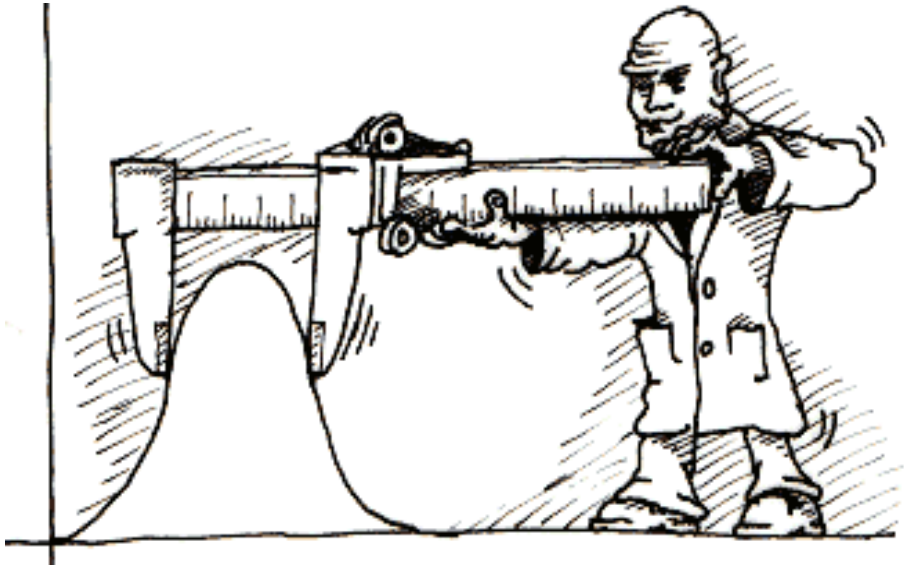


Normy a standardizace testu

PSYB2590: ZÁKLADY PSYCHOMETRIKY (PŘEDNÁŠKA 5)

27. 4. 2020 | HYNEK CÍGLER


10
Joštova
218
Brno-město
2019
VYMÝVÁNÍ MOZKŮ
pondělí - pátek: 8 - 16 hod.
(Individuální zákroky po dohodě s rektorem)
FAKULTA SOCIÁLNÍCH
STUDIÍ


2020
MASARYKOVA UNIVERZITA
NAŠE ŠKOLA SE PYŠNÍ
PRŮMĚRNÝM STUDENTSKÝM
IQ: 63,15 %
FAKULTA SOCIÁLNÍCH
STUDIÍ

2021

Standardizace testu?

Soubor veškerých postupů, které slouží jako podklad a důkazy pro *standardní* rozhodování o jednotlivcích na základě testových metod.

- Proces tvorby podkladů.
- Proces dokazování, že tyto podklady jsou validní.

Pojem standardizace je proto poměrně široký a zahrnuje:

- Důkazy validity, reliability.
- Vytvoření norem, standardních skórů.
- Kodifikace postupů pro standardní *administraci*, *skórování* a *interpretaci* skórů.
- ...

Standardizace testu

Tomáš Urbánek (2010): 3 pojetí standardizace.

- **I. povrchní:** *„... **metoda je přesně popsána** [...] jak má vypadat např. testový materiál, pomůcky nebo testový sešit, na jakém papíře, v jakých barvách a jakým písmem apod. Kromě toho je jasně definováno, **jak má být metoda používána**, tzn. komu, kým a za jakých podmínek smí být administrována, **jak má být vyhodnocována** a co znamenají získané výsledky.“*
- **II. klamavé:** *„... se spokojuje s existencí **jakýchkoli norem** ve smyslu popisu, jakých výsledků dosahují respondenti z nějakých jasně definovaných skupin. Ani tento požadavek není obtížné splnit, stačí jen použití metody spojit se sběrem dat a elementární statistickou prezentací výsledků...“*
- **III. komplexní:** *„... Součástí tohoto pojetí jsou i obě pojetí předchozí [...] je přinejmenším nutné prokázat, zda metoda měří daný atribut (**validita** a validizace) a s jakou přesností (**reliabilita**). Současně je nutno vyřešit **všechny speciální otázky**, které mohou nastat v souvislosti s testováním specifických charakteristik. To je pojetí uváděné např. ve Standardech pro pedagogické a psychologické testování (AERA, APA, NCME, 2001), ale je doporučováno i [..., EFPA].“*

Manuál diagnostického testu

Teoretická východiska: Co je měřeno, jaké jsou známé souvislosti, proč se to měří.

- Účel metody: Komu, kdy, proč, kým, kde...

Postup administrace a skórování: jak přesně se metoda zadává a skóruje.

- Tvorba hrubých skórů a převod na standardní/vážené skóry.

Postup interpretace: co výsledky znamenají.

- Součástí zpravidla i kazuistiky.

Psychometrický manuál: Dokládá výše uvedené na vzorku z cílové populace.

- Standardizační soubor, postupy konstrukce norem.
- Důkazy validity, reliability vzhledem k účelu metody.

Co z výše uvedeného je možné pouze přeložit ze zahraniční verze testu?

Tvorba testu

Značné rozdíly mezi metodou určenou pro výzkum a pro individuální diagnostiku.

Tvorba nové testové metody.

- + Kulturně adekvátní metoda.
- - Tvorba může selhat, vysoké nároky na přípravu...
- - Vysoké finanční náklady na průběžné pilotáže, analýzy...

Adaptace zahraniční metody.

- Překlad vs. adaptace.
- + Zpravidla ověřená metoda, nižší nároky na velikost vzorku, pilotáže, méně práce.
- + Lze využít zahraniční důkazy validity, většinou rozsáhlejší teorie.
- - Cena licence (i několik milionů Kč), často časově omezená, poplatky...
- - Standardizační studie stejně musí být realizována.

Design standardizační studie 1

Volba výběrové populace (pro koho je test určený)?

- Mezinárodní, národní, lokální, místní...

Kognitivní pilotáž (kvalitativní metodologie).

Způsob výběru vzorku a administrátorů.

- Náhodný, stratifikovaný, clusterový, příležitostný... Plánovaně chybějící data.
- Tvorba adekvátních clusterovacích proměnných (ČSÚ).
- Inkluzivní a exkluzivní kritéria.
- Zaškolení a výběr administrátorů.

Sběr dat.

- Párování respondentů s administrátory. Jak zajistit ortogonalitu?
- Kódování dat.

Design standardizační studie 2

Přepis dat, kontrola správnosti.

- Vyčištění dat, spárování datasetů atd.

Vážení respondentů?

- Clusterový/stratifikovaný výběr. Bude váženo vše/nic?

Položkové analýzy, analýzy reliability, validity.

- Uvnitř napříč kohortami?

Tvorba norem.

- Vytvoření vyhodnocovacího softwaru, normalizačních tabulek...

Zkompletování manuálu.

Prodej. Spotřební materiál volně dostupný?

Příklad nákladů: BACH

Školní dovednosti, cca 25 subtestů (čeština a matematika), cílová populace 5-22 + 60-80 let.

Konormace s WJ-IV, resp. TOMAL-II-S; 2,5 roku vývoje, 3 roky standardizace v rámci projektu.

Pilotáže a vývoj (velmi hrubě): **1.100.000 Kč**

- Náklady na vývoj testu: 500.000 Kč.
  - Tahle částka reálně nebyla vyplacena.
  - Zčásti je „sanována“ náklady na jiné personální náklady v rámci projektu.
- 500 individuálních administrací: 400.000 Kč.
- přepis dat: 50.000 Kč.
- tisk, poštovné: 150.000 Kč

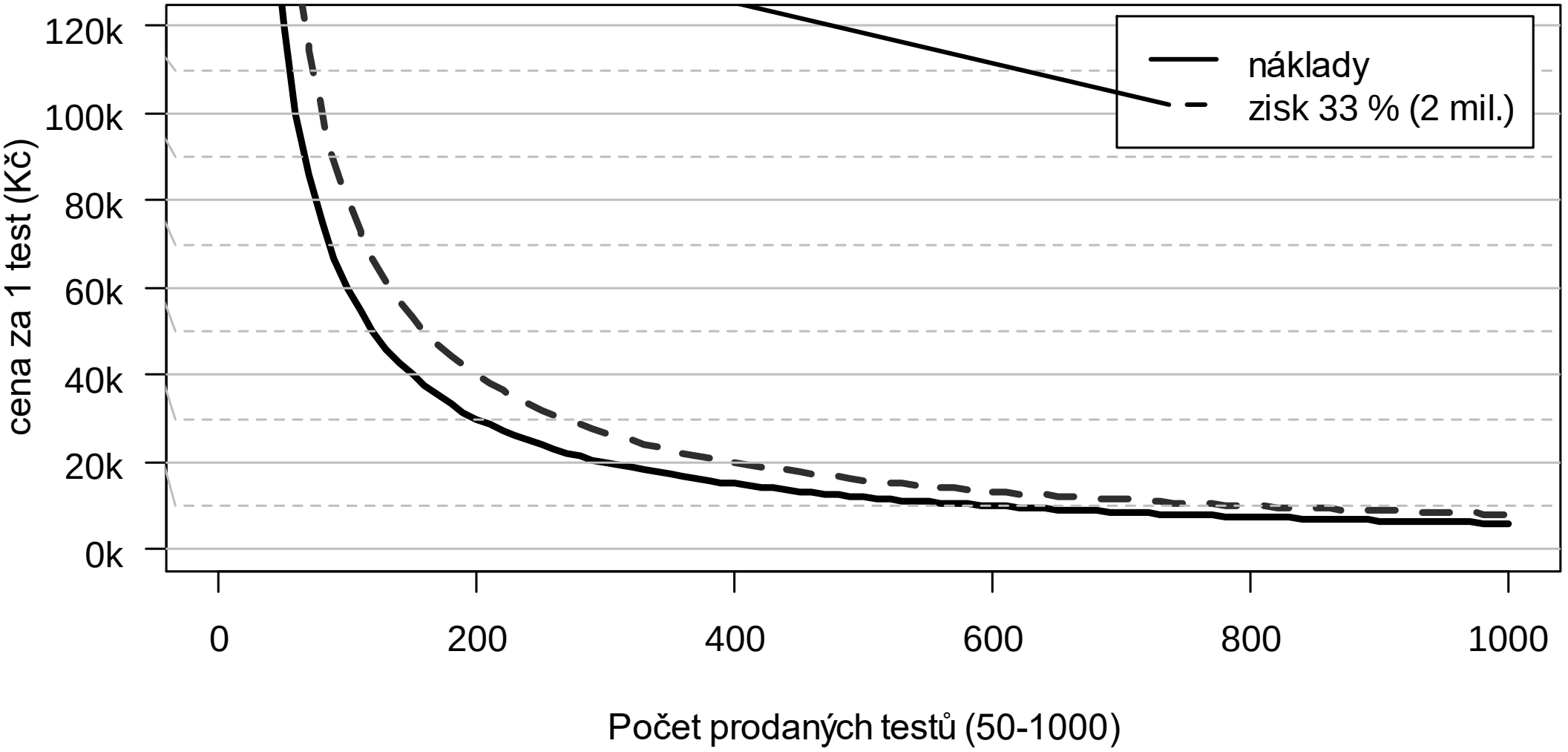
Standardizace (TAČR): **4.906.248 Kč**

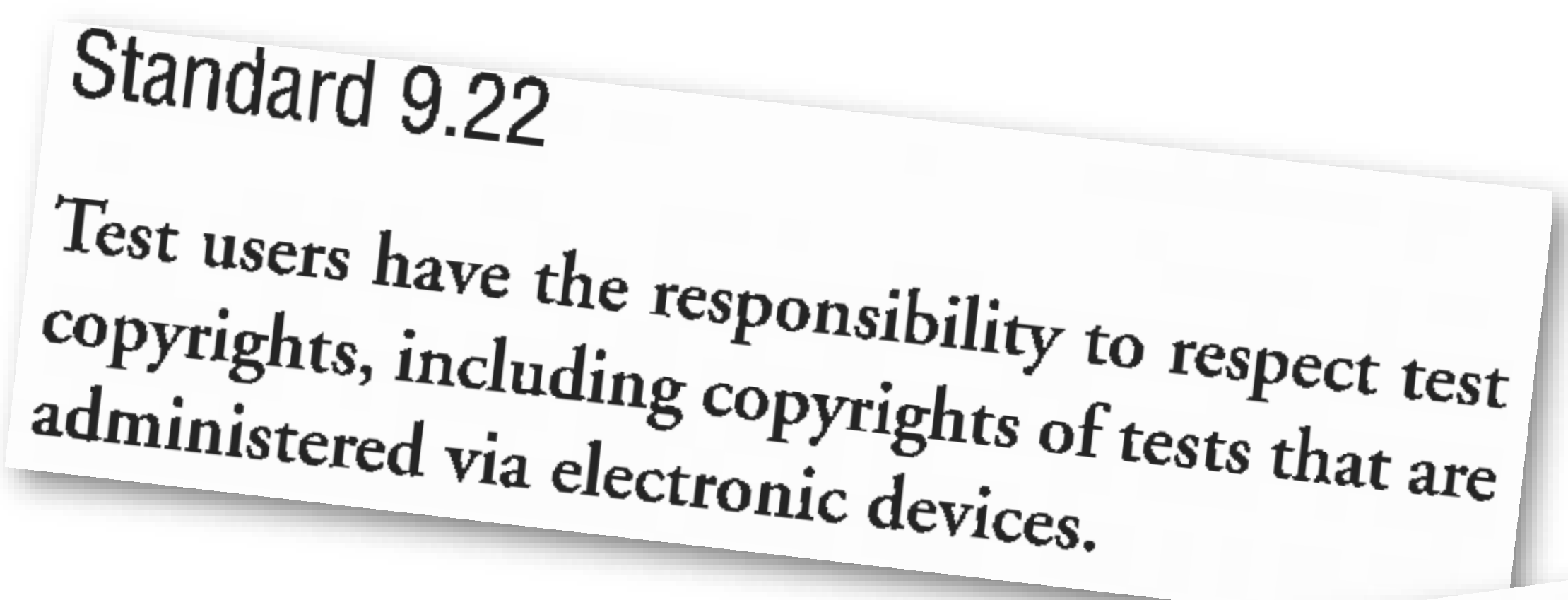
- Sběr dat: 2.000.000 Kč.
- Odměna respondentům: 140.000 Kč
- Přepis dat: 170.000 Kč
- Školení (lektoři): 50.000 Kč
- Tisk, nahrávací studio, grafika apod.: 440.000 Kč
- Poštovné: 80.000 Kč
- Personální náklady jiné: 1.000.000 Kč.
- Režie, nájmy...: 800.000 Kč

Celkové náklady: **cca 6 milionů Kč.**

- (V personálních nákladech se zčásti překrývá pilotáž a standardizace).

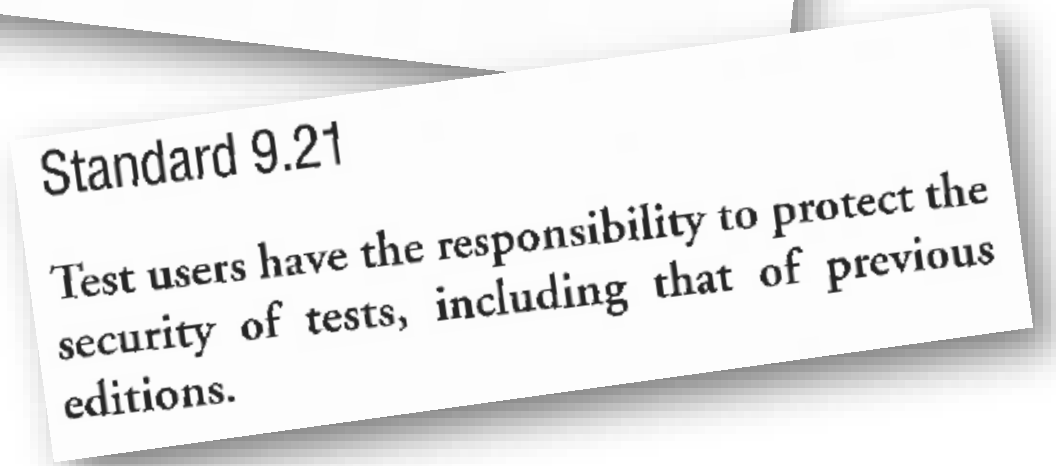
Cena 1 testu podle množství prodeje





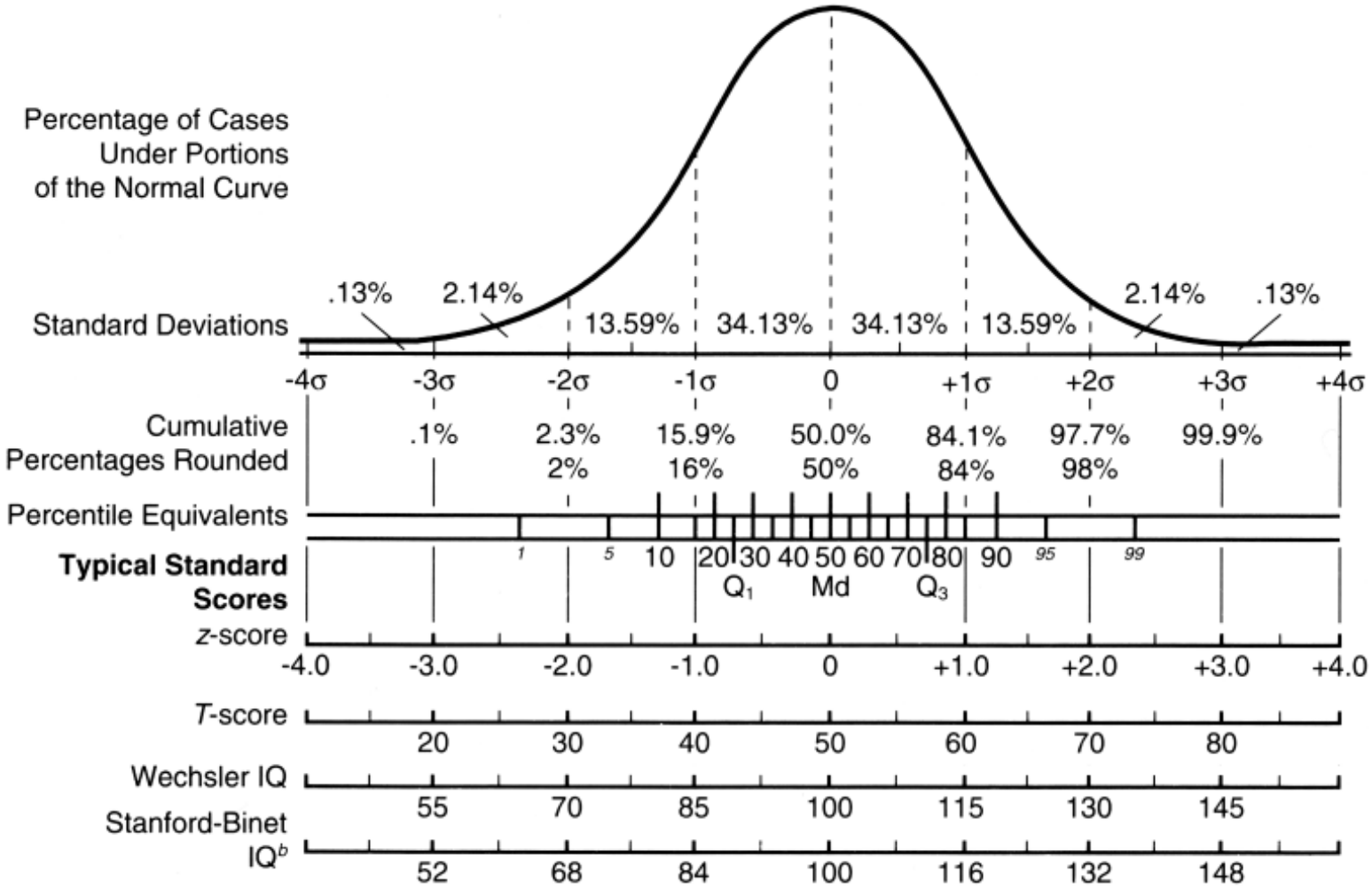
„1. přikázání psychologické diagnostiky:“

Nezkopíruješ!



AERA, APA, & NCME. (2014).
Standards for Educational and Psychological Testing.
Washington: American Educational Research Association.

Normy


Percentage of Cases Under Portions of the Normal Curve
.13%
2.14%
13.59%
34.13%
34.13%
13.59%
2.14%
.13%
Standard Deviations
-4σ
-3σ
-2σ
-1σ
0
+1σ
+2σ
+3σ
+4σ
Cumulative Percentages Rounded
.1%
2.3%
2%
15.9%
16%
50.0%
50%
84.1%
84%
97.7%
98%
99.9%
Percentile Equivalents
1
5
10
20
30
40
50
60
70
80
90
95
99
Typical Standard Scores
Q1
Md
Q3
z-score
-4.0
-3.0
-2.0
-1.0
0
+1.0
+2.0
+3.0
+4.0
T-score
20
30
40
50
60
70
80
Wechsler IQ
55
70
85
100
115
130
145
Stanford-Binet IQb
52
68
84
100
116
132
148

K čemu jsou normy

Snaha vyhnout se **chybám intepretace**. Normy dávají smysl výsledkům testování:

- Porovnáním výsledku s výsledky populace;
- porovnáním výsledku s kritériem;
- porovnáním výsledků navzájem napříč testy.

Zamezení osobních chyb, svévolné intepretaci „čísel“.

Částečné „překonání“ problému měření v sociálních vědách.

- Normy jsou tím, co udává „škálu“ měření. Tvorba „jednotek“ (IQ apod.).
- Důsledkem je často neoprávněná reifikace výsledku měření
  - Např. ztotožnění IQ = inteligence.

Proč vlastně normy?

Proč jsou normy v psychologii nezbytné?

I pokud by měření v psychologii bylo intervalové, není poměrové.

Neexistuje tedy jasně definovaný referenční bod.

- Referenční bod je nutné stanovit arbitrárně.

Je nutné zvolit i jednotku; typicky je závislá na vzorku (populaci).

- Na čem bývá založena jednotka např. ve fyzice?
- Proč je jednotka závislá na vzorku problém?
- Šlo by to řešit jinak?

Typy norem

„Klasické“ normy

- Mezinárodní, národní normy, místní normy.
- Nahodilé normy (více různých specifických populací v případě, že není dostupný reprezentativní vzorek).
- Uživatelské normy.
- Lokální normy, normy pro specifické populace.

Referenční (normy) vs. **kriteriální** (arbitrární kritérium) testování.

Expektační tabulky – odhady pravděpodobnostího výskytu jevu, klinické odpovědi apod.

- Nepředpokládá náhodný vzorek, spíše vypárovaný oproti pacientům.
- Často v podobě grafu pravdivě vs. falešně pozitivní odpovědi. Podobné ROC analýze.

„**Typologie**“ – specifický příklad ipsativních skórů. Pozor na ně!

- Specifické nároky na data, na měřený atribut.
- Kontinuální rys by měl mít bimodální rozdělení.

Druhy skórů[1]

HRUBÉ SKÓRY

Sumační indexy – prostý součet položek.

- Nebo průměr, který má výhody i nevýhody.

Lineární kombinace – každá pol. má jinou váhu, např. na základě faktorové analýzy.

- Někdy též vážené nebo kompozitní skóry.

IRT odhady (Analogie hrubých skórů v CTT – theta, EAP/MAP, W-skóry)

ODVOZENÉ SKÓRY (VŠE OSTATNÍ)

Percentilové skóry: Percentily, decily, percentilové pořadí a další (kvantily, percentilové rozpětí...), steny, staniny...

Standardní skóry: IQ(100;15), T(50;10), T(500; 100), z-skóry, Wechslerovy vážené skóry W(10;3)...

Vývojové skóry: Mentální věk (age-equivalent score, grade-equivalent score), index relativní výkonnosti (RPI), zóna vývoje

Ipsativní skóry

[1] http://prirucka.ujc.cas.cz/?slovo=skóre ☺

Standardní skóre

Lineární transformace hrubých skórů na odvozené. **Z-skór:**

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

Standardní skór:

$$S = \sigma_S \cdot z + \bar{S} = \frac{\sigma_S}{\sigma_X}(X - \bar{X}) + \bar{S}$$

- S – standardní skór, σ_S – cílová SD, z – z-skór, $\bar{S}$ – cílový průměr, σ_X – SD HS, $\bar{X}$ – průměr HS, X – hrubý skór.

Předpoklady:

- Průměrně/přiměřeně obtížné položky a tedy i **normální rozdělení hrubého skóru**.
- Pokud předpoklad neplatí: nelineární transformace podle tabulky. Kde se vezme ta tabulka? ☺

Normalizace rozložení (mírné zešikmení)

McCallova plošná standardizace.

- Každému X je přiřazeno percentilové pořadí.
- Percentilům je přiřazen T-skór za předpokladu normálního rozdělení.
- + teoreticky „dobré“ vyhlazení.
- - percentily jsou zatížené vysokou výběrovou chybou.

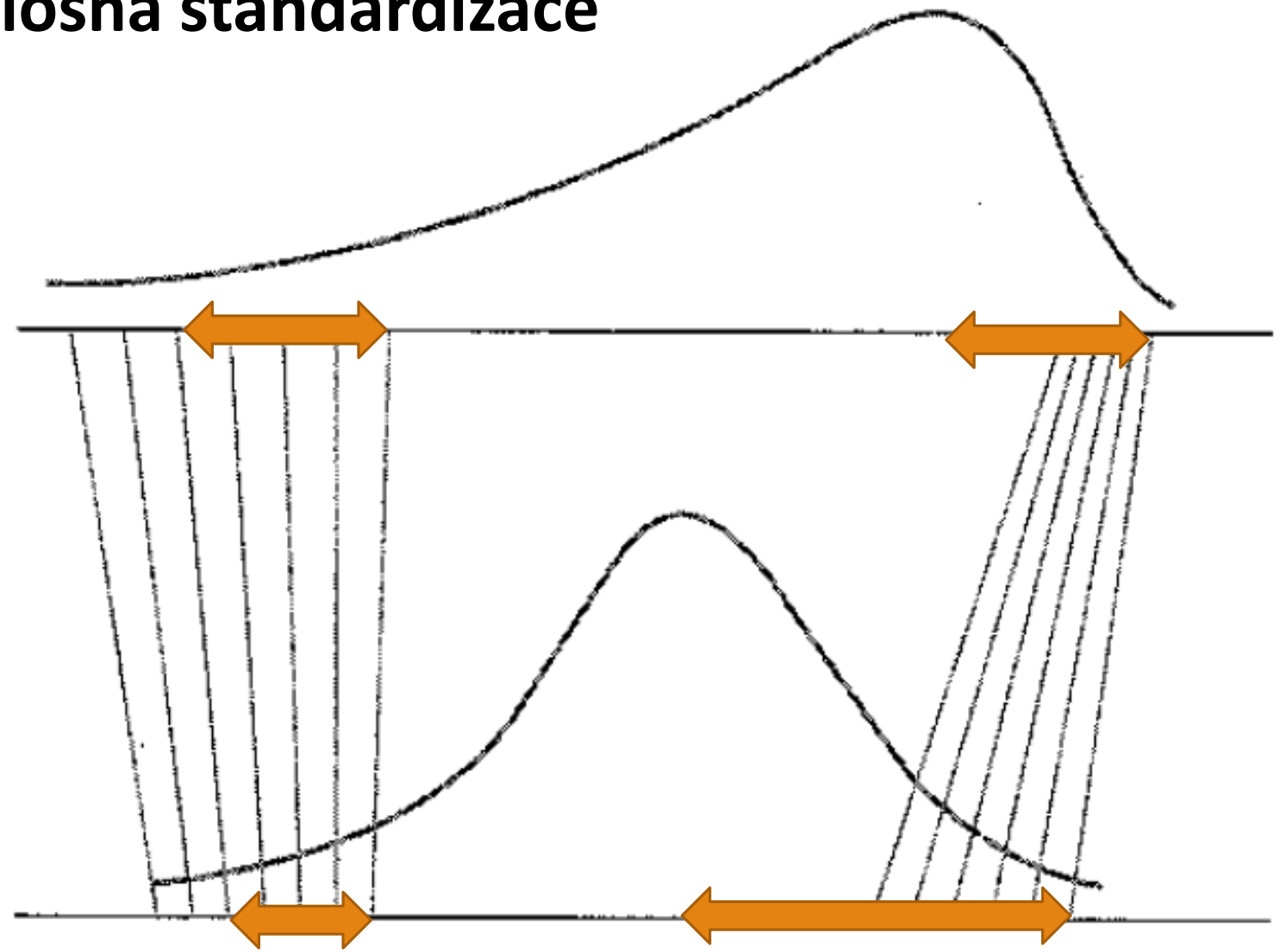
Normalizace podle mediánu.

- Samostatná SD pro lepší a horší respondenty.
- + odhad jen 3 parametrů (M, SD_{lower}, SD_{upper}), menší výběrová chyba.
- - slabší vyhlazení, obtíže s konstrukcí CI.

Jiné nelineární transformace včetně kontinuálního normování.

Transformuje se nejen skór, ale i jeho SE!

McCallova plošná standardizace



Normalizace podle mediánu

Nejjednodušší způsob normalizace skórů.

Předpoklady:

- Normální rozdělení má průměr (přibližně) shodný s mediánem.
- Předpokládáme, že každá polovina rozložení sama o sobě odpovídá přibližně normálnímu nezešikmenému rozložení, jen s jinými parametry.

Postup:

- 1. Rozdělíme respondenty na dvě poloviny podle mediánu.
- 2. Ručně spočítáme SD horní a dolní poloviny.
  - Nejde o SD uvnitř poloviny, ale odhad SD napříč polovinami, když by druhá polovina měla stejné, avšak zrcadlově otočené rozložení.
- 3. SD použijeme zvlášť pro výpočet SE a SS v obou polovinách.
  - Co s přechody přes medián a SE, CI?

Např. Woodcock-Johnson IV us (výhodné při vyhlazování skórů).

Extrémní zešikmení = problém

Výrazný efekt stropu nebo podlahy.

- Velká komplikace – ideálně by žádný respondent neměl mít max. nebo min. skóre.
- V případě těžkého testu i zisk jediného bodu hrubého skóre posune respondenta velmi výrazně na všech škálách (percentil, stand. skóre).

Extrémně snadný/obtížný test.

- Např. při měření patologie.

Neexistuje kontinuální latentní proměnná, ale kvalitativní latentní „třída“.

V těchto případech je standardní skóre nevhodné.

- Percentil nebo kriteriální skórování.

Doporučení ke standardním skórům

Veškeré skóry jsou zaokrouhleny na celá čísla (kromě z-skórů, ty na 2 desetiny).

APA doporučuje T-skóry; IQ skóry výhradně pro měření výkonu v kognitivních testech.

Se skórem je vždy reportována chyba, např. formou CI (doporučuje se 90%).

- Vyjma stenů a staninů.

Steny a Staniny jsou považovány za „rozpětí“, konstruovány jsou na základě plošné transformace.

- Steny N(5,5; 2), staniny N(5; 2).
- Spíše marginální použití.

Percentily

Procento osob, které mají **horší** hrubé skóre než hrubé skóre daného člověka.

- U škál s malým množstvím možných skórů prakticky nejde dosáhnout percentilu 100.
- Percentilové pořadí (percentil rank) – **stejné nebo horší** hrubé skóre.
- U dlouhých škál je rozdíl zanedbatelný, u krátkých je potřeba vědět, s čím pracujeme.
  - V případě nespojité proměnné (v psychologii prakticky vždy) se liší percentily a percentilové pořadí liší.

Odhad většinou na základě pozorovaného rozložení a ne normální distribuční funkce.

- Naopak standardní skóry často založené na percentilu (viz McCallovu plošnou transformaci).
- Ale co chyba měření a výběrová chyba? Může vést k rozdílu percentilu a standardního skóre.
- Je zvykem „vyhladit“.

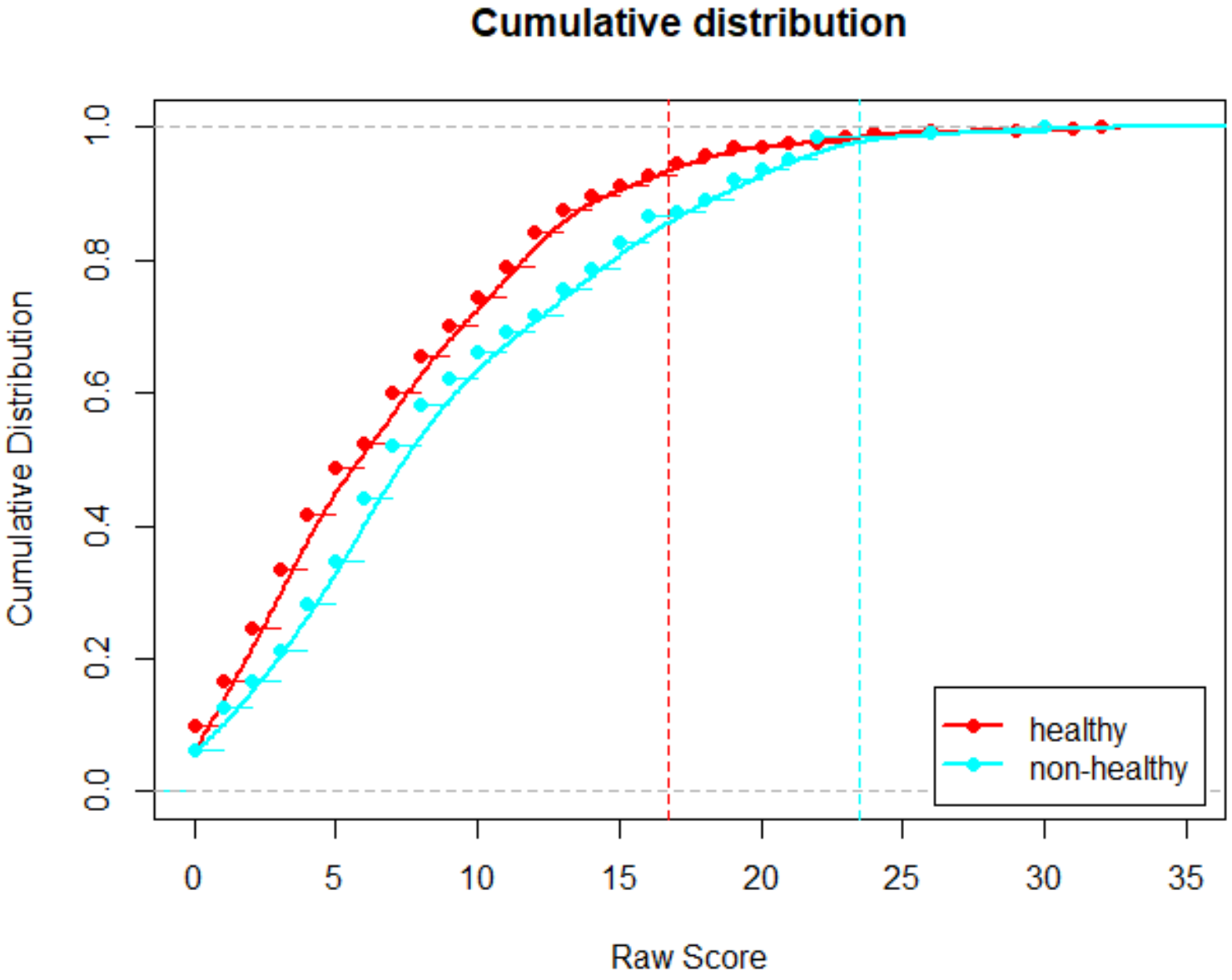
Příklad na vyhlazení percentilových norem

Beckův inventář depresivity (BDI)

- Svislé části označují dva kritické skóry.
- N_h = 450
- N_n = 127

balíček ks v R

- Kernel cumulative distribution



Vývojové skóry

Věkové ekvivalenty (age equivalent) – jakému věku odpovídá dané skóre?

- Věk, v němž respondenti průměrně dosahují daného skóre.
- Analogie „mentálního věku“ (Binet) – dnes se tento termín nepoužívá.

Ročníkový ekvivalent – totéž, ale pro ročník/třídu.

Zóna vývoje – věkové skóre v podobě rozsahu.

- Rozsah na základě chyby měření, nebo častěji na základě stadiální křivky vývoje.

Raschovské skóry

Kategorie skórů založená na Teorii odpovědi na položku (IRT), konkrétně 1parametrovém (Raschově) IRT modelu.

- Viz poslední přednáška.

Analogie hrubého skóre v CTT. Výhodnější např. pro sledování vývoje.

W-skóre.

- Referenční bod: Právě 10leté děti mají průměrně W=500.
- Univerzální *jednotka*: Pokud někdo s W=A má 50% pravděpodobnost na správnou odpověď na určitou položku, pak někdo jiný s W=A+10 má 75% pravděpodobnost, resp. W=A-10 25%.

Index relativní výkonnosti

- RPI – Relative Proficiency Index
- Ve formátu XX/90, např. 47/90 nebo 94/90.
- „S jakou pravděpodobností respondent zvládne úkol, který jeho vrstevníci zvládnou s 90% pravděpodobností?

Ipsativní skórování

Nejsou zpravidla skóry v pravém slova smyslu:

- Standardní skóry srovnávají interindividuální variabilitu.
- Ipsativní skórování srovnává intraindividuální variabilitu.

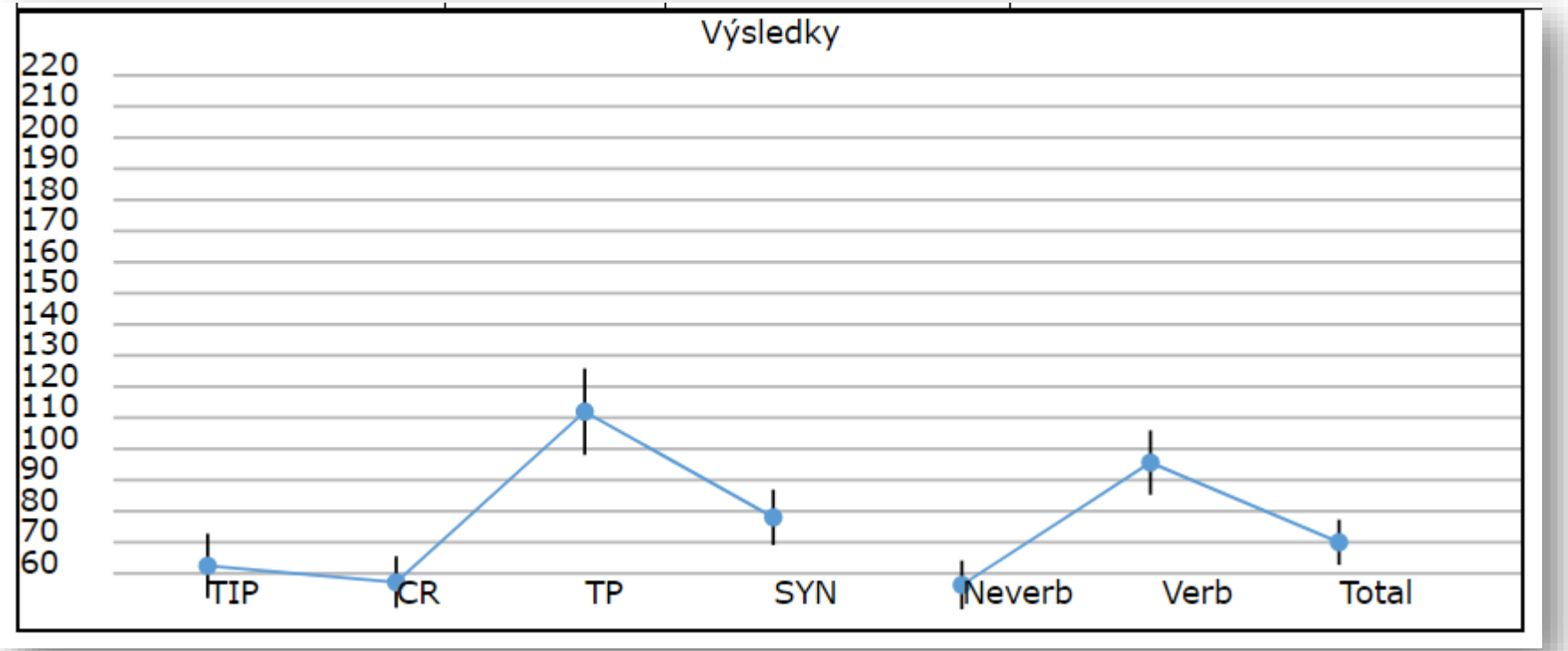
Založené na diskrepanci

- Předpokládáme, diskrepance mezi subtesty/faktory v inteligenčním testu může ukazovat na SPU.
- Kariérní poradenství – dotazník volby povolání („co člověka baví víc?“).

Používá se standardní chyba rozdílu, případně je rozdíl subtestů přímo standardizován.

Analýza profilu.

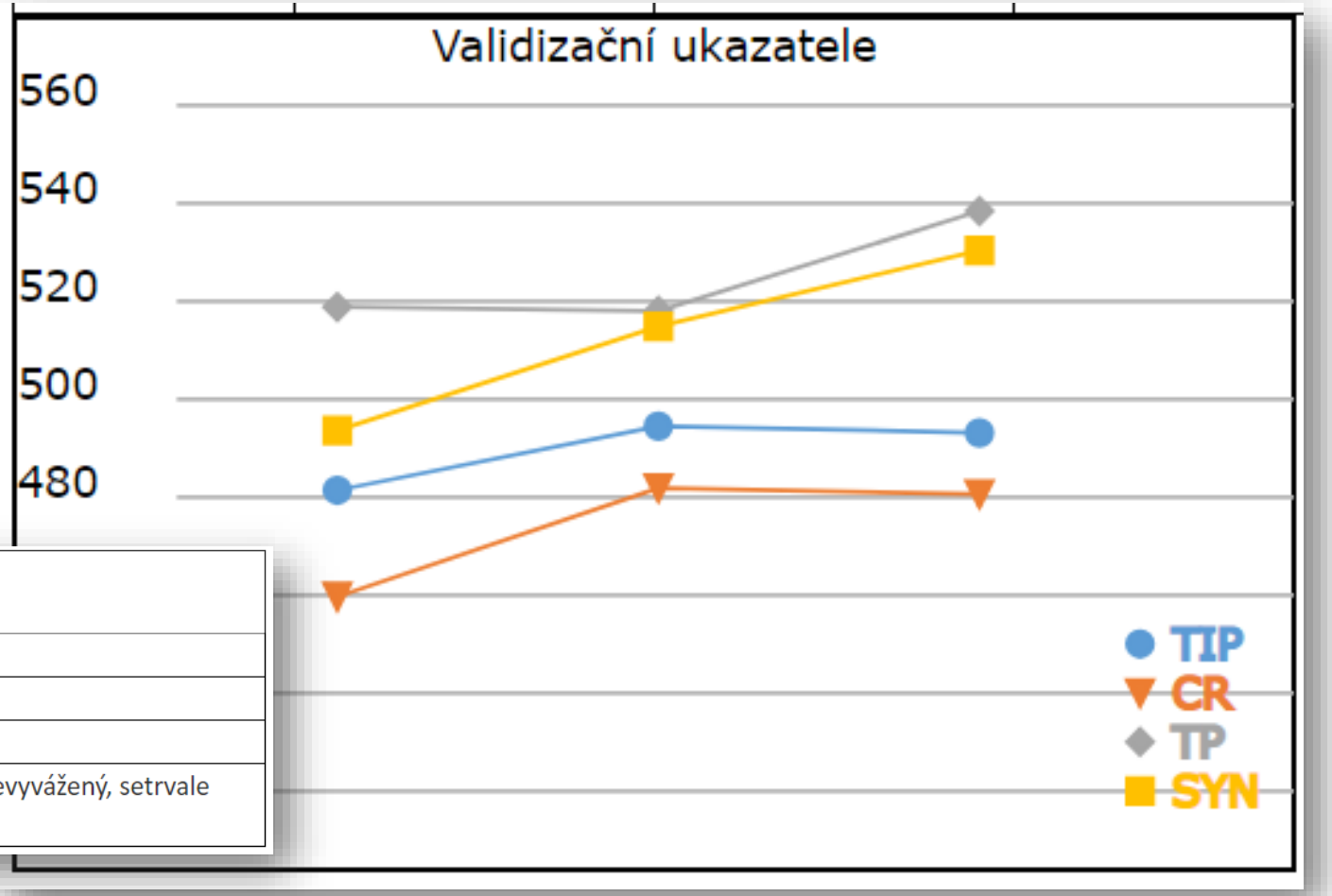
Ipsativní skórování (více testů)



Krátký inteligenční test

Ipsativní skórování (v rámci testu)

Krátký inteligenční test



Validizační ukazatele	Část 1	Část 2	Část 3	χ^2(df = 2)	p	pozn
TIP	481,37	494,44	493,09	1,6	0,449	
CR	459,63	481,83	480,58	3,816	0,148	
TP	518,76	517,97	538,37	2,11	0,348	
SYN	493,55	514,81	530,31	12,677	0,002	Pozor, v subtestu je patrný nevyvážený, setrvale rostoucí výkon.

Součtové/vážené skóry

Příklad: Máme inteligenční test. Chceme spočítat celkový skór (g-faktor). Můžeme:

- 1. sečíst všechny položky napříč subtesty.
- 2. standardizovat každý test a pak sečíst subtesty.
- 3. standardizovat každý test a vzít jejich vážený součet.

Výhody? Nevýhody?

Hlavní komplikace:

- Nelze sčítat nevážené subtesty (a tedy ani položky), mají jinou SD.
- Nelze předpokládat, že všechny vážené subtesty mají stejný vztah s g-faktorem. Na rozdíl od položek nepředpokládáme „náhodný výběr“ z domény.
- Efekt stropu, podlahy. U dětí různé *„váhy“* pro různé referenční skupiny.
- Vliv chyby měření (testy s nižší reliabilitou mají nižší váhu). Různá chyba pro různé referenční skupiny.

Např. Wechsler: součet standardizovaných subtestů.

Např. Woodcock-Johnson: vážený průměr nestandardizovaných subtestů.

Součtové/vážené skóry

Formativní vs. reflektivní měření na druhé úrovni vzhledem k chybě měření.

A. Reflektivní celkový skór.

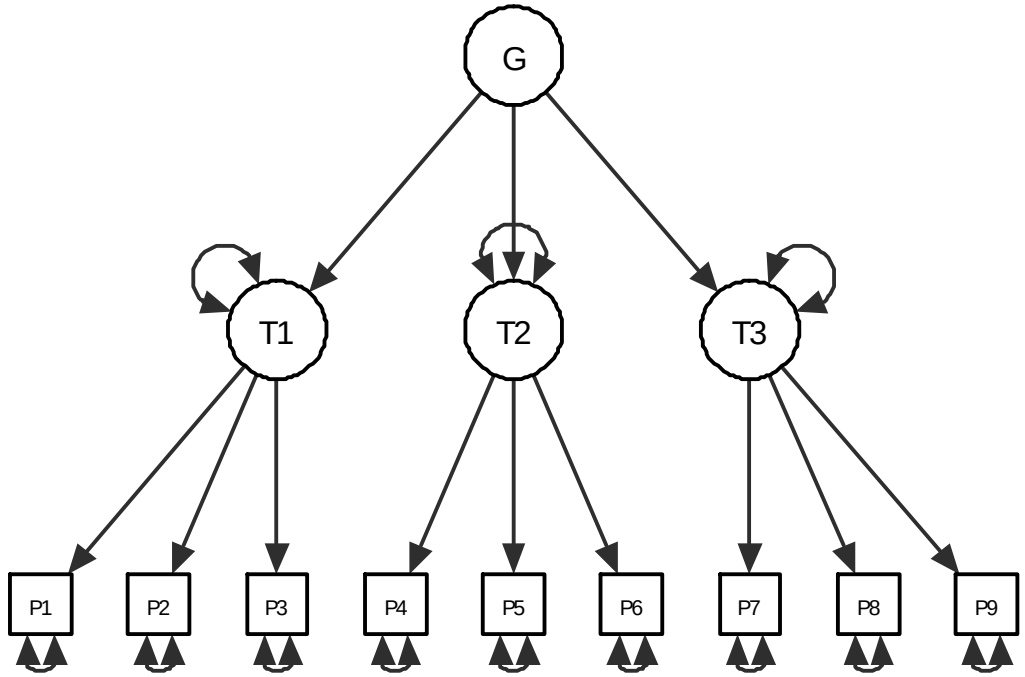
- Celkový skór je odhad g-faktoru.
- Specifické rozptyly považovány za chybu.
- Vyšší míra chyby měření.

B. Formativní celkový skór.

- Celkový skór je jednoduše průměrem subtestů.
- Specifické rozptyly nehrají roli.
- Nižší míra chyby měření.

Zpravidla testy používají variantu B.

Různé odhady reliability/chyby měření.



Konstrukce norem

Protože je důležitá reprezentativita, normy by měly být *vážené*.

- Každý jednotlivý respondent přispívá jinou váhou ke tvorbě norem (výpočtu M, SD...) či dalším analýzám (zejm. FA) podle toho, jak často jsou jeho demografické charakteristiky zastoupené ve vzorku.
- Např. velikost sídla, vzdělání (rodičů), věk, typ SŠ, pohlaví...

Zpravidla pro každou věkovou/referenční kategorii (kohortu) spočítáme zvlášť.

- Včetně odhadu reliability.

Tabulka přepočtů HS na odvozené skóry + chyby měření pro každou kohortu.

Občas, např. v klinické psychologii, se pracuje i s regresními funkcemi.

- Predikce skóru na základě pohlaví, vzdělání, věku...

Čím „podrobnější“ referenční kategorie, tím vyšší výběrová chyba. Vznikají nepřesnosti.

- Např. stejné hrubé skóre odpovídá *vyššímu* výkonu u starších dětí.

Vyhlazení norem (kontinuální normování)

Pro zpřesnění norem, zmenšení výběrové chyby či snížení požadavků na velikost vzorku se používá vyhlazení.

- **Horizontální vyhlazení** – *napříč* referenčními kategoriemi.
- **Vertikální vyhlazení** – *uvnitř* referenčních kategorií.

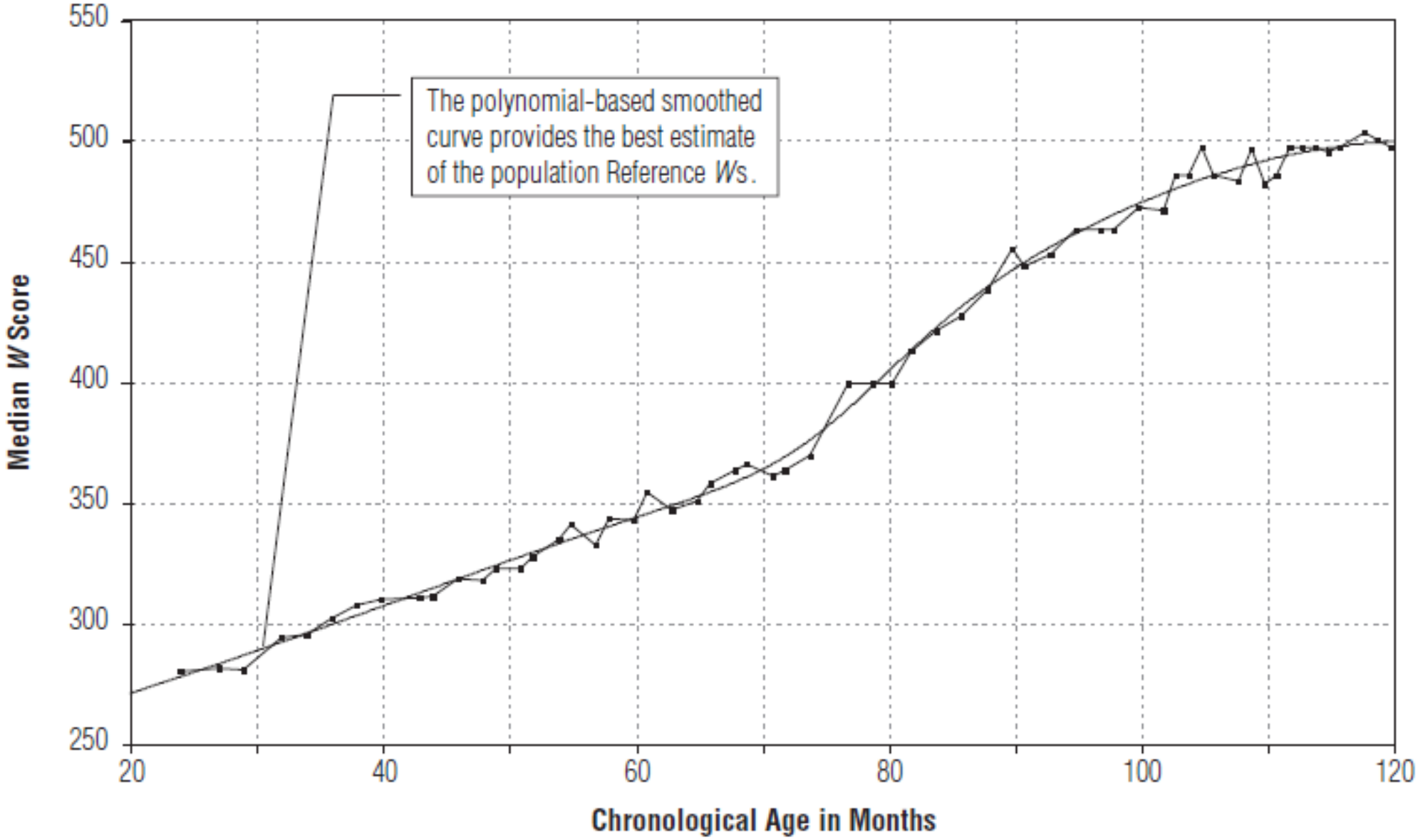
Princip: V případě 20 kategorií bychom potřebovali 20×2 (M+SD) = 40 parametrů.

- Vyhlazení odhadne všech 20 průměrů s pomocí 3 parametrů. Ušetří se informace.
- Dílčí odchylky od reprezentativity souboru dostávají nižší význam.

Celá řada postupů pro různé druhy vyhlazení.

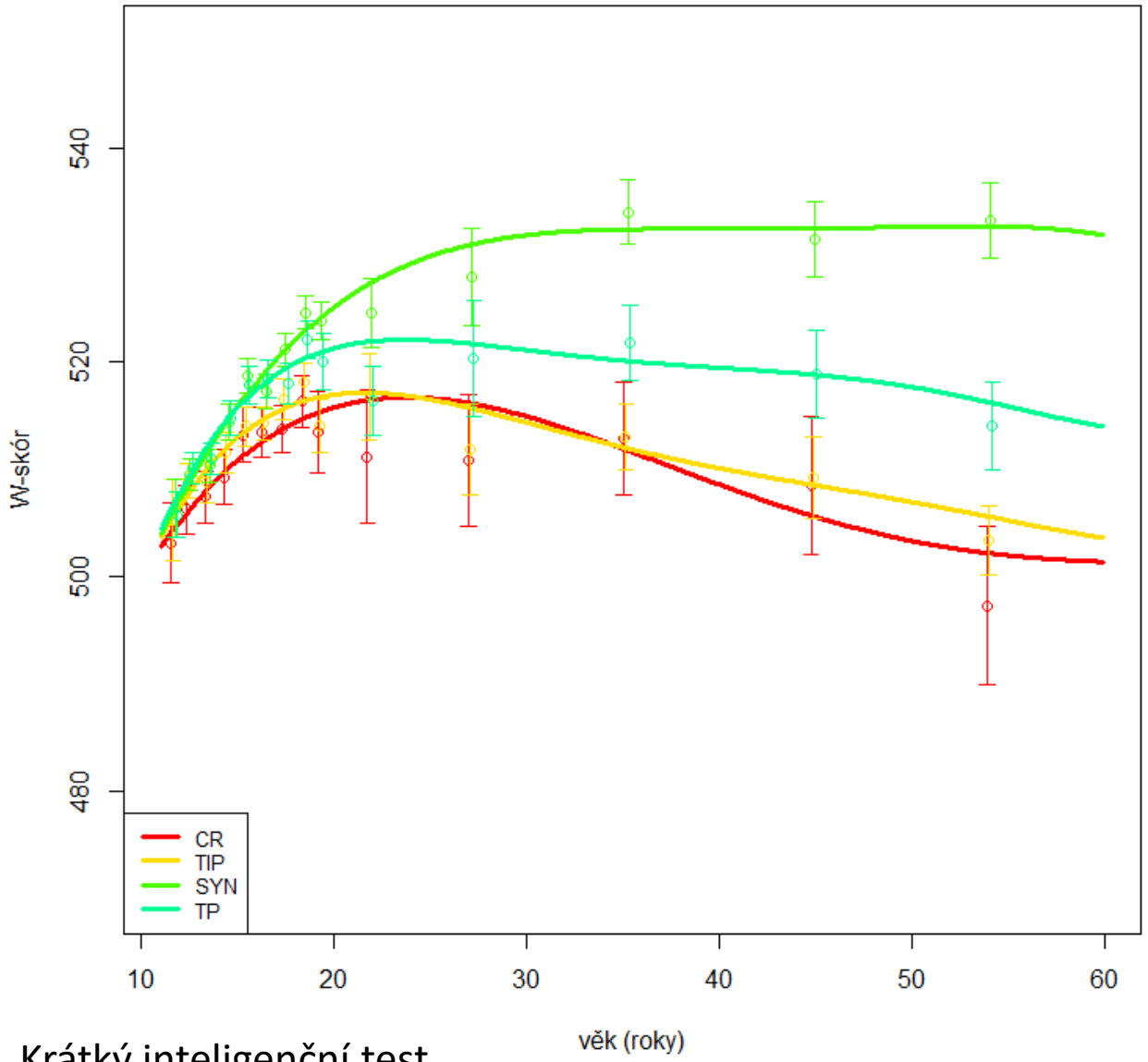
- „Ruční“ korekce/vyhlazení ☺
- Kernel density smoothing – pro vyhlazení percentilů uvnitř kategorie.
- Polynomy, frakční polynomy, spline smoothing – vyhlazení M, SD napříč kategoriemi.
- Vyhlazení prostřednictvím Taylorových polynomů (R balíček cNORM).

Další výhoda: Normy s přesností na měsíc či den (není nutné mít „široké“ kohorty).

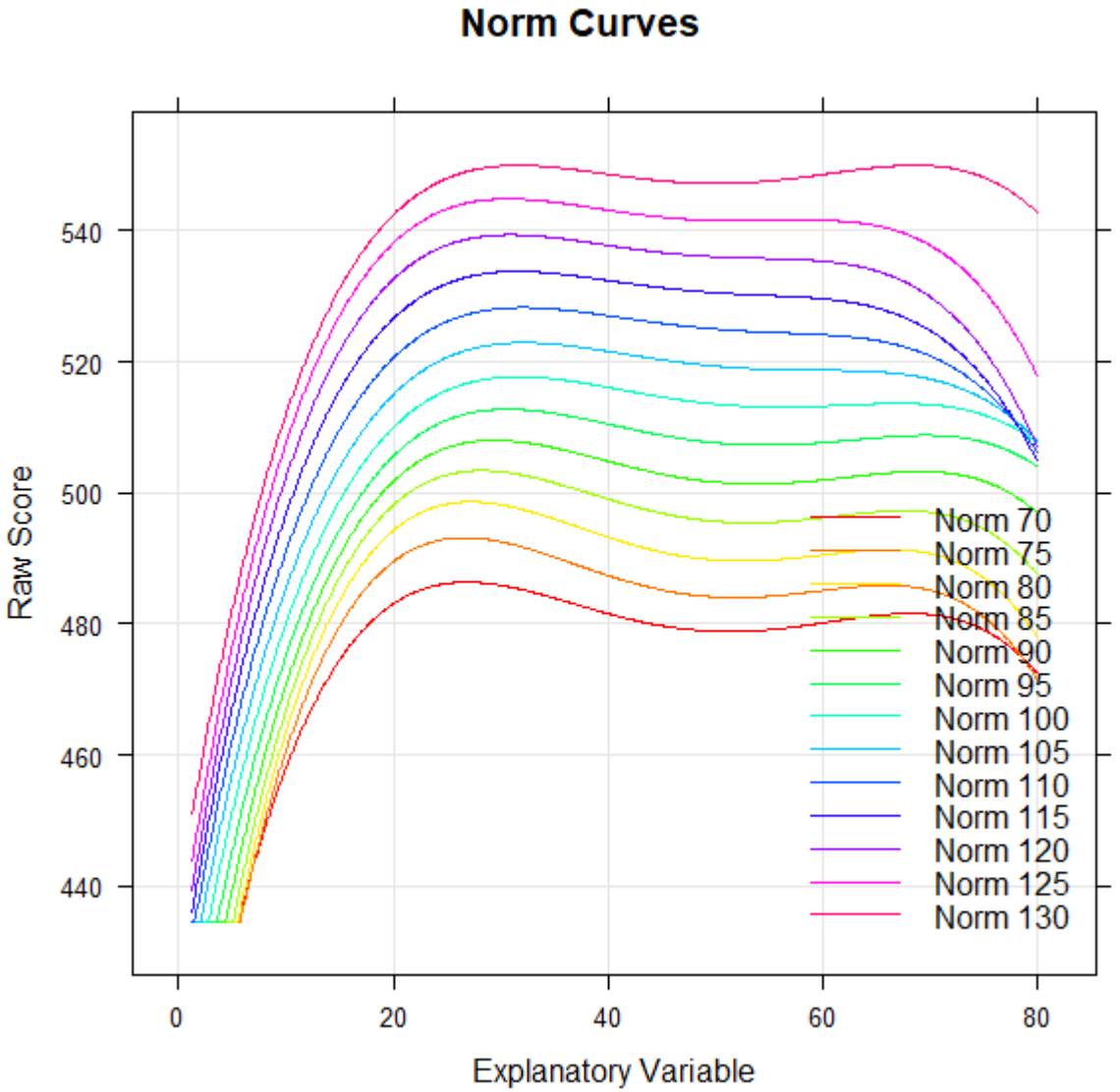


Note. A similar process is completed with the standard deviations (*SD*s) for Letter-Word Identification.

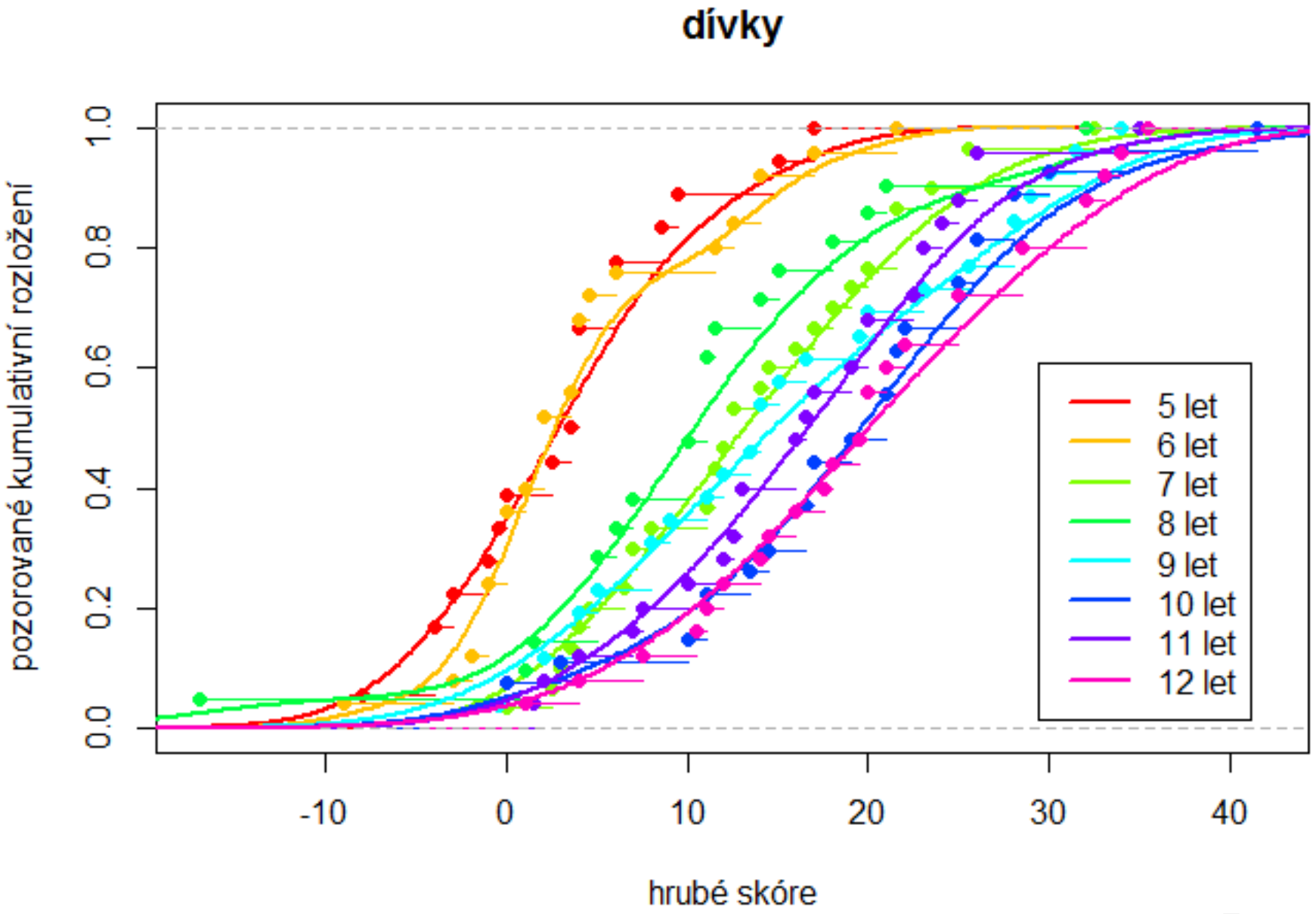
Technický manuál testu WJ-IV



Krátký inteligenční test



WJ-IV COG CZ (pracovní analýzy)
cNORM package



Test vytváření příběhů