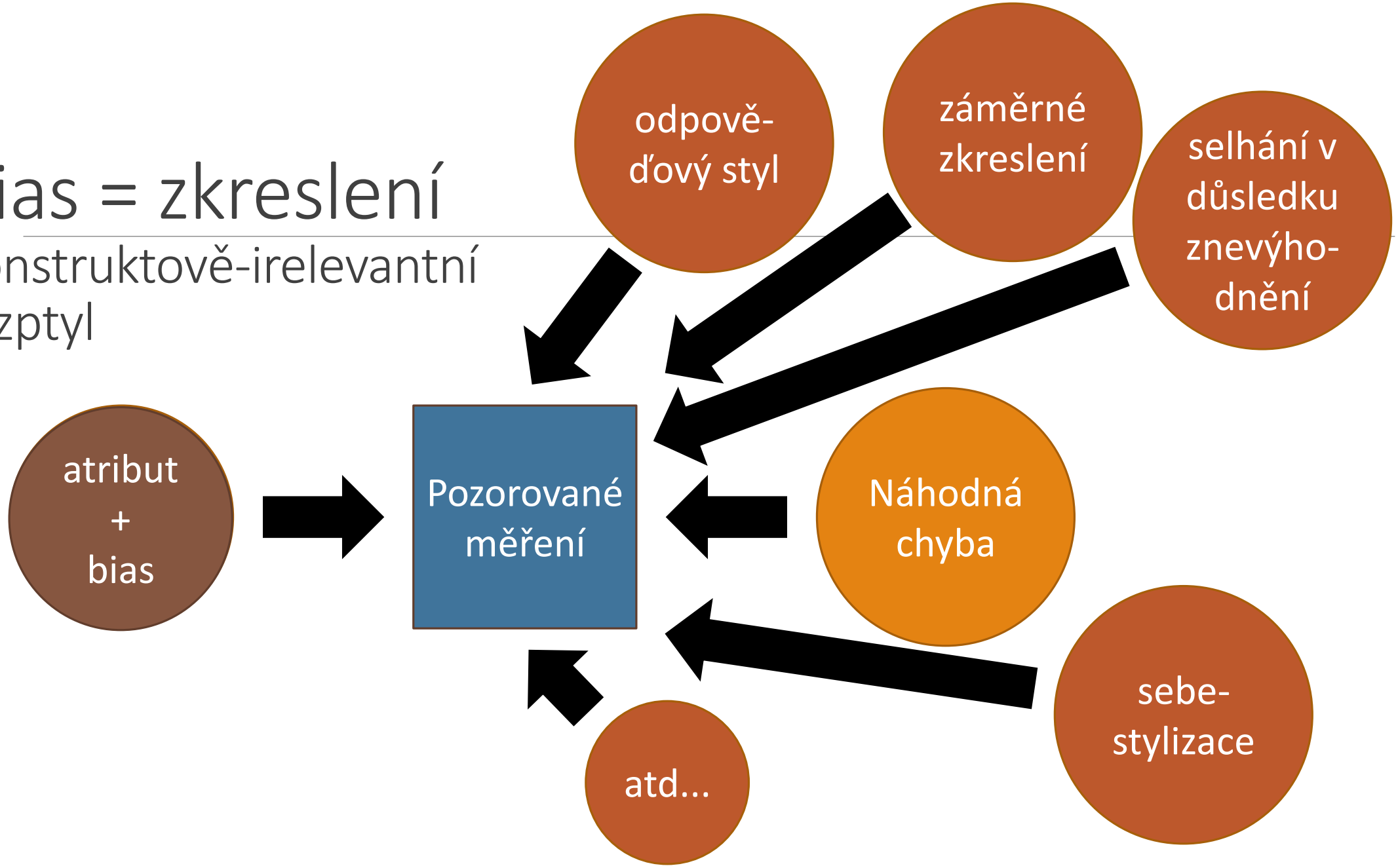
Bias = zkreslení

Konstruktově-irelevantní rozptyl



Bias = zkreslení

Konstruktově-irelevantní rozptyl



Potenciální oblasti systematického zkreslení:

Zkreslení na úrovni examinátora a testové situace, nestranné zacházení.

- Je zacházeno se všemi respondenty stejně?

Response bias

- Záměrné i nezáměrné zkreslení odpovědi respondentem na úrovni položky.

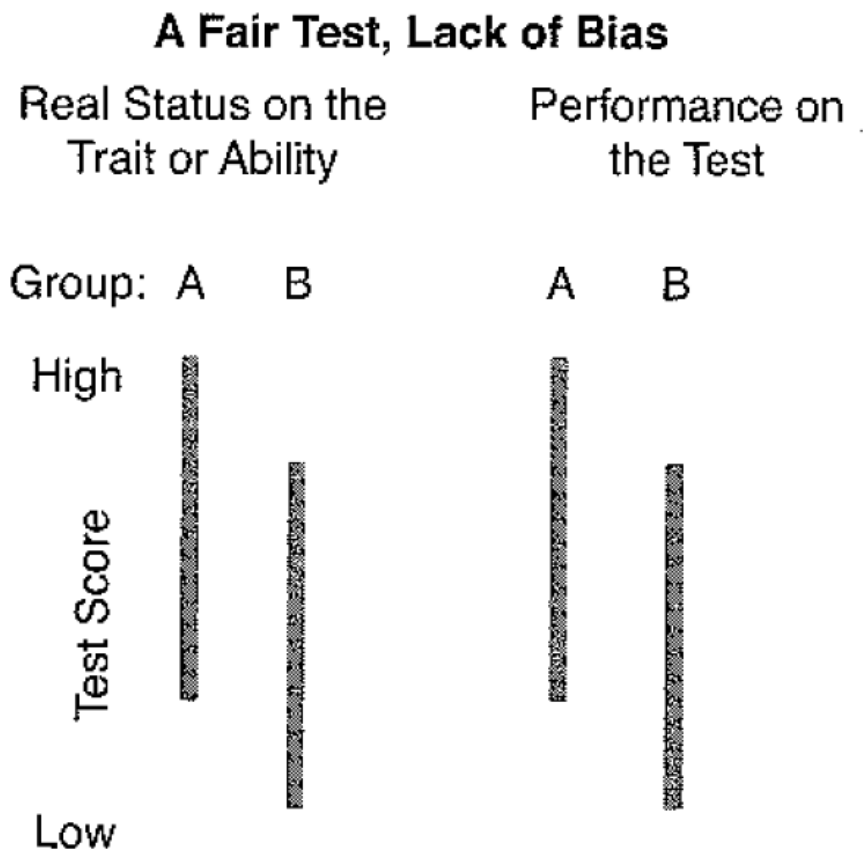
Item bias

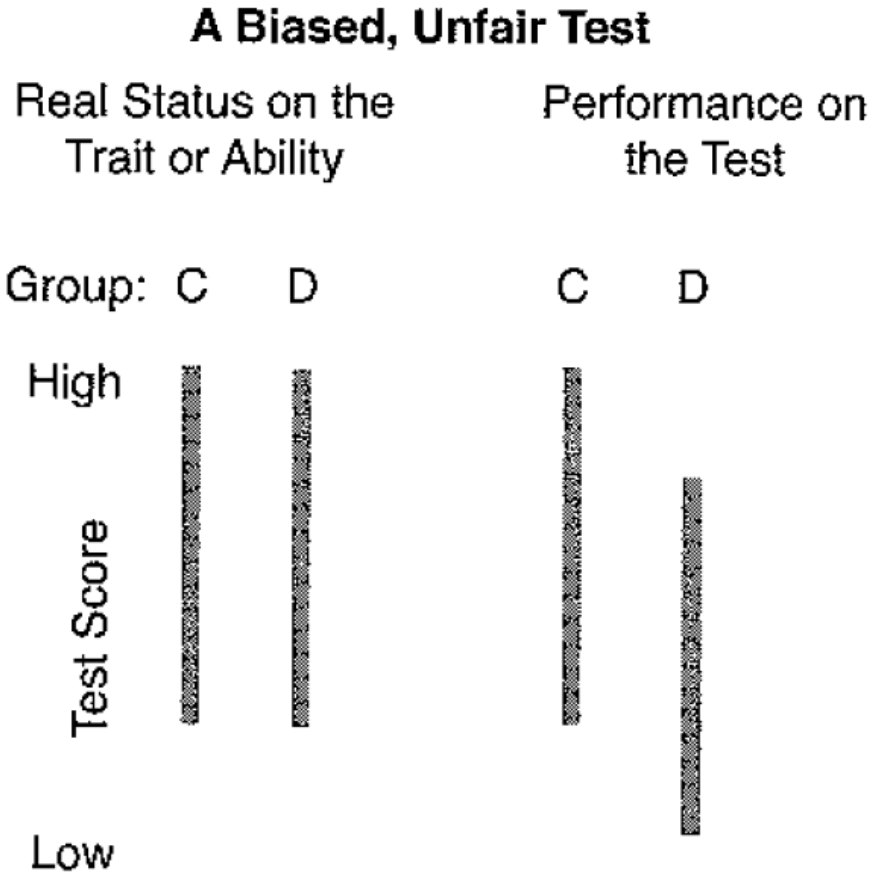
- Systematické rozdíly mezi osobami/skupinami v odpovědi na položku, nevysvětlitelné rozdílnými úrovněmi latentního rysu.

Test bias

- Týká se systematického zkreslení celkových skórů/výsledků testu napříč skupinami respondentů.
- Je otázkou „férovosti testu“, oba pojmy jsou zaměnitelné.

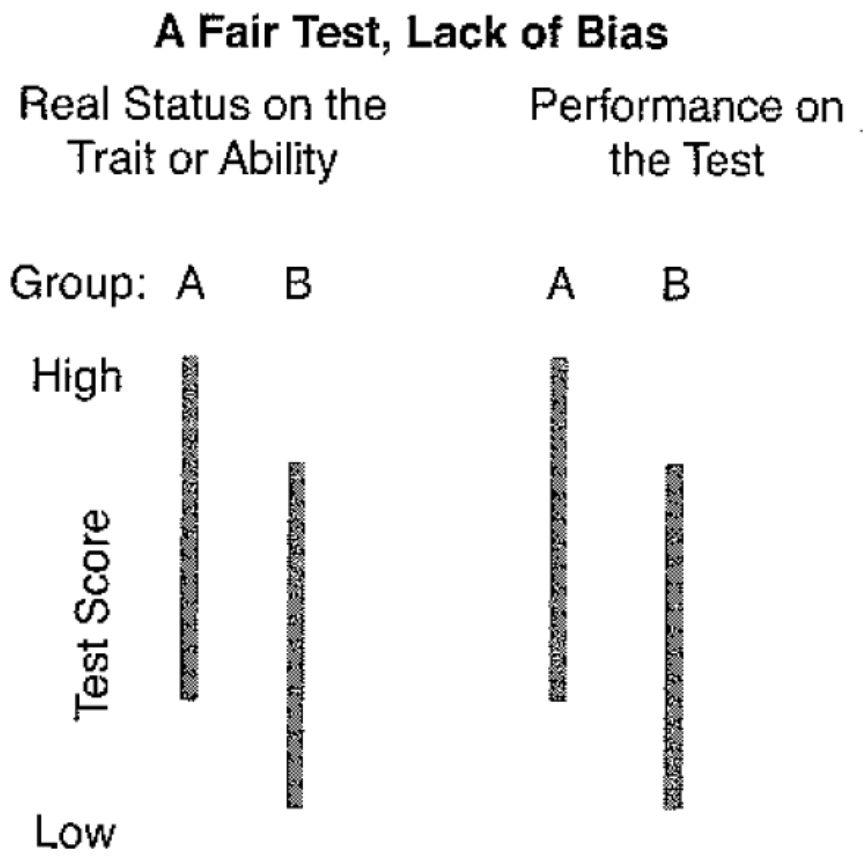
Test bias, test fairness

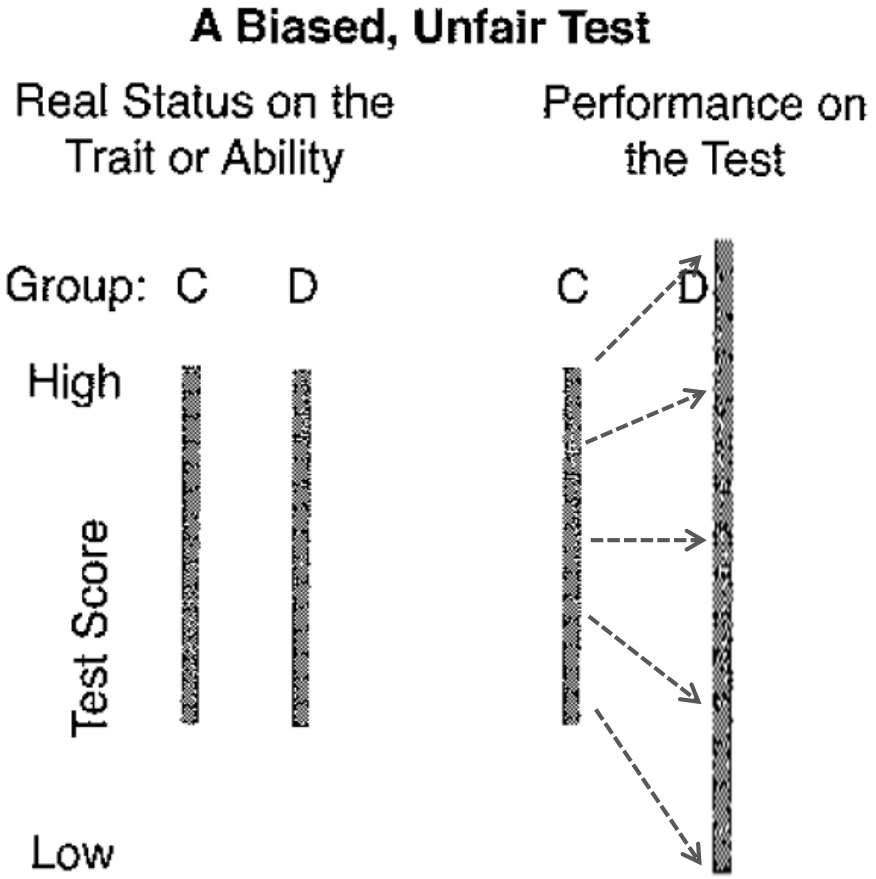




FURR, s. 233

Test bias, test fairness





FURR, s. 233

Zdroje ohrožení férovosti testování

Obsah testu

- Který znevýhodňuje některé skupiny, osoby atd.

Kontext testové situace

Odpovědi na položky

- Formát položek, interpretace při kvalitativním skórování výsledků...

Příležitost k přípravě na test

Kontext testové situace

Tohle je otázka spíše do psychologické diagnostiky/etiky.

Cílem je zajistit, aby každý respondent měl možnost projevit ty stejné schopnosti ve stejné míře.

- APA standard 7.12: *„Testování nebo hodnocení by mělo probíhat takovým způsobem, aby se všem testovaným osobám dostalo stejného nebo srovnatelného zacházení během všech fází testování.“*

Administrátor testu rovněž musí být kompetentní s konkrétním testem pracovat (školení, zácvik...).

Možnost přípravy

Všichni respondenti musí mít shodné možnosti zácviku, poučení o cíli testování...

- Na tohle pozor! Běžná praxe neomlouvá...

„Tajné“ informace o způsobu dopravně-psychologického vyšetření.

Placené (a drahé) přípravné testy na přijímačky.

Neformálně dostupné informace o průběhu forenzního vyšetření.

Různý způsob informování před zahájením vyšetření.

Férovost jako přístupnost

Příklad: Přijímačky do bc studia na FSS

2 testy: studijní předpoklady (váha 0,4), ZSV (váha 0,6)

Studijní předpoklady – na výběr:

- SCIO (až 5 pokusů, bere se nejlepší)
- TSP od MU (1 pokus)

ZSV – jediná možnost:

- SCIO (až 5 pokusů, bere se nejlepší)

Jaké jsou nevýhody daného designu z hlediska psychometriky?

Co byste studentům řekli, aby měli rovné podmínky?

Simulace: http://fssvm6.fss.muni.cz/prijimZk/

Response bias

Souvisí s „response style“.

- Jde o určitý styl odpovídání specifický konkrétnímu respondentovi v konkrétní situaci, který znehodnotí/zneplatní testové výsledky.

Nahodilé odpovědi a záměrné zneplatnění výsledků.

Zkreslení v užším významu (práce Paulhuse a kol.[1]).

- Simulování a sebeznevýhodňování (záměrné). Tzv. „impression management“.
- Sociální žádoucnost a nezáměrné zkreslení. Tzv. „self-deception“.

Response style

- Tendence k souhlasu nebo nesouhlasu.
- Tendence k extrémním nebo průměrným odpovědím.

Hádání, tipování.

[1] Řada dílčích publikací o self-presentation, overclaiming, self-management atd.

Response bias – řešení

Změna setingu testové situace, aby respondent nebyl motivován výsledky zkreslovat.

- Anonymita, redukce stresu, srovnání úrovně motivace...

Úprava položek

- Jednoduché položky – krátké jednoznačné stimuly, krátké jednoznačné a „ne-extrémní“ distraktory.
- U delších odpověďových (Likertových škál) je větší prostor pro zkreslení.
- Zajištění absence chybějících odpovědí.
- Rozdílná valence položek (negativní skórování).

Odhalení zkreslení

- Tzv. „validizační škály“ či „lži škály“ (např. v případě MMPI-II 6 různých škál).
- Dodatečné testy (Malingering scale – máme v KDM).
- Netestová detekce ☺.

Metody ověření zkreslení

Expertní panelová review: Obsahová validita.

Diferenciální fungování položek: Vnitřní struktura testu.

Testová invariance: Vnitřní struktura testu.

Diferenciální predikce testu: Prediktivní/kriteriální validita

Panel review

Zejména v případě high-stakes testů.

Pečlivá volba tzv. expertního panelu (Subject Matter Experts, SME).

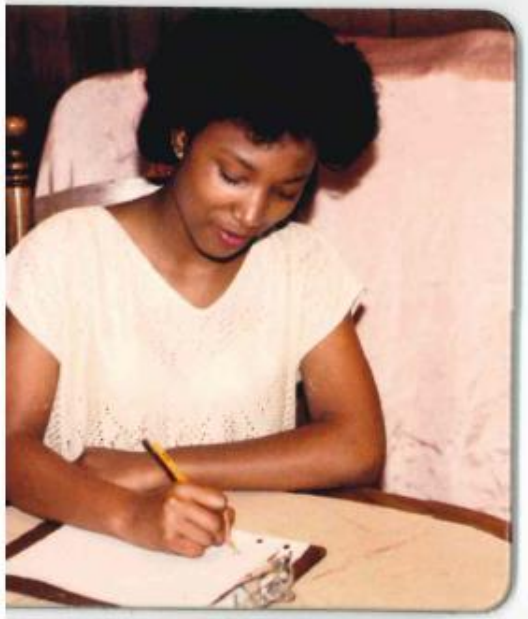
Expertní panel vytváří, reviduje a připomínkuje položky a složení testu (zejm. didaktické a edukativní testy).

- Experti musí být experty na měřený konstrukt.
- Zároveň by ale měli dobře reprezentovat testovanou populaci.
- Muži i ženy, minority...
- Jsou ale SME z určité minority dobrými reprezentanty této minorit?

Test bias, item bias:
Férovost z hlediska psychometriky.

DIFFERENTIAL ITEM FUNCTIONING (DIF)

DIFFERENTIAL TEST FUNCTIONING (DTF)



Test/item bias

Nelze odvodit bez dat (jen odhadovat).

- Empirické důkazy a technická řešení.

Respondent se snaží odpovídat pravdivě, ale test měří v různých skupinách něco jiného.

WAIS-III: *„Co uděláte, když najdete na zemi zalepenou poštovní obálku s napsanou adresou, známkou, ale bez razítka?“*

WISC-III: *„Co uděláte, když chcete uvařit čaj?“*

Skupiny: etnikum, pohlaví, jazyk, socio-ekonomický status, region...

Dva hlavní empirické přístupy k férovosti

Na úrovni položky (item bias analysis).

- Které položky (a zda ta která položka) vykazují rozdílný styl odpovídání napříč skupinami, který nelze přičíst rozdílům v úrovni latentního rysu?
- DIF analýza (differential item functioning)

Na úrovni testu (test bias analysis).

- Do jaké míry test jako celek (soubor mnoha různých položek) měří ten stejný rys pro různé skupiny?
- Lze srovnávat naměřené skóry napříč skupinami?
- Invariance testu.

Logika ověření zkreslení

Předpoklad férovosti:
Atribut (latentní rys) „způsobuje“ pozorované odpovědi.

Systematické zkreslení znamená:
Příslušnost ke skupině moderuje tento vztah.

- Zvyšuje/snižuje intercept závislé proměnné.
- Zvyšuje/snižuje regresní koeficient.
- Zvyšuje/snižuje reziduální rozptyl.

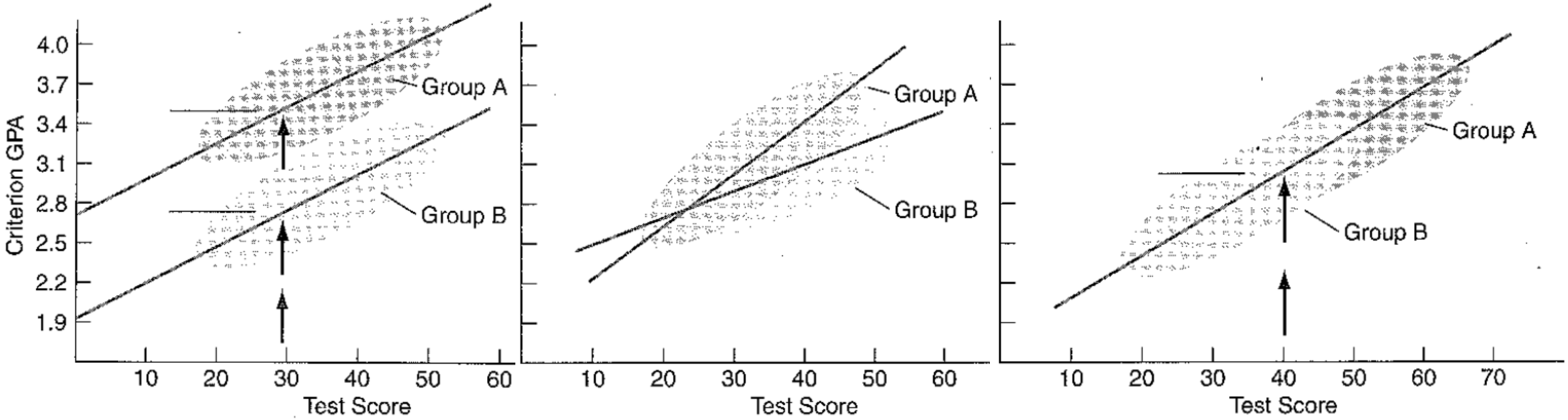
Bias/zkreslení = moderace.

Test bias (zkreslení predikce)
např. přijímací zkoušky vs. státnice

situace A
rozdíl v průměru predikce

situace B
rozdíl v přesnosti (a průměru) predikce

situace C
férový test



Test bias

Ověření typicky pomocí moderačního modelu (lineární i logistická regrese).

Krok 1: vytvoření interakční proměnné součinem prediktoru a moderátoru.

Krok 2: prostá lineární regrese

- Prediktivní nebo kriteriální validita.

$$Y = aX + b$$

- *Y* – kritérium, *X* – výsledek testu
- *a* – směrnice/slope, *b* – průsečík /int.

Krok 3: přidání moderátoru do regrese.

$$Y = aX + b + \left(cM + d(M \cdot X)\right)$$

- M – moderátor (skupina osob...)

Signifikantní F-test rozdílu 1. a 2. modelu (ΔR^2) → přítomnost test bias.

- sig. *c* → rozdíl v průměru predikce.
- sig. *d* → rozdíl v přesnosti predikce.

Srovnáváme nestandardizované koeficienty!

- Standardizované jsou ovlivněné populačními charakteristikami, které se lišit mohou.

DIF v CTT

Jen obtížné, protože CTT a FA nedobře modeluje odpovědi na položku v závislosti na HS.

Komparace ULI indexů: Rozdělíme vzorek pro výpočet ULI napříč skupinami.

- ULI následně spočítáme pro celý vzorek, pro jednu i druhou skupinu.
- Jsou stejné? Jaká je korelace ULI napříč skupinami?

Komparace popularit položek.

- Je pořadí položek dle obtížnosti stejné napříč skupinami?
- Korelují popularity položek napříč skupinami?
- Spearmanova korelace obtížností položek.

DIF v CTT

Mantelův-Haenszelův test. Chí-kvadrát pozorovaných odpovědí pro každou úroveň HS a následná agregace výsledků.

Postupy založené na logistické regresi.

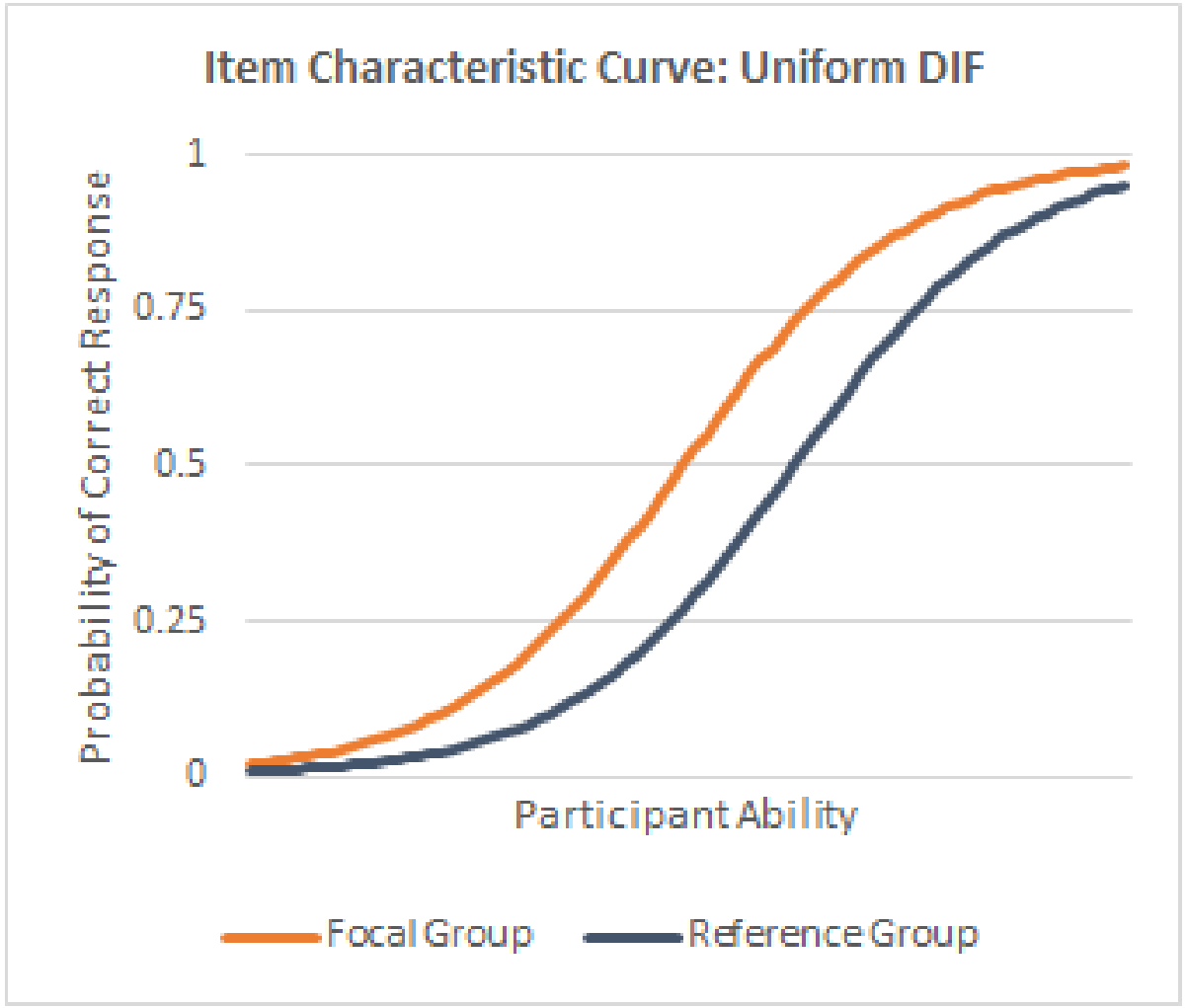
- Podobný přístup jako v případě test bias.
- Prediktorem je hrubé skóre, závislou odpověď na položku, moderátorem příslušnost ke skupině.
- Binární pol.: logistická regrese; ordinální pol.: ordinální log. regrese.

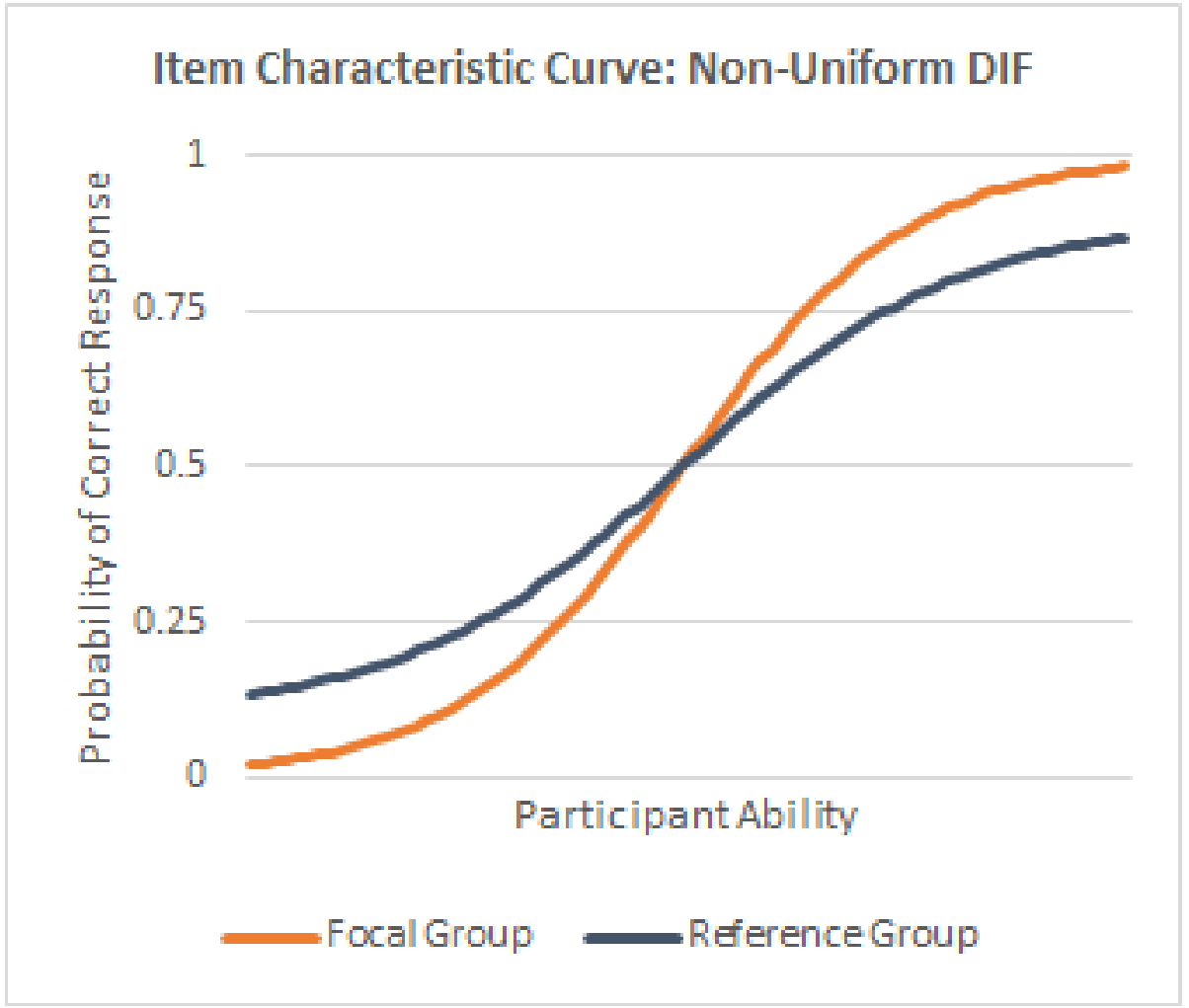
IRT: Differential Item Functioning

DIF analýza se používá se zejména v rámci IRT.

Obecný framework pro usuzování na neférovost jednotlivých položek.

- Některé postupy aplikovatelné i v CTT, ale IRT je výrazně vhodnější.
- V CTT je např. problematické testovat non-uniform DIF (viz dále), nebo DIF mezi skupinami, které se výrazně liší svým výkonem.





Zdroj: http://blog.questionmark.com/wp-content/uploads/2013/12/Item-Characteristic-Curve.png

Příklad: Žádné DIF

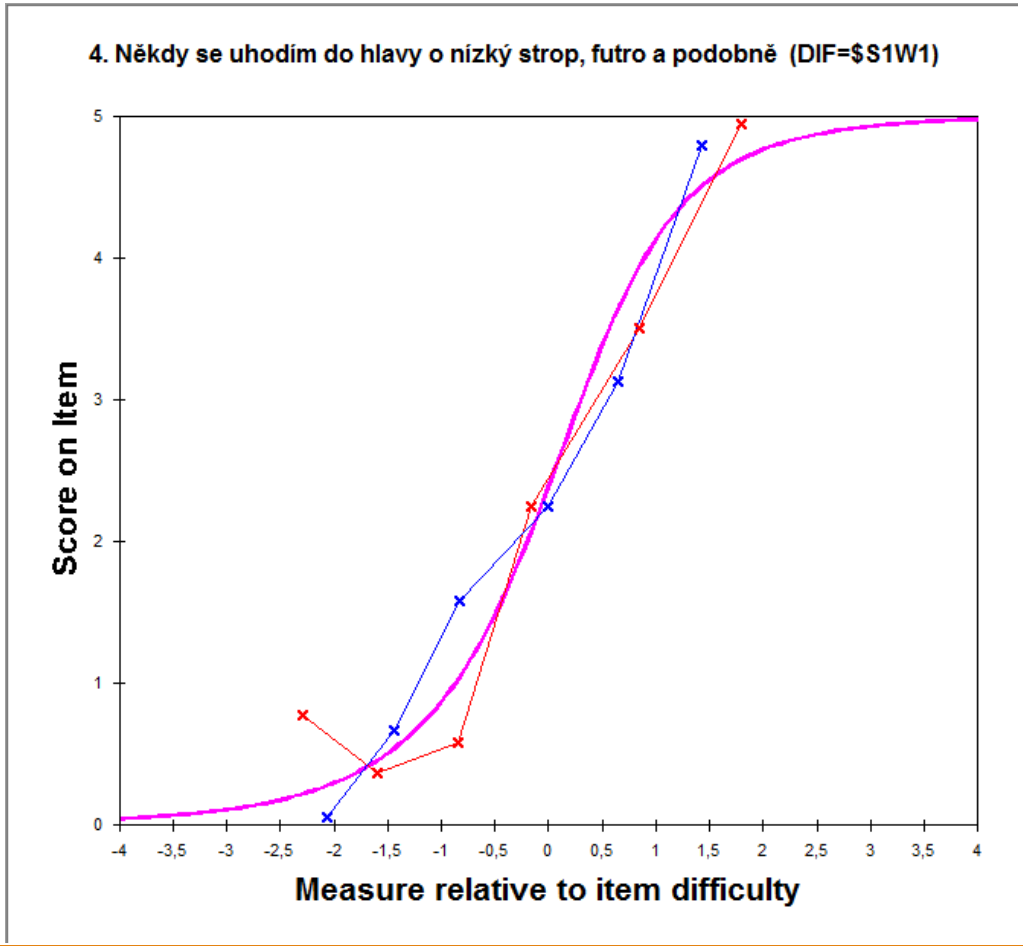
Dotazník výšky:

Někdy se uhodím do hlavy o nízký strop, futro a podobně.

DIF:

- t-test: t(86) = -0,31, p=0,756
- M-H: $\chi^2(1)=0,44$, p=0,508.

Modrá muži, červená ženy.



Příklad: Uniform DIF

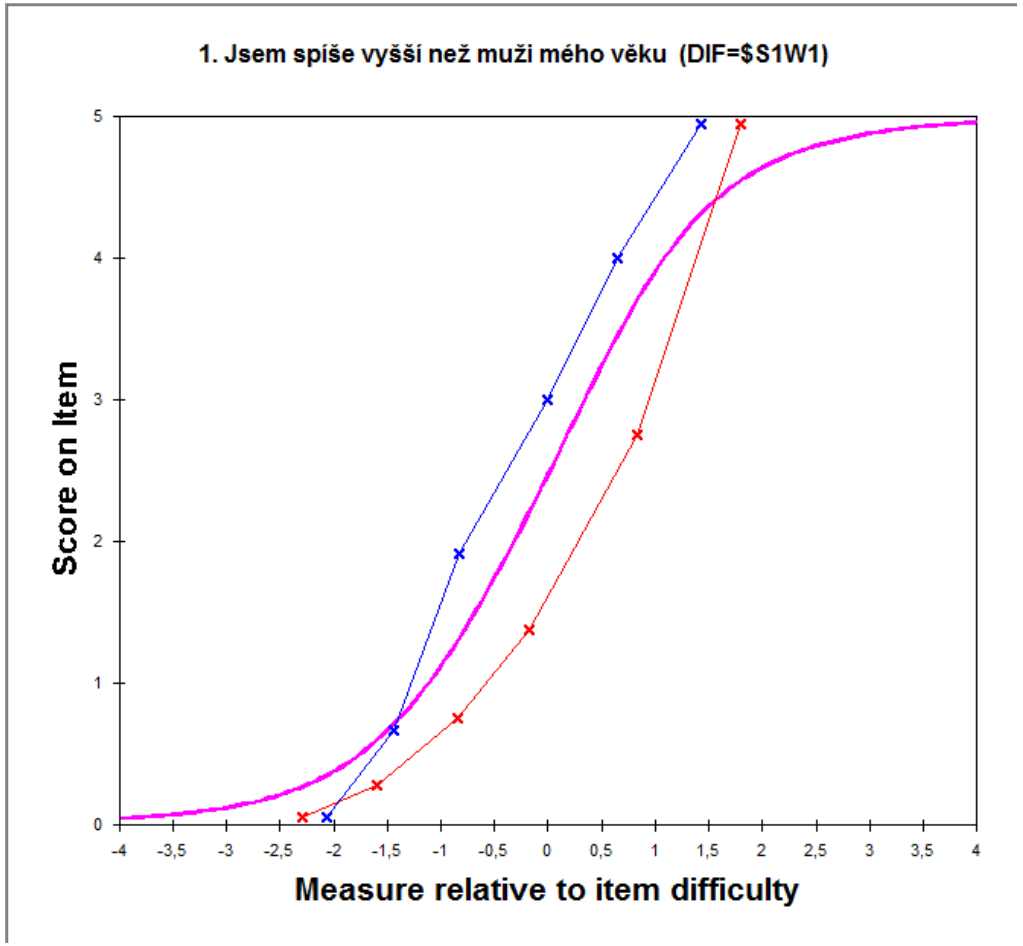
Dotazník výšky:

Jsem spíše vyšší
než muži mého věku.

DIF:

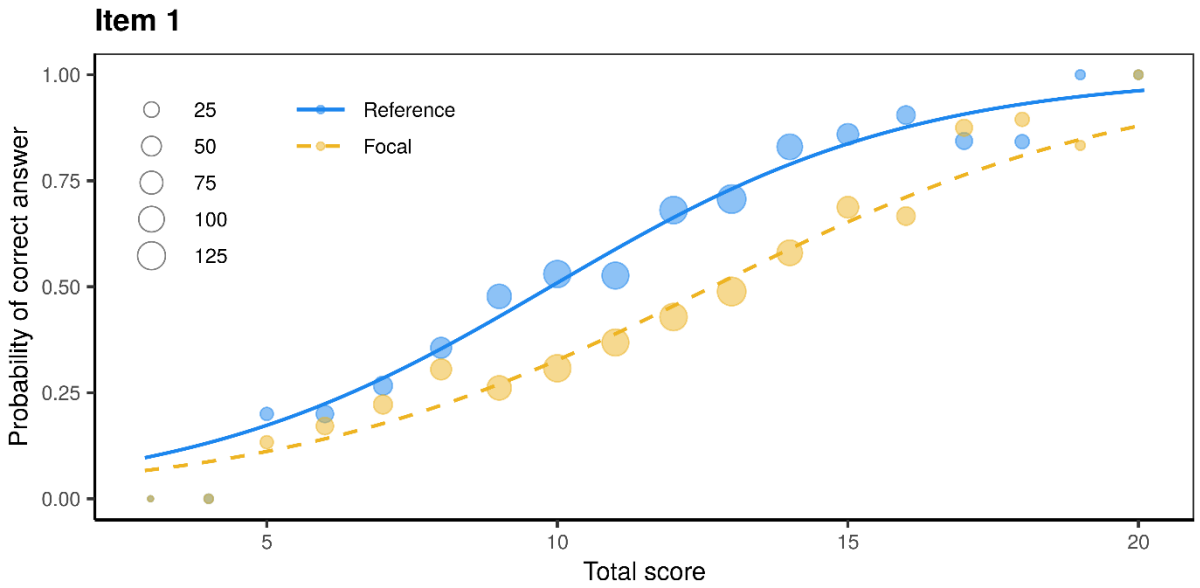
- t-test: t(86) = -4,63, p<0,001
- M-H: $\chi^2(1)=18,7$, p<0,001.

Modrá muži, červená ženy.

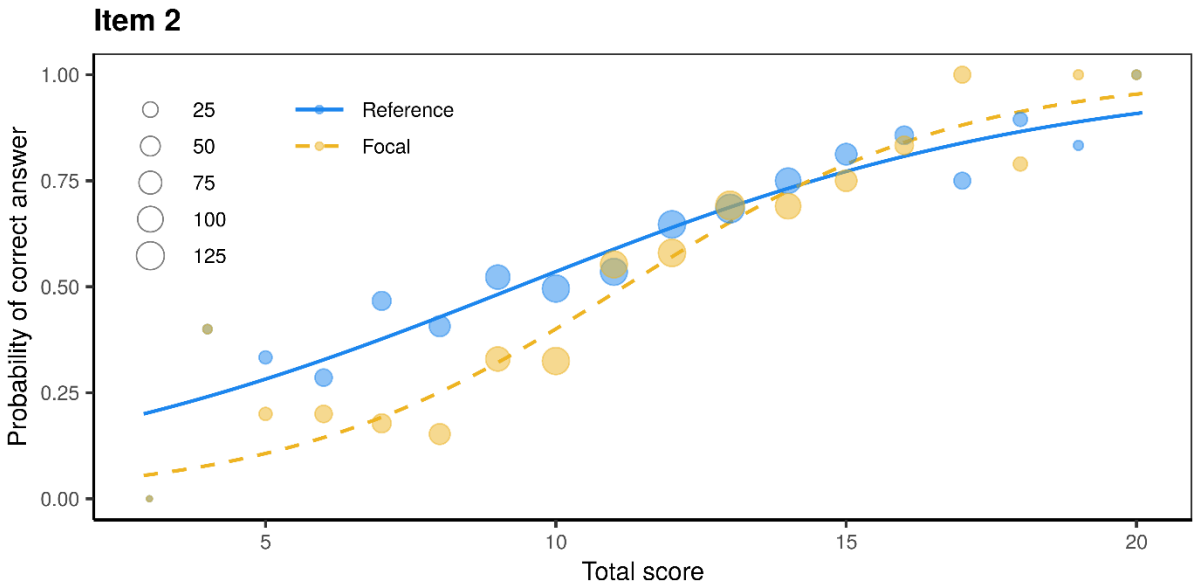


Příklad: uniform vs. non-uniform DIF

Uniformní DIF



Non-uniformní DIF



https://shiny.cs.cas.cz/ShinyItemAnalysis/

Software

Zejména R, různé balíčky

- difNLR package, mirt, difR, lordif, DIFlasso, DIFtree...

On-line aplikace: https://shiny.cs.cas.cz/ShinyItemAnalysis/

Jakýkoli statistický program, který disponuje modulem pro (ordinální) logistickou regresi.

Test bias: Invariance měření

Postup založený na konfirmační faktorové analýze, ale je použitelný i v IRT.

- Tzv. multiple-group CFA/IRT (MG CFA, MG IRT).

Ověřuje shodnost faktorové struktury (modelu měření) napříč skupinami.

- Rozdílné úrovně invariance umožňují rozdílné možnosti srovnání skupin.

Typicky se řeší při:

- Při konstrukci diagnostických metod: je test jako celek „férový“ pro různé skupiny respondentů?
- Large-scale assessment: Do jaké míry mohu srovnávat skóry respondentů napříč zeměmi/státy/kulturami atd.?
- Teoreticky při každém použití t-testu by měla být vyargumentovaná invariance napříč oběma skupinami, aby je bylo možné srovnat.

Doporučuji: Wicherts, J. M., Dolan, C. V. (2010). Measurement Invariance in Confirmatory Factor Analysis: An Illustration Using IQ Test Performance of Minorities. *Educational Measurement: Issues and Practice 29*(3), 39–47. doi: 10.1111/j.1745-3992.2010.00182.x

4 (5) stupňů invariance:

Základní:

- 1. Konfigurální invariance.
- 2. Metrická (slabá invariance).
- 3. Skalární (silná invariance).

Další:

- 4. Reziduální (striktní) invariance.
- 5. Paralelní skupiny.

Jednotlivé stupně/úrovně:

- Vyšší úrovně zahrnují všechny požadavky úrovní nižších.
- Nižší úrovně jsou předpokladem úrovní vyšších.

Analogie k „paralelním položkám.

- Paralelní položky: srovnání různých položek navzájem uvnitř jedné skupiny.
- Invariance: srovnání stejných položek napříč skupinami.

4 (5) stupňů invariance:

1. Konfigurální invariance:
   - Test má stejnou strukturu (počet faktorů, přiřazení položek faktorům atd.) napříč skupinami.
   - Měří tedy obsahově „ty stejné rysy“, ale klidně úplně „jinak“.
   - Přesná definice rysů se může mírně lišit.
   - Nelze srovnávat M a SD napříč skupinami, měřítko metody je jiné.
2. Metrická (slabá) invariance:
   - Faktorové náboje v CFA jsou shodné (intercepty se mohou lišit).
   - „Definice“ latentního rysu je stejná, je měřený „ve stejném měřítku“.
   - Umožňuje srovnávat korelace latentních skórů napříč skupinami, dávat škály do jednoho modelu atd.
   - Analogie tau-ekvivalentních položek.

4 (5) stupňů invariance:

3. Skalární (silná) invariance
   - Intercepty v CFA, jsou stejné.
   - Umožňuje srovnávat průměry latentních skórů napříč skupinami.
     - Např.: Češi mají vyšší skóre v PISA testech než Slováci.
     - Např.: Pacienti v dotazníku dosahují nižšího skóre než neklinická populace.
   - Analogie paralelních položek.
   - V tomto případě má prostý součet položek stále trochu jiný „význam“ (kvůli rozdílným reziduálním rozptylům).
     - Lze ale zanedbat, má vliv jen na signifikanci srovnání skupin a velikost efektu, nikoliv na „možnost“ takového srovnání.

4 (5) stupňů invariance:

4. Reziduální (striktní) invariance
   - Položky mají v CFA modelu stejný chybový rozptyl.
   - Analogie striktně-paralelních položek.
     - Vztah součtu položek a latentního rysu je napříč skupinami stejný.

5. Paralelní skupiny
   - Na rozdíl od předchozího není vlastností jen testu, ale i skupiny.
   - Jednotlivé skupiny respondentů mají stejné průměry a rozptyly.
   - Jinými slovy, neexistuje žádný pozorovatelný rozdíl napříč skupinami v odpovědích na test.

Typické stupně invariance

- Alternativně lze fixovat vybraný faktorový náboj, nikoliv lat. rozptyl.
- Pořadí není zcela pevně dané, jen 1. a 2. krok jsou nezbytné pro všechny další;
- Krok 3 je předpokladem pro 5a a 6; 5a a 5b lze přeskočit a rovnou testovat 6.
- Pozor, v ordinální CFA a v IRT jsou určité odlišnosti!

	náboje	intercepty	rezidua	lat. průměry	lat. rozptyly
1. konfigurální	volné	volné	volné	fixované (0)	fixované (1)
2. metrická (slabá)	omezené	volné	volné	fixované (0)	ref. skup. fixované (1) další skup.: volné
3. skalární (silná)	omezené	omezené	volné	ref. skup. fixované (0) další skup.: volné	ref. skup. fixované (1) další skup.: volné
4. reziduální (striktní)	omezené	omezené	omezené	ref. skup. fixované (0) další skup.: volné	ref. skup. fixované (1) další skup.: volné
5a. ekvivalence průměrů	omezené	omezené	omezené	fixované (0)	ref. skup. fixované (1) další skup.: volné
5b. ekvivalence rozptylů	omezené	omezené	omezené	ref. skup. fixované (0) další skup.: volné	fixované (1)
6. ekvivalentní skupiny	omezené	omezené	omezené	fixované (0)	fixované (1)

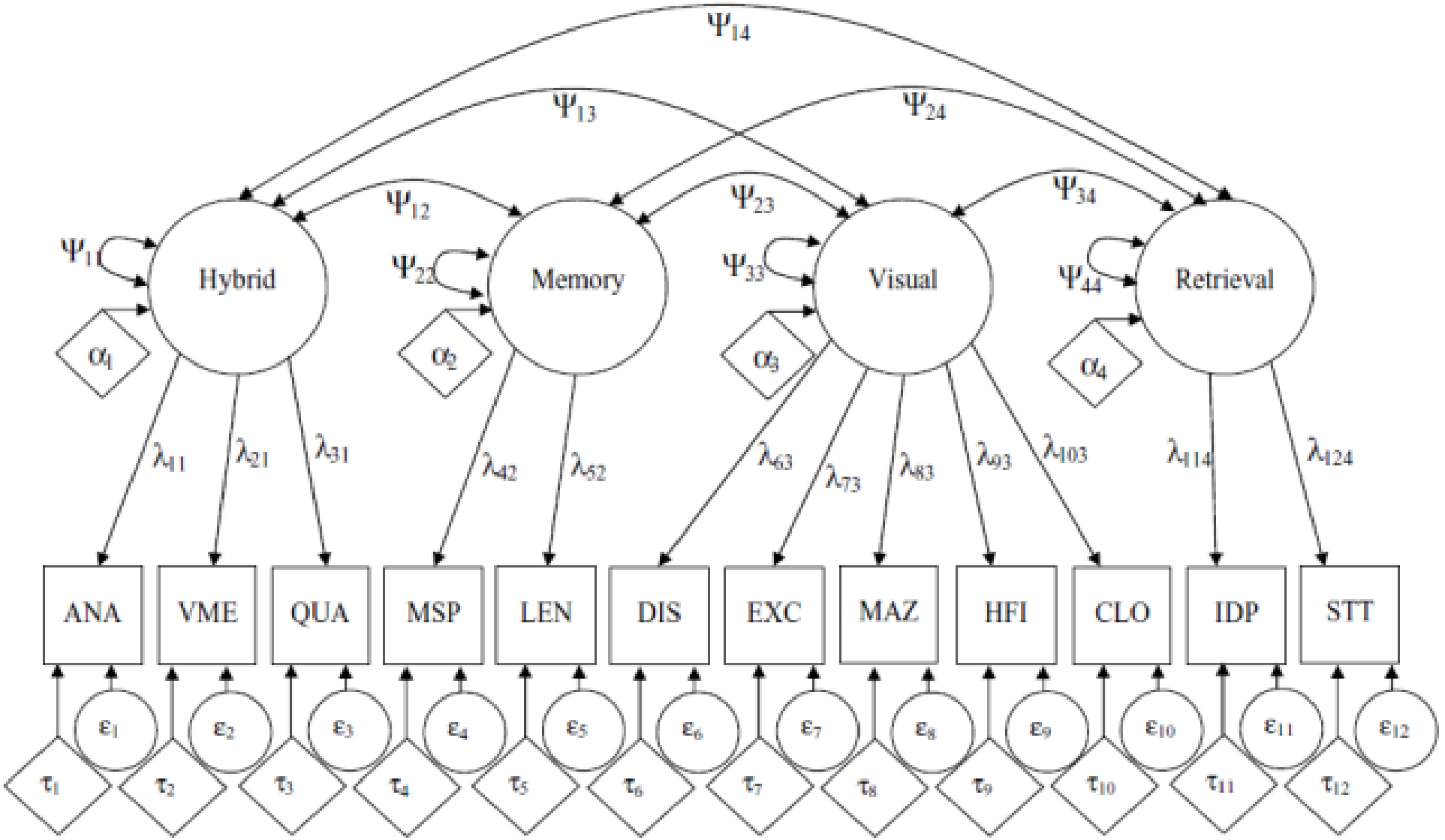


FIGURE 2. Factor model for RAKIT subtests.

Model měření pro jednu skupinu: λ_{if} – náboj pol. i na faktoru f; τ_i – intercept (průměr) pol. i; ε_i – reziduální rozptyl pol. i; α_f – průměr faktoru f; Ψ_{ff} – rozptyly nebo kovariance mezi faktory.

Alternativní způsoby ověření invariance

Multi-group CFA není jediným postupem.

Přehled všech postupů předkládá Kim, Cao, Wang and Nguyen (2017).

- Multiple group confirmatory factor analysis (MG CFA).
- Multilevel confirmatory factor analysis (ML CFA).
- Multilevel factor mixture modeling (ML FMM).
- Bayesian approximate M.I. testing (using BSEM).
- Alignment optimization.

Není nutné znát. MG CFA je zlatý standard a v psychologii postačuje.