**Univerzita Karlova v Praze**

**Filozofická fakulta**

**Katedra psychologie**


# Diplomová práce

Jana Dlouhá

## Počítačové adaptivní testování a možnosti jeho využití v psychodiagnostice

## Computerized Adaptive Testing and its Use in Psychodiagnostic

Praha 2013

Vedoucí práce: doc. MUDr. Mgr. Radvan Bahbouh, PhD.

Na tomto místě bych chtěla poděkovat vedoucímu své diplomové práce doc. MUDr. Mgr. Radvanu Bahbouhovi, PhD. Děkuji za podporu, cenné připomínky, rady a doporučení.

Ráda bych poděkovala také svým rodičům a prarodičům, bez nichž bych tuto práci nejspíš nikdy nedokončila. Svému příteli za odbornou pomoc při programování počítačového adaptivního testu. Svým kamarádkám za psychickou podporu v těžkých chvílích a odborné rady a názory.

Děkuji také všem respondentům, kteří se výzkumu zúčastnili.

Prohlašuji, že jsem diplomovou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

V Praze dne 17.11.2013 ……

Jana Dlouhá

**Abstrakt**

Teoretická část práce je zaměřena na **počítačové adaptivní testování (CAT) a teorii odpovědi na položku (IRT)**. Zařazena je také kapitola **srovnávající IRT s běžně používanou klasickou testovou teorií (CTT)**. Je zde také krátká zmínka o **počítačovém a online testování**, protože tyto typy administrace se v mnohém liší od klasických tužka-a-papír testů.

Cílem této práce bylo **zhodnotit jednotlivé způsoby administrace eEOD testu a porovnat je s testy eEPQ a sebehodnocením**. V praktické části byly **kalibrovány položky** škály extraverze Eysenckova osobnostního dotazníku (eEOD) na skupině 124 respondentů Na získaných datech byla následně provedena **simulace počítačového adaptivního testování**, která jasně ukázala **výhody tohoto typu testování** v porovnání s klasickou podobou testu.

Tyto **výsledky byly porovnány s výsledky reálné administrace** CAT testu na původním vzorku a na nové skupině respondentů ($N_p$=69, $N_n$=68). Výsledky silně korelovaly s výsledky simulovaného testu. Pro **ověření validity** počítačové adaptivní verze eEOD byly navíc výsledky respondentů v tomto testu **porovnány s výsledky v testu eEPQ** a v **krátké sebeposuzovací škále**. Nakonec bylo provedeno srovnání výsledků počítačové adaptivní administrace s **výsledky administrace tužka-a-papír.**

**Klíčová slova:**

Počítačové adaptivní testování, online testování, teorie odpovědi na položku, psychodiagnostika, Eysenckův osobnostní dotazník

**Abstract**

The theoretical part of the paper focuses on **computerized adaptive testing (CAT) and item response theory (IRT)**. Also included is a chapter **comparing IRT with the commonly used classical test theory (CTT)**. There is also a brief mention of **computerized and online testing**, as these types of administration differ in many aspects from conventional paper & pencil tests.

The goal of this paper was to **evaluate the individual ways of eEPI test administration and to compare them with eEPQ tests and self-evaluation**. In the practical part the items of the extraversion scale of the Eysenck Personality Inventory (eEPI) were **calibrated** using a group of 124 respondents. The acquired data were subsequently used to carry out a **simulation of computerized adaptive testing**, which clearly demonstrated the **benefits of this type of testing** in comparison to the classical test form.

**These results were compared with the results of real** CAT test administration using the original sample and a new group of respondents (Np=69, Nn=68). The results were highly correlated with the results of the simulated test. Moreover, to **verify the validity** of the computerized adaptive version of eEOD, the respondents' results in this test **were compared with the results in the eEPQ test** and in a **short self-assessment scale**. Finally, comparison of the results of computerized adaptive administration and **the results of paper & pencil administration** was performed.

**Keywords:**

Computerized Adaptive Testing, Online Testing, Item Response Theory, Psychodiagnostic, Eysenck Personality Inventory

# Obsah

## Seznam použitých zkratek

| | |
|---|---|
| IRT | Item response theory - teorie odpovědi na položku |
| CTT | Classical test theory - klasická testová teorie |
| CAT | Computerized adaptive testing - počítačové adaptivní testování |
| DIF | Differential item functioning - odlišné fungování položek |
| EOD | Eysenckův osobnostní dotazník |
| EPQ | Eysenck personality questionnaire |
| eEOD | Škála extraverze Eysenckova osobnostního dotazníku (forma a + b) |
| eEPQ | Škála extraverze Eysenck personality questionnaire |

# Úvod

Snad každý psycholog je v současnosti obeznámen v oblasti psychometrie a psychodiagnostiky s tzv. klasickou testovou teorií (CTT). Ve své práci bych se ale chtěla zabývat zcela jinou teorií, a to **teorií odpovědi na položku (IRT), která se od CTT výrazně liší**

Vždy bylo třeba volit mezi skupinovým a individuálním testováním. **Individuální testování je výhodné pro své dobré přizpůsobení se** testovanému jedinci. Můžeme si být jisti, že respondent rozuměl položkám, předcházet některým příčinám zkreslení výsledků apod. Oproti tomu je **skupinové testování méně nákladné**, a je možné otestovat i velký počet respondentů.

**Velký přínos ve skupinovém testování znamenal příchod počítačů**, které nejdříve umožnily zpracování odpovědí z papírových testů počítačem a následně i možnost provedení samotné administrace přímo na počítači. **Zvláštním typem počítačových testů jsou počítačové adaptivní testy**, které se přizpůsobují respondentovi podobně, jako to dělá zkoušející při individuálním testování. Tyto testy jsou založené právě na zmíněné **teorii odpovědi na položku**.

IRT má oproti CTT mnoho výhod, které budou v práci zvažovány. Lze si klást otázku **proč je teorie odpovědi na položku hlavně u nás stále mnohem méně využívaná než klasická testová teorie**. Podle Jelínka a kol. (2011a) může mít vliv hlavně **vysoká matematická obtížnost** této teorie. S tím související nutnost používat většinou **uživatelsky nepřívětivý software**, který vyžaduje nadstandardní počítačové znalosti uživatele a je dostupný prakticky pouze v angličtině. Tyto nedostatky jsou ale dostatečně kompenzovány zmíněnými výhodami IRT (McKay, 2008).

Svou roli zde může hrát i jistá **konzervativnost** - vědci jsou trénovaní v klasické teorii, jsou zvyklí používat její metody a jsou zběhlí v interpretování jejích výsledků (Reeve & Fayers, 2005). Vytvoření norem u CTT testu je časově i technicky méně náročné, než kalibrace CAT testu (Halama, 2005). Omezením je také závislost adaptivních testů na dostupnosti počítačů, a to hlavně při skupinovém testování. Tento problém se ale postupně zmenšuje. Za poměrně přijatelnou cenu je dnes možné získat výkonné a **poměrně malé**

**počítače, notebooky, ale i tablety, které jsou snadno přenosné**. Přesto jde stále o větší investici, než u papírových testů.

**Počítačové adaptivní testování**, které je **hlavní aplikací IRT**, se ve světě používá již několik desítek let a je velmi rozšířené. Je již k dispozici řada výzkumů, zabývajících se jeho efektivitou, možnostmi využití i různými způsoby vylepšení a inovací. V roce 2011 byl založen časopis ***Journal of computerized adaptive testing*** JCAT, což je recenzovaný elektronický časopis zaměřený na počítačové adaptivní testování (The International Association for Computerized and Adaptive Testing (IACAT), 2010). Je oficiálním časopisem ***International Association for Computerized Adaptive Testing*** a jeho šéfredaktorem je David **J. Weiss** z *University of Minnesota*, který se počítačovému adaptivnímu testování dlouhodobě věnuje. Dosud vyšla jen 4 čísla, poslední v září 2013.

# I. Teoretická část

# 1 Teorie odpovědi na položku

Kapitola o teorii odpovědi je zde hlavně z důvodu, že je tato teorie **naprosto nezbytná pro počítačové adaptivní testování** - tvoří jeho statistický základ. V tomto případě **nemůže být nahrazena klasickou testovou teorií**. Nemám ale v úmyslu popsat základy IRT vyčerpávajícím způsobem, to je úkol mnohem rozsáhlejších monografií, chci zde jen stručně zmínit některé důležité body této teorie. Čtenáře, který by měl zájem o podrobnější výklad, odkazuji například na publikace Baker (2001), Embretson and Reise (2000), Jelínek, Květon a Vobořil (2011b).

Přestože je teorie odpovědi na položku **mnohem mladší než klasická testová teorie**, je ve světě již velmi rozšířená a oblíbená, což dokládá i velké množství studií na toto téma. U nás je prozatím známá jen velmi málo. První monografie u nás, která se zabývá touto problematikou, vznikla až v roce 2011. V tomto roce byl v ČR k dispozici jediný CAT test - Woodcock-Johnsonovy testy kognitivních schopností (Urbánek, Denglerová, & Širůček, 2011).

Podle Hendla (2009) můžeme rozdělit teorie testů na dvě skupiny - klasické a pravděpodobnostní.

- **Klasická teorie testování** se u nás běžně používá. Výsledkem testování je bezprostřední projev rysu ovlivněný náhodnou chybou. Hrubé skóre je většinou vypočítáno jako součet bodů získaných za jednotlivé položky (Baker, 2001).

- Mezi pravděpodobnostní modely pak patří právě **teorie odpovědi na položku IRT**, kde je důležitá pravděpodobnost, se kterou respondent s určitou úrovní rysu odpoví na určitou položku správně (Hendl, 2009). Díky znalosti této pravděpodobnosti dokážeme předvídat, jestli respondent odpoví na položku správně, nebo špatně (Jelínek, Květon, & Vobořil, 2011b). Je zde tedy důležitější **výsledek u každé položky** než celkový výsledek testu (Baker, 2001).

IRT je jinak známá také jako „latent trait theory“ neboli teorie latentního rysu (Embretson & Reise, 2000).

## 1.1 Historie IRT

Zatímco klasická teorie testů vznikla přibližně v roce 1917 (Jelínek, Květon, & Vobořil, 2011a), kdy vznikly testy Army Alpha a Army Beta, za vznik IRT bývá považováno až vydání knihy **George Rasche** - *Probabilistic models for some inteligence and attainment test* v roce 1960, ve které popsal modely dnes zahrnované právě pod IRT (Jelínek, Květoň, & Denglerová, 2006). Vzniku IRT předcházelo vyvození ***logistic function*** (logistická funkce) v roce 1844 (Baker, 2001). Ta byla nejdříve používána hlavně v přírodních vědách pro modelování růstu rostlin a živočichů. Logistická funkce později poskytla základ právě Raschovým modelům. Raschovy modely se pak staly základem, na němž R. A. Fisher vytvořil způsob odhadu schopnosti respondenta pomocí ***maximum likelihood method*** (metody maximální věrohodnosti), která se používala dříve v toxikologii a biologii (Bock, 1997).

První známkou IRT už mohla být i **Thurstonova analýza Binetova a Simonova testu** inteligence v roce 1925, kde pro každou položku spočítal procento úspěšných dětí a graficky znázornil vztah věku a úspěšnosti. Položky pak uspořádal na věkovou stupnici, která již připomíná položkovou funkci v IRT (Bock, 1997).

V USA je za začátek IRT považováno až vydání knihy **Lorda a Novicka** - *Statistical theories of mental test scores* z roku 1968 (tedy až 8 let po vydání Raschovy knihy) (Jelínek, Květon, & Vobořil, 2011a; Embretson & Reise, 2000). Kniha obsahuje 4 kapitoly o IRT napsané Allanem Birnbaumem (Embretson & Reise, 2000). O další vývoj se pak postarali například Fumiko Samejim, David Thissen, Darrel Bock, Robert J. Mislevy, Wim J. Van der Linden nebo Cees A. W. Glas (Jelínek, Květon, & Vobořil, 2011a).

## 1.2 Předpoklady IRT

Pro jednodušší IRT modely platí jisté předpoklady (Reeve & Fayers, 2005;(Amarnani, 2009; Jelínek, Květon, & Vobořil, 2011b), které **musí být splněny, aby mohl být model spolehlivě použit**:

- Unidimenzionalita (nebo alespoň přítomnost jednoho dominantního faktoru)
- Lokální nezávislost
- Shoda modelu s daty (model fit)

**Předpoklad unidimenzionality** znamená, že pravděpodobnost odpovědi na položku je ze strany respondenta ovlivněna jen jednou z jeho charakteristik, kterou test měří. Ve skutečnosti ale takto přísnou podmínku prakticky nelze splnit, a tím spíš ne v psychologii (Jelínek, Květon, & Vobořil, 2011b). Vzhledem k tomu, že IRT modely jsou poměrně robustní vůči drobným odchylkám, stačí většinou zajistit, aby byl v testu přítomen pouze **jeden dominantní faktor**, a test je považován za unidimenzionální. Pokud není možné ani přes tento ústupek předpoklad unidimenzionality splnit, je vhodné zvážit **použití multidimenzionálních modelů**. I ty však vyžadují přiměřený počet dominantních faktorů. Unidimenzionalita bývá testována faktorovou analýzou (Kingsbury & Wise, 2000).

Pokud je test unidimenzionální, lze předpokládat i to, že je **lokálně nezávislý** (Jelínek, Květon, & Vobořil, 2011a). Porušení předpokladu lokální závislosti může nastat například ve chvíli, kdy **předchozí položky ovlivňují obtížnost řešení těch dalších**. I zde existují speciální modely, které shlukují položky, u kterých byla nalezena lokální závislost a každý tento shluk je pak považován za samostatnou položku.

Posledním předpokladem je **předpoklad shody modelu s daty**, který je obvykle posuzován pomocí chí kvadrát testu (Jelínek, Květon, & Vobořil, 2011a). Tato shoda je **velmi závislá na velikosti vzorku** respondentů - u malého vzorku se často objevuje falešná negativita, u větších vzorků zase falešná pozitivita. Vhodná velikost vzorku ve vztahu ke shodě modelu s daty byla zkoumána ve studii (Hula, Fergadiotis, & Martin, 2012), kde byly ze skutečných dat vytvořeny vzorky pro simulaci používající 1PL, 2PL a 3PL modely. Například pro odhad parametrů a lepší výsledky shody modelu s daty byl pro 1PL model vhodnější menší vzorek. Obecně je **nejvhodnější model, kde co největší počet položek vykazuje co nejnižší hodnoty chí** (Reise, Widaman, & Pugh, 1993).

## 1.3 Modely

V IRT je měření založené na modelech, ve kterých úroveň odhadu rysu záleží na odpovědích respondenta a na vlastnostech položek, které jsou mu administrovány (Embretson & Reise, 2000). Podle Jelínka, Květona a Vobořila (2011b) dělíme modely podle posuzovaných dimenzí na **unidimenzionální** a **multidimenzionální**. V unidimenzionálních je odpověď na každou položku ovlivněna jedním obecným faktorem a platí pro ně výše uvedená pravidla a omezení. Multidimenzionální modely jsou určeny pro testy, ve kterých nelze splnit požadavek unidimenzionality, například u EOD nebo Big five dotazníků.

### 1.3.1 Unidimenzionální

Dělíme na **dichotomní**, pro položky skórované pomocí dvou kategorií (např. ano/ne, 1/0 apod.) a **polytomní**, pro položky skórované pomocí více než dvou kategorií (např. Likertova škála).

#### 1.3.1.1 Dichotomní

Rozlišujeme 4 (většinou jsou uváděny 3) unidimenzionální dichotomní logistické modely podle počtu parametrů (Wainer & Dorans, 2000).

**1PL** (one parameter logistic model) - zahrnuje pouze **parametr obtížnosti b**. Ten dosahuje většinou hodnoty mezi -3 a 3, ale může nabývat hodnot od minus nekonečna po nekonečno. Čím je větší, tím je úloha obtížnější. Pokud je obtížnost položky 1, má respondent s úrovní rysu 1 právě 50% šanci zodpovědět položku správně. Amarnani (2009) podotýká, že tento model není vhodný, pokud nejsou skóry jednotlivých položek stejně korelovány s celkovým skóre v testu.

$$P(\theta) = \frac{1}{1 + e^{-(\theta - b)}}$$

**2PL** - k parametru obtížnosti přidává ještě **parametr citlivosti (diskriminační) a**. Ten dosahuje nejčastěji hodnot 0 až 2,8. Opět může dosahovat hodnot od minus nekonečna po nekonečno. Dobrá je hodnota nad 0,382, ideálně vyšší než 1.

$$P(\theta) = \frac{1}{1 + e^{-a(\theta - b)}}$$

**3PL** - přidává další parametr, **parametr (pseudo)uhádnutelnosti c**. Ten určuje pravděpodobnost uhádnutí správné odpovědi. Hodnota je většinou mezi 0 a 0,35. Teoreticky může dosáhnout hodnot 0 až 1.

$$P(\theta) = c + \frac{1 - c}{1 + e^{-(\theta - b)}}$$

**4PL** - přidává ne vždy zmiňovaný **parametr ledabylosti d** (Jelínek, Květon, & Vobořil, 2011b). Hodnotí možnost zodpovězení velmi jednoduché položky špatně u respondenta s vysokým theta, například z důvodu neopatrnosti nebo nepozornosti. Snaží se tak omezit ovlivnění výsledku. Zejména v adaptivních testech může mít totiž tato nepozornost

mnohem větší dopad, než v testech klasických, hlavně proto že bývají adaptivní testy kratší a každá položka zde má mnohem větší význam. Výsledky experimentu ukazují, že tento model nejen snižuje problematické podhodnocení respondentů, kteří udělali chybu z nepozornosti, ale ještě zlepšuje přesnost měření (Yen, Ho, Laio, Chen, & Kuo, 2012).

$$P(\theta) = c + (d - c)\frac{1}{1 + e^{-Da(\theta - b)}}$$

Rovnice 1 - 4PL model (Jelínek, Květon, & Vobořil, 2011b)

#### 1.3.1.2 Polytomní

Modely většinou vyžadují ordinálně uspořádané položky, některé dokonce umí pracovat i s položkami s nominální povahou (Reeve & Fayers, 2005).

**Graded response model GRM** (model odstupňovaných odpovědí) je vhodný pro položky s uspořádanými odpověďmi, jeho autorkou je F. Samejima (Embretson & Reise, 2000). Jde o úpravu 2PL modelu. Základem tohoto modelu je odhad parametru obtížnosti pro každou z možných odpovědí minus jedna (parametr b tvoří jakýsi práh). Odpovědi jsou tak rozděleny do určitého počtu dichotomních položek. Pak lze podle 2PL modelu vypočítat pravděpodobnosti správné odpovědi pro každou možnost. Podle studie De Ayala a kol. (1992), je však lepších výsledků dosahováno pomocí partial credit modelu.

$$P(\theta) = \frac{1}{1 + e^{-Da(\theta - b_{ij})}}$$

**Modified graded response model MGMR** (modifikovaný model odstupňovaných odpovědí), jehož autorem je E. Muraki, je založen na odhadu parametru b pro každou položku, místo odhadování prahů jako u klasického modelu odstupňovaných odpovědí.

$$P(\theta) = \frac{1}{1 + e^{-Da(\theta - (b_i - b_j))}}$$

**Partial credit model PCM** (model pro odstupňovaný kredit) rozšiřuje 1PL model (Reeve & Fayers, 2005). Je vhodný například v případě testování kognitivních schopností, osobnosti nebo hodnocení postojů (názorů) (Koch & Dodd, 1989).

$$\frac{P_{iu}(\theta)}{P_{iu-1}(\theta) + P_{iu}(\theta)} = \frac{1}{1 + e^{-(\theta - d)}}$$

**Nominal cathegory model NCM** (model pro nominální kategorie) se kromě klasického použití v testech dá použít třeba i pro analýzu odpovědí v některých testech, jako například ve studii Květona, Jelínka, Vobořila a Klimusové (2012). Analýza je prováděna na testu prostorové představivosti, který je součástí testu studijních předpokladů na Masarykově univerzitě v Brně. Díky tomuto modelu lze například zjistit, že respondenti s určitou úrovní testované schopnosti si u analyzovaných položek volí určité distraktory mnohem častěji.

$$P(\theta) = \frac{e^{c+a\theta}}{\sum e^{c+a\theta}}$$

Dalším polytomním modelem je také **rating scale model RSM** (model pro posuzovací škály) apod. (Reeve & Fayers, 2005).

Polytomní modely jsou vhodné například pro známý test NEO-PI-R, kde byla jejich použitelnost již několikrát testována například ve studii Jelínka a kol. (2011a) nebo studii Makaransky a kol. (2013), které ukázaly, že polytomní modely jsou poměrně efektivní a adaptivní administrace testu může přinést jeho výrazné zkrácení při žádném nebo minimálním snížení přesnosti. Je zde však třeba navíc použít multidimenzionální metody.

#### 1.3.1.3 Multidimenzionální

Multidimenzionální modely (MIRT) jsou vhodné pro testy, u kterých nelze dosáhnout požadovaného předpokladu unidimenzionality, ani předpokladu jednoho dominantního faktoru (Urbánek, Denglerová, & Širůček, 2011). Můžeme je dělit na MIRT modely pro položky se dvěma a modely pro položky s více skórovacími kategoriemi (Reckase, 2009).

Patří sem například rozšířený **multidimenzionální 2PL model** nebo **multidimenzionální 3PL model**.

## 1.4 Odhad parametrů

### 1.4.1 Odhad parametrů položek

Jinak také nazývaný **kalibrace položek**. Částečné se podobá analýze položek v klasické teorii testů (Urbánek, Denglerová, & Širůček, 2011).

### 1.4.2 Odhad úrovně rysu (theta)

Pro odhad úrovně rysu (theta) se obvykle používá několik metod (Jelínek, Květoň, & Vobořil, 2011a): **maximum likelihood estimation MLE** (metoda maximální věrohodnosti) a Bayesovské metody odhadu maximum a **posteriori MAP** a **expected a posteriori EAP**. Metodu maximální věrohodnosti nelze použít v případě, že všechny položky byly zodpovězeny správně, nebo naopak špatně. Doporučuje se proto v případě adaptivního testu použít u prvních položek pro odhad Bayesovské metody a dále metodu maximální věrohodnosti.

$$P = \prod P(\theta)Q(\theta)$$

Rovnice - maximum likelihood estimation method (Wainer & Dorans, 2000)

Ke zjištění očekávané úrovně schopnosti **není potřeba srovnání s normami, jako u klasické teorie** (Amarnani, 2009). Při skórování 10 položkového testu dokáže CTT podle Weisse (2004) vyprodukovat obvykle 11 skórů, zatímco metoda maximální věrohodnosti dokáže z toho samého testu vyprodukovat až $2^{10}$ (tedy asi 1 024) různých odhadů theta.

## 1.5 Chyba odhadu a informační přínos

Protože v IRT jsou parametry položek a úrovně schopnosti **pouze odhady**, je dobré doplnit je informací o jejich chybě, případně informačním přínosu položek a testu (Baker, 2001).

### 1.5.1 Informační přínos položky a testu

Pokud je informační přínos položky nebo testu příliš malý, znamená to, že úroveň schopnosti nemůže být odhadnuta s dostatečnou přesností (Baker, 2001). Hodnota **informačního přínosu** je obvykle uváděna jako hodnota mezi 0 a 10, ačkoli může dosahovat

i vyšších hodnot. **Informační funkce položky** znázorňuje přesnost odhadu na každé úrovni schopnosti. Nejvyšších hodnot informačního přínosu dosahuje obvykle položka v bodě, kde její obtížnost odpovídá úrovni rysu respondenta a pravděpodobnost jejího správného zodpovězení je tedy zhruba poloviční (to může být částečně ovlivněno použitím dalších parametrů). U extrémních hodnot theta bývá naopak informační přínos velmi malý.

$$I_i(\theta) = P_i(\theta)Q_i(\theta)$$

**Rovnice - informační přínos položky pro 1PL model**

$$I_i(\theta) = a^2 P_i(\theta)Q_i(\theta)$$

**Rovnice - informační přínos položky pro 2PL model**

$$I_i(\theta) = a^2 \left[\frac{Q(\theta)}{P(\theta)}\right]\left[\frac{P(\theta) - c^2}{(1 - c^2)}\right]$$

**Rovnice - informační přínos položky pro 3PL model**

**Celkový informační přínos testu** je vypočítán jako suma informačních přínosů použitých položek pro danou úroveň rysu (Baker, 2001). **Informační funkce testu** nám pak ukazuje, jak přesně měří náš test na každé úrovni schopnosti.

$$I(\theta) = \sum I_i(\theta)$$

**Rovnice - informační přínos testu**

Pokud by byla v testu zahrnuta pouze jedna položka, byl by jeho celkový informační přínos velmi malý a odhad úrovně schopnosti velmi nepřesný (Baker, 2001). Teoreticky by se mohlo zdát, že čím více položek je v testu zahrnuto, tím lepší je jeho informační přínos. To ale nemusí být pravda v případě, že budeme položky opatrně volit tak, aby **jejich obtížnost co nejlépe odpovídala úrovni schopnosti respondenta**, čímž dosáhneme vyšších informačních přínosů každé položky a tedy i **vyššího celkového informačního přínosu testu**

**i při použití menšího počtu položek**. Na tomto principu si lze zjednodušeně představit fungování počítačových adaptivních testů.

### 1.5.2 Standardní chyba odhadu parametrů položek a úrovně schopnosti

Standardní chyba je jednoduše vypočtena pomocí následujícího vzorce (Baker, 2001):

$$SE(\theta) = \frac{1}{\sqrt{I(\theta)}}$$

Rovnice - Standardní chyba odhadu theta

Pokud je tedy zjištěná úroveň informačního přínosu 2,383 na úrovni schopnosti 0, je standardní chyba rovna 0,65 (Baker, 2001). Z toho vyplývá, že přibližně 68% odhadů této schopnosti bude mezi -0,65 a 0,65.

### 1.5.3 Informační přínos, standardní chyba a reliabilita

Standardní chyba v IRT je **ekvivalentem k reliabilitě v CTT** (Wainer & Dorans, 2000). Reliabilita v CTT je fixní hodnotou pro daný test nezávislá na respondentovi a závislá na vzorku, na kterém byl test standardizován. S**tandardní chyba je ale odvozena pro každou úroveň schopnosti, protože na každé úrovni schopnosti má test jinou přesnost**. Standardní chybu testu si tedy lze představit jako funkci. Standardní chybu v IRT však lze převést na klasickou reliabilitu podle vzorce:

$$\rho = 1 - SE^2$$

Rovnice - výpočet reliability ze standardní chyby

Pokud informační přínos testu dosáhne hodnoty 10, standardní chyba je podle předešlých vzorců přibližně 0,3 a reliabilita dosahuje hodnoty 0,9.

# 2 Porovnání teorie odpovědi na položku (IRT) s klasickou testovou teorií (CTT)

Teorie odpovědi na položku je **výpočetně mnohem složitější, než klasická testová teorie** (Amarnani, 2009). Měla by ale poskytovat **přesnější informace** o úrovni schopnosti jedince a dalších faktorech. Nejlepších výsledků je dosaženo v případě, že je zvolen vhodný model, u kterého je prokázána dostatečná shoda s daty a v případě unidimenzionálních modelů jsou dodrženy další požadované předpoklady.

V klasické testové teorii je význam jednotlivé položky mnohem menší, než v teorii odpovědi na položku (Amarnani, 2009). Položky v CTT tvoří jednotný celek a nejsou oddělitelné od testu. Naopak v **IRT má každá položka své vlastní parametry a není na testu závislá**. Položky v IRT lze bez zásadního vlivu na celý test přidávat, odebírat nebo upravovat. Velkou výhodou IRT je také **dostupnost standardních chyb** parametrů položek i odhadů úrovně schopnosti (Schmettow & Vietze, 2008).

Parametry položek v IRT mají jednu zajímavou vlastnost a tou je jejich nezávislost na vzorku (Baker, 2001). **Nezávislost parametrů položek na kalibračním vzorku** znamená, že pokud budeme mít například dva odlišné vzorky respondentů s odlišnými úrovněmi měřené schopnosti, parametry položek vyjdou přibližně stejné. Díky této vlastnosti **není pro vývoj testu a kalibraci položek nutné získat reprezentativní vzorek**, jako je tomu v CTT.

V CTT jsou důležité normy, s jejichž srovnáním získávají testové skóry svůj význam (Embretson & Reise, 2000). V **IRT jsou testové skóry porovnávány vzhledem k jednotlivým položkám**, normy nejsou potřeba. Pro pretestování metody není v IRT potřeba sbírat data od skutečných respondentů, jak je tomu v CTT, protože je možné tato data simulovat (Mills & Stocking, 1996). I tak je samozřejmě nezbytná následná reálná studie.

Kvůli matematické obtížnosti IRT je nezbytné mít dostupnou **výpočetní techniku** a samozřejmě také vhodný **software**. To lze chápat jako jednu z nevýhod IRT. Software pro IRT je většinou velmi drahý, ale existují i různě kvalitní freeware programy. Je možné získat rozšíření jazyka R pro IRT a CAT (catR) (Magis & Raiche, 2011). Dalšími možnostmi je software společnosti SSI, která poskytuje programy pro dichotomní (BILOG, BILOG-MG, IRTPRO) i polytomní IRT modely (MULTILOG, PARSCALE). Program IRTPRO je

podobný známému Microsoft Excel a je poměrně uživatelsky přívětivý a jednoduchý. K dispozici jsou také simulační programy, např. WinGen autora Chrise Hana.

## 2.1 Studie srovnávající CTT a IRT

Porovnáváním CTT a IRT přístupů se zabývá mnoho studií. Z nich bych chtěla uvést například studii slovenských autorů Halamy a Matúse (2006), kteří prováděli pomocí obou přístupů psychometrickou analýzu Rosenbergovy škály sebehodnocení. Použili archivní data z více výzkumů, kde většinu respondentů tvořili studenti středních a vysokých škol na Slovensku. V**ýsledky CTT a IRT jsou si zde v určitých ohledech podobné.** Složitější IRT přístup zde nemá takový přínos, jaký by se od něj dalo očekávat. Podobné srovnání proběhlo také v roce 2011. Autoři se zajímali o citlivost obou přístupů při měření psychoterapeutické změny. **IRT přístup se ukázal jako citlivější**, protože identifikoval více osob se signifikantní změnou než CTT (Halama & Biescad, 2011).

K názoru, že jsou výsledky obou přístupů podobné, došlo i několik zahraničních studií. Například podle studie Fan (1998) **produkují obě teorie velmi podobné výsledky** jak v parametrech položek, tak v parametrech osob. Autor ale dodává, že **IRT má potenciál k rozvoji**, zatímco CTT už ne. Limitem, omezujícím využití plného potenciálu IRT, zde může být také použitá položková banka. Vyšší citlivost IRT přístupu, v tomto případě při hodnocení školní efektivnosti, se prokázala ve Fox (2004) studii. Autor dochází k závěru, že **IRT přístup je výhodnější**, a to hlavně proto, že chyba měření může být hodnocena nezávisle na vzorku.

Z hlediska srovnání CTT a IRT je velmi zajímavá také **série studií Christiny Stage z Umeå university ve Švédsku**, které byly provedeny v rozmezí let 1996 až 2003. Všechny studie z této série jsou zaměřeny na **porovnávání přístupů při jejich aplikaci na test Swedish scholastic aptitude test (SweSAT)** (Stage, 2003). SweSAT má několik části, na které se zaměřují jednotlivé skupiny studií (Stage, 1996): Diagrams, maps and tables - DTM (Stage, 1997a), English reading comprehension - ERC (Stage, 1997b; Stage, 1998b), Swedish reading comprehension - READ (Stage, 1997c; Stage, 1999a) a Vocabulary - WORD (Stage, 1997d; Stage, 1998a). IRT přístup se rozhodli vyzkoušet hlavně z důvodu nedostatků CTT, jako je závislost obtížnosti položky a diskriminačních schopností položky na použitém vzorku respondentů nebo závislost skórů na náročnosti testu. Problém byl i

s odhadem schopnosti v CTT, který byl horší pro respondenty s vysokou nebo naopak nízkou úrovní schopnosti (Stage, 1996). Jako nejvhodnější se pro všechny subtesty ukázal 3PL model, naopak 1PL model se pro SweSAT nehodí (Stage, 2003). V**ýsledky obou srovnávaných teorií se i v tomto případě téměř shodovaly** (Stage, 1999b). V žádné z výše zmíněných studií se **nepodařilo prokázat významnější přínos IRT pro SweSAT test**. Tuto skutečnost autorka připisuje i velikosti vzorku - **pokud by byl vzorek menší, CTT přístup by měl pravděpodobně omezenější možnosti a IRT přístup by se tak ukázal výhodnější**. Záměna dosud používaného CTT přístupu za IRT přístup tedy zřejmě nebude pro tento test nijak významně přínosná, a to alespoň do doby, než se Umeå university rozhodne pro přepracování testu do počítačové adaptivní formy, u které nelze CTT přístup použít (Stage, 2003).

Jako jiné studie, i Amarnani (2009) dochází k závěru, že **ideální je zkombinovat výhody obou přístupů**. Ačkoli ve zmíněných studiích nebyly prokázány jasné výhody IRT přístupu v porovnání s klasickým přístupem, což mohlo být způsobeno prací s velkými vzorky nebo nedostatečným ověřením vhodnosti použitého modelu, bude pravděpodobně **IRT používána a rozvíjena i nadále, a to hlavně v oblasti počítačového adaptivního testování (CAT), kde je nenahraditelná**.

# 3 Počítačové a online testování

Podle Embretson (1992) dělá hlavní změny v psychologickém testování nově dostupná počítačová technologie. **Počítačové testy jsou sestrojovány tak, aby jejich výsledky byly srovnatelné s testy klasickými.** Zdá se ale, že ve skutečnosti zobrazuje počítač položky testu trochu jiným způsobem, než při klasické administraci, a že má tato odlišnost své důsledky pro srovnatelnost výsledků. Žitný (2011) rozlišuje tři úrovně počítačových testů v psychologické diagnostice:

**CAPA - computer-assisted psychological assessment**, což znamená jakoukoli diagnostiku využívající počítač (např. Katz & Dalby, 1981).

**CBTI - computer-based test interpretation**, zahrnuje diagnostiku využívající počítač i k interpretaci získaných údajů (např. Moreland, 1985).

**CAT - computerized adaptive tests**, což jsou již známé počítačové adaptivní testy, kde se daný test přizpůsobuje respondentovi na základě určitých algoritmů.

## 3.1 Ekvivalence počítačových a klasických testů

Srovnáváním klasických *paper&pencil* (tužka-a-papír) testů s testy počítačovými se zabýval například Chen a kol. (2011). Autoři srovnávali tři úlohy, jejichž **analýza následně odhalila rozdíly ve výkonu, mezi respondenty, kteří dostali klasickou formu testu a těmi, kterým byly úlohy administrovány počítačově**. Obecně měli respondenti, kterým byla administrována klasická forma, lepší výsledky. Podle autorů může počítačové testování zvýhodnit některé respondenty vzhledem k jejich počítačovým schopnostem. Podle mého názoru se ale počítačová gramotnost rapidně zlepšuje a práce s počítačem je pro většinu lidí již samozřejmostí, takže význam tohoto zvýhodňování bude postupně klesat. Navíc tento vliv bude pravděpodobně **patrný pouze u výkonových testů**.

Podle Žitného (2011), **není možné automaticky považovat papírovou a počítačovou verzi testu za ekvivalentní.** K rozhodnutí o ekvivalenci takových testů je třeba empirických studií, protože zobrazení položek na počítači a práce s osobním počítačem má prokazatelný vliv na výsledky testů. Tento vliv je největší u **testů výkonových**, zvláště pokud obsahují **grafické úkoly a mají časové omezení** (Květon & Klimusová, 2002). Například **u testu IST**

**byla prokázána neekvivalence** některých subtestů, které obsahovaly podněty v grafické formě (Květon, Martin, Vobořil, & Klimusová, 2003). V**ýkony respondentů při počítačové administraci byly horší**, než u klasické administrace. Autoři to přisuzují náročnějšímu ovládání a jinému způsobu zobrazování.

Neekvivalence byla odhalena také u **Bourdnova testu a Testu koncentrace pozornosti** (Květon, Jelínek, Vobořil, & Klimusová, 2007). V této studii se autoři navíc zabývají **vlivem změny barvy pozadí a písma** na výkon respondenta u Bourdnova testu. Byl zde prokázán signifikantní vliv na výkon testovaného. Při zobrazení tmavých znaků na světlém pozadí dosahovali respondenti lepších výsledků než při inverzním zobrazení. Vliv na rozdíly ve výkonech respondentů ve srovnání počítačové a klasické verze mohla mít ale i modifikace zobrazení testu. U počítačové administrace nebylo možné zobrazit celý test najednou, proto byl vždy zobrazen pouze jeden řádek testu rozdělený na několik kratších řádků, a po vypršení časového limitu byl zobrazen další řádek na nové obrazovce.

Naopak u testu **Ravenovy standardní progresivní matice** byla prokázána poměrně dobrá ekvivalence obou forem administrace (Williams & McCord, 2006). Tato studie neprokázala vyšší úroveň úzkosti při počítačové administraci, která bývá někdy uváděna.

## 3.2 Výhody

Respondenty je jako největší výhoda chápáno **okamžité vyhodnocení**, které jim počítač zobrazí bezprostředně po vyplnění testu nebo testové baterie (Parshall, 2002). Pro administrátora může být výhodou třeba to, že počítač umožňuje **sledovat úmyslně vynechané nebo nezobrazené položky**, měřit délku rozhodování u každé položky nebo zaznamenávat změny v odpovědích. Počítačové testy také znemožňují omylem zaškrtnout dvě odpovědi u položky, kde je možné odpovědět jen jednou možností. Výsledkem jsou čistší data, **prevence chyb** a tím pádem i lepší hodnotitelnost výsledků testů.

Hornke (2000) se zabýval právě zaznamenáváním **reakční doby** u každé odpovědí a tím, jaký má tato informace vlastně smysl. V jeho výzkumu podstoupilo 5 912 respondentů adaptivní test. Souvislost reakčního času a odhadu schopnosti se neprokázala. Přestože průměrná délka u špatných a správných odpovědí neukázala signifikantní rozdíl, u **špatných**

**odpovědí byla délka reakce o něco větší**. Je pravděpodobnější, že by odlišnosti v reakčním čase mohly souviset s osobnostními charakteristikami.

Podobný výzkum byl proveden také ve studii Chang a kol. (2011). Porovnávali **množství času, které respondenti na různých úrovních schopnosti stráví nad správně a špatně zodpovězenými položkami**. Zkoumali také vztah této reakční doby a hádání. Byly odhaleny **unikátní časové odpověďové modely** pro šest skupin úrovně schopnosti. Tyto modely mohou být použity například pro odhalování hádání, nebo zjišťování úrovně spolehlivosti testu.

Počítače také umožňují vytvářet **nové podoby položek** (Wainer & Dorans, 2000). Lze využít různá multimédia - zvuk, video, různý hardware - klávesnici a myš, páky, volanty, kreslící tabule, dotykové obrazovky, simulátory apod. Vzniká tak **inovativní a mnohem realističtější testovací prostředí**. Do budoucnosti lze navíc očekávat ještě **další rozvoj** počítačové techniky, který může přinést ještě zajímavější podoby položek a testů.

## 3.3 Slabé stránky

Je třeba zohlednit nutnost kontroly respondentů proti případnému podvádění, a to hlavně v případě online testů (Wainer & Dorans, 2000). Vzhledem k nutnosti naprogramovat potřebný software bude také pravděpodobně **cena takového testu mnohem vyšší**, než u testu klasického (Parshall, 2002). Problém nastává také při **zajišťování stejných podmínek pro všechny respondenty** tak, aby byly jednotlivé testy srovnatelné mezi sebou. Je třeba zohlednit například emoční reakce respondentů na testování počítačem a **úroveň počítačových znalostí. V**elké rozdíly mezi testy může způsobit také **efekt individuálního nastavení počítače**, na kterém je test administrován (velikost, barevnost, kontrast a rozlišení monitoru), a to zvláště v oblasti online testování a testů s časovým limitem (Wainer & Dorans, 2000).

Pro samotnou administraci je třeba vytvořit informační i položkové rozhraní tak, aby bylo jasné jak se po stránce pohybovat, jak přejít na další položku, změnit odpověď apod. (Parshall, 2002). S**právné testovací rozhraní musí být jednoduché, jasné a intuitivní** - často se používá termínu *user-friendly* (uživatelsky přívětivé). Rozhraní by mělo zahrnovat **jasné instrukce, tutoriály a procvičovací položky**, případně také možnost přivolat

administrátora nebo jinak získat pomoc. Může být nastavena i možnost zopakování instrukcí, pokud respondent nesprávně zodpoví několik velmi jednoduchých položek za sebou (Wainer & Dorans, 2000). Počítač může dokonce v některých případech určit, co dělá respondent špatně (nesprávné přečtení instrukce položky, záměna negativu) a respondenta cíleně navést k opravě.

V neposlední řadě je nezbytné **dostatečné zabezpečení získávaných dat**. To je opět větší problém hlavně u online testů, které jsou mnohem náchylnější k napadení hackery. U testu, převedeného z klasické podoby do počítačové, by měla být **znovu ověřena reliabilita a validita**.

# 4 Počítačové adaptivní testování

Počítačové adaptivní testování je **jednou z nejvýznamnějších aplikací IRT** (Jelínek, Květon, & Vobořil, 2011b). Podle Wainer a Dorans (2000) mnozí věří, že adaptivní testování je *raison d'etre* (důvod bytí) IRT. Základní myšlenkou je, že pro získání co největšího množství informace by bylo ideální, aby každý respondent dostal položky, u kterých má právě **poloviční pravděpodobnost správné odpovědi**. CAT se tedy snaží z položek dostupných v položkové bance vybírat právě tyto položky, a omezit množství těch, které jsou pro respondenta příliš jednoduché, nebo příliš obtížné. Pokud totiž respondent dostane příliš jednoduché položky, pravděpodobně je zodpoví všechny správně, pokud dostane příliš těžké, pravděpodobně zodpoví všechny špatně (Urbánek, Denglerová, & Širůček, 2011). Tento výsledek nám ale řekne velmi málo o jeho skutečné úrovni schopnosti.

Adaptivní test funguje tak, že pokud je položka na určité úrovni obtížnosti **zodpovězena špatně, úroveň se sníží, pokud je zodpovězena správně, úroveň se zvýší** (Wainer & Dorans, 2000). Položky jsou obvykle pokládány do doby, než je dosaženo určité úrovně chyby nebo přesnosti.

## 4.1 Historie adaptivních a počítačových adaptivních testů

Adaptivní testování je v podstatě tak staré, jako ústní testování (van der Linden & Glas, 2000). Adaptivní test má napodobovat úkol zkoušejícího, tedy vybírat vhodné položky pro individuálního respondenta (Wainer & Dorans, 2000). Samotná myšlenka adaptivních testů pochází již z **Binet-Simonova inteligenčního testu**, tedy někdy kolem roku 1905 (van der Linden & Glas, 2000). V tomto testu ale výběr položek prováděl sám zkoušející, na základě odhadovaného mentálního věku. V testu byly zvoleny položky pro každý rok věku podle toho, že přibližně polovina dětí v tomto věku zodpověděla položku správně (Weiss, 2004). Vznikly tak sety položek pro každý rok od 3 do 11 let. Tento soubor položek by mohl být považován za jakousi položkovou banku.

Bohužel byl Binet-Simonův test posledním adaptivním testem až do 50. let 20. Století, kdy **začaly vznikat jednoduché počítačové programy pro adaptivní testování** (Filípková & Byčkovský, 2008). Nejdříve vznikaly jen lineární programy, větvené rozhodovací struktury se začaly objevovat od 60. let a od 70. let, kdy už byly počítače schopné vytvořit sadu otázek

z různých úloh banky. Přizpůsobivost takových testů byla však stále značně omezená, protože šlo o předem dané větve a smyčky otázek. Počítače byly, kromě toho, také stále častěji používány pro skórování klasických testů (Wainer & Dorans, 2000).

V roce 1968 vyšlo dílo Lorda a Novicka *Theories of mental test scores*, které se zabývalo **teorií odpovědi na položku**. S vývojem IRT pak bylo najednou možné snadno hodnotit každou položku podle její náročnosti a každého respondenta podle jeho úrovně schopnosti. To přineslo možnost administrovat respondentovi jen některé položky z položkové banky, které nejlépe hodnotí jeho schopnost a vynechat ty položky, které jsou příliš jednoduché nebo obtížné.

V 80. letech se dočkal Binet-Simonův test své počítačové varianty *The stratified adaptive computerized ability test* (Filípková & Byčkovský, 2008). V této době už **běžně probíhají administrace testů na počítači** (Wainer & Dorans, 2000). Postupně jsou **zlepšovány adaptivní algoritmy**. Průkopníkem počítačových adaptivních testů bylo *US Department of Defense*, které vytvořilo **první vojenský prototyp počítačového adaptivního testu**, který byl původně určen pro počítače firmy Apple (van der Linden & Glas, 2000). Šlo o test *ASVAB - Armed Services Vocational Aptitude Battery*.

## 4.2 Výhody a nevýhody

Za hlavní výhodu CAT je považována jeho **efektivita**, důležité je ale i **zpřesnění hodnocení** respondentů s extrémními úrovněmi schopnosti. Kromě těchto hlavních výhod přináší i další vedlejší výhody jak pro respondenty, tak pro samotné tvůrce testů. Největší nevýhodou jsou naopak **vysoké prvotní náklady**. Respondenty je také často negativně hodnoceno omezení v přeskakování položek a oprav v testu.

### 4.2.1 Výhody

#### 4.2.1.1 Efektivita

Největší výhodou počítačových adaptivních testů je pravděpodobně jejich efektivita. Ta znamená hlavně **nutnost administrace menšího počtu položek k dosažení stanovené míry přesnosti** (Jelínek, Květon, & Vobořil, 2011b). Psychometrická efektivita se zde může

projevovat buď na snížení délky testu, často až na polovinu klasického testu, nebo zvýšením přesnosti měření při zachování původního počtu položek (Embretson, 1992).

Například v testu MMPI pro adolescenty byl poměr ušetřených položek mezi 10,7-26,4% (Forbey, Handel, & Ben-Porath, 2000). U testu MMPI-2 dokonce až mezi 20-35% (Forbey & Ben-Porath, 2007). Při pilotní studii celostátního matematického testu *Michigan Educational Assessment Program* byl test zkrácen přibližně o čtvrtinu (Schermis & Stemmer, 1996). Vos (2000) provedl porovnání u 3 variant testů: 10, 52 a 50 položek. Dosáhl zkrácení testů až o 76,4%, 57,9% a 34,8%. Efektivnějších a přesnějších výsledků v porovnání s klasickými testy dosáhli i Vispoel a kol. (Vispoel, Rocklin, & Tianyou, 1994). Takováto úspora času a položek je výrazná zvláště u **testování časově náročnými bateriemi testů**, kdy časová úspora dosahuje většinou 40-60% (Linden & Glas, 2010). Pokud je tedy třeba otestovat respondenty ve velmi omezeném čase, CAT je dobrou volbou (Parshall, 2002).

Z dalších osobnostních testů byla efektivita zkoumána například u **MPQ, NEO PI-R nebo EOD**. U metody MPQ bylo dosaženo **redukce až o polovinu položek při vysoké korelaci** r = 0,97 (Waller & Reise, 1989). Při poloviční redukci položek bylo vysoké korelace dosaženo i **u testu NEO PI-R, korelace dosahovala r = 0,92** (Reise & Henson, 2000). Při administraci škály neuroticismu EOD testu byla nastavena chyba měření SE = 0,5 (Květon, Jelínek, Denglerová, & Vobořil, 2008). Téměř polovina respondentů zodpověděla k dosažení této přesnosti **pouze polovinu položek**. Naopak u některých respondentů nebylo možné dosáhnout stanovené úrovně chyby ani při administraci všech položek, přičemž šlo většinou o **jedince s extrémní úrovní rysu**. Pro zlepšení odhadu schopnosti a efektivity testu by bylo vhodné vyvinout ještě položky měřící právě v oblasti obou extrémních pólů, které v dostupném EOD testu chybí.

U nás zkoumali efektivitu počítačových adaptivních testů Žitný a kol. (2012). Jedna skupina respondentů dostala **klasický papírový test, druhá skupina počítačový nebo počítačový adaptivní test**. U testu TIP (Test inteligenčního potenciálu) bylo dosaženo 55% úspory, u testu VMT (Vienna matrices test) 54% úspory. Výsledky CAT testu byly i při takovémto snížení počtu položek **srovnatelné s výsledky papírových testů**. Podle Žitného je efektivita testu velmi závislá na podmínkách pro jeho ukončení a jejich správném nastavení (Žitný, 2011).

#### 4.2.1.2 Vyšší přesnost v extrémních pólech

Konvenční testy většinou výrazně ztrácí na přesnosti v extrémních pólech distribuce skóru (Parshall, 2002). V**ětšina položek v klasických testech se totiž, kvůli normálnímu rozložení, pohybuje kolem průměru**, kde předpokládá nejvyšší podíl respondentů (Wainer & Dorans, 2000). Zařazení většího počtu obtížnějších a jednodušších položek by pravděpodobně neúměrně prodloužilo délku testu. Počítačové adaptivní testy ale ve svých položkových bankách mají dostatek položek pro každou úroveň rysu, a ty jsou použity jen, když je jich potřeba, takže prodlužování testu není problém. Přesnost testu záleží na vhodnosti rozložení otázek napříč různými úrovněmi schopnosti. Příliš mnoho velmi jednoduchých položek může respondenta nudit, unavovat a zvyšuje tak nebezpečí chyb z nepozornosti. Naopak příliš mnoho obtížných položek zase zvyšuje frustraci, snižuje motivaci respondenta a přináší nebezpečí hádání (Jelínek, Květoň, & Denglerová, 2006).

#### 4.2.1.3 Výhody pro respondenty

Jak již bylo řečeno, CAT test je obvykle tak akorát těžký, takže respondenta **nenudí ale ani nefrustruje** (Wainer & Dorans, 2000). Je tak zachována **dostatečná motivace respondenta**. Testy jsou poměrně **snadno dostupné**, u online testů je možné podstoupit je prakticky kdykoli a odkudkoli, kde je k dispozici potřebné počítačové vybavení a připojení k internetu. Velkou výhodou pro respondenty je také možnost **okamžitého zobrazení výsledků** po vyplnění testu. Samozřejmě lze používat multimediální položky jako u počítačových testů, což umožňuje vytvářet testy, které jsou pro respondenty zábavnější.

#### 4.2.1.4 Snadné úpravy a aktualizace testu

Počítačové adaptivní testy a testy založené na IRT obecně lze **snadno upravovat** (Jelínek, Květoň, & Denglerová, 2006). Položky v položkové bance mohou být přidávány, odebírány nebo nahrazovány bez ohrožení integrity testu. Parametry položek mohou být stále **zpřesňovány a aktualizovány**. U online testů je například možné přepočítat parametry položek po každém vyplnění testu. Test je tak možné bez zvyšování nákladů udržovat stále aktuální. U klasických testů jakékoli úpravy obvykle vyžadují novou standardizaci testu.

#### 4.2.1.5 Větší bezpečnost testu, možnost opakování testování

Další výhodou adaptivních testů, je jejich **možnost opakování bez nežádoucího efektu učení** i po krátké době (Jelínek, Květon, & Vobořil, 2011b). Čím větší je položková banka testu, tím méně nežádoucí efekt učení hrozí. Respondent totiž prakticky **podstupuje při opakování úplně jiný test**, Díky tomu, že je vždy **používána jen část položkové banky**, zvyšuje se její bezpečnost proti ukradení položek (Wainer & Dorans, 2000). Protože každý respondent podstupuje trochu odlišný test, s jiným pořadím položek, zamezuje se také opisování.

### 4.2.2 Nevýhody

Počítačové adaptivní testování má samozřejmě i své nevýhody. Jednou z nich jsou **vysoké prvotní náklady**, hlavně na pořízení software a hardware (Wainer & Dorans, 2000). Je také potřeba vyvinout mnohem větší počet dostatečně kvalitních položek pro položkovou banku, aby se mohly naplno projevit zmíněné výhody CAT. Tyto položky je vhodné dostatečně chránit proti odcizení. Samozřejmostí jsou také **vysoké nároky na teoretickou připravenost** provozovatele testu (Jelínek, Květon, & Vobořil, 2011b). Počítačové adaptivní testy se stále vyvíjejí, ale náklady už nejsou takové jako ty prvotní (Wainer & Dorans, 2000).

CAT testy mohou u klienta vyvolat **pocit nespravedlnosti**, protože může svůj set položek vnímat jako obtížnější nebo příliš krátký. Může mít také dojem, že výsledek v testu neodpovídá jeho procentuální úspěšnosti. Což je vlastně pravda, protože stejné procento správných odpovědí v testu může vést k odlišným odhadům schopnosti (Jelínek, Květon, & Vobořil, 2011b).

Jednou z velkých nevýhod počítačových adaptivních testů je také **nemožnost hodnocení otevřených odpovědí** (Eggen, 2004). I hodnocení polytomních modelů a multidimenzionálních testů může být poměrně komplikované. Kingsbury a Wise (2000) zmiňují také stížnosti respondentů, že **není možné během testu zjistit, v jaké fázi se právě nachází** a jak velká část testu ještě zbývá. Respondent tak neví, jak rozložit své úsilí. Podle mého názoru by možná stačilo - u testů, kde je ukončovacím pravidlem dosažení určité chyby měření - graficky znázornit momentálně dosaženou úroveň chyby v porovnání s tou

očekávanou. Sice by ani tak nebylo přesně určitelné kolik práce má ještě respondent před sebou, ale přesto by to pro něj mohlo být příjemnější.

#### 4.2.2.1 Problém přeskakování položek a jejich kontroly (item review)

Respondenti za jeden z největších nedostatků počítačových adaptivních testů zmiňují právě **nemožnost přeskakování položek a jejich zpětné kontroly a oprav**. Obecně totiž respondenti preferují testy, ve kterých mají co nejvíce kontroly a informací (Vispoel, Rocklin, & Tianyou, 1994). I v Schermiz a Stemmer studii (1996) vadilo respondentům hlavně to, že se nemohou vracet k předchozím položkám, jako u toho bylo u porovnávaného klasicky administrovaného testu.

Podle Kingsbury a Wise (2000) **opravy položek zvyšují délku testu a často i zvyšují konečnou standardní chybu**. Kromě toho jich respondenti mohou zneužít pro zvyšování svého skóre. Chápou nemožnost opravovat odpovědi jako výhodu proti klasickým testům, kde jde spíše o neplánovaný a nekontrolovatelný aspekt skupinové administrace. Nemožnost kontrol ale u respondentů zvyšuje úzkost.

Přesto existují **CAT testy, které tyto kontroly a opravy umožňují** (Olea, Revuelta, Ximénez, & Abad, 2000). U jednoho takového testu možnost kontrol využilo velké procento respondentů, kteří díky nim získávali vyšší úrovně schopnosti než respondenti z kontrolní skupiny. Podle autorů **je umožněním kontrol a oprav snižována efektivita** testu, která je obvykle považována za největší výhodu CAT. Při možnosti oprav stoupla úroveň chyby signifikantně z 0,25 až na 0,8.

Snížení přesnosti testu potvrzuje i studie Han (2013), která se zabývá *Item pocket method IP*, což je metoda umožňující **kontrolu a změny odpovědí během CAT testování**. Díky představované IP metodě nebyla tak výrazně snížena efektivita testu, než při jiných způsobech umožnění kontroly a oprav položek v CAT.

## 4.3 Vytváření CAT testů

Tvorba kvalitního CAT testu je mnohem větší výzvou, než tvorba testu klasického. (Wainer & Dorans, 2000). Je možné buď **převést klasický test** do počítačové adaptivní

podoby, nebo **vytvořit zcela nový**. V obou případech je potřeba kalibrovat položky a vytvořit dostatečně velkou položkovou banku.

### 4.3.1 Položková banka

Počítačové adaptivní testy vlastně vytvářejí **spoustu jednotlivých testů**, každý z nich je jiný (Wainer & Dorans, 2000). Čím lepší je položková banka, tím efektivnější budou adaptivní algoritmy. U klasických testů jsou položky konstruovány tak, aby měřily co nejpřesněji v oblasti průměru, mají tedy obtížnost ideálně 0,5. Položky v položkové bance CAT jsou ale konstruovány tak, aby **měřily v co největší šíři oblastí theta**.

Je potřeba, aby byla položková banka **dostatečně velká**, určitě mnohem větší, než položková banka u konvenčních testů (Wainer & Dorans, 2000). Při určování velikosti položkové banky je důležité zvažovat počet dimenzí v testu, použitý model a případné využití kontroly expozice položek a dalších faktorů. Nejméně by měla položková banka obsahovat kolem 100 položek, většinou jde ale spíš o **stovky a tisíce** (Kujal, 2008; Jelínek, Květon, & Vobořil, 2011b; Kingsbury & Wise, 2000). Potřebnou velikost lze přibližně odhadnout pomocí simulací (Thompson & Weiss, 2011). Je také potřeba zohlednit informační funkci testu, která by měla být nejvyšší v bodě, kde očekáváme nejvyšší četnost odhadů schopnosti. Tato očekávaná četnost nemusí nutně odpovídat průměru v normálním rozložení, ale může být nakloněna k oběma extrémům. Obtížnost položek v položkové bance je tedy vhodné volit **podle účelu testu**. Lze vytvořit hybridní položkové banky, ve kterých jsou jednotlivé položky kalibrovány různými IRT modely, a to včetně kombinace unidimenzionálních a multidimenzionálních modelů (Kingsbury & Wise, 2000).

V položkové bance by měl být **dostatek položek s různými obtížnostmi**, aby pokryly různé úrovně theta a měřily tak ve všech jeho úrovních dostatečně přesně (Wainer & Dorans, 2000). Je nezbytné, aby každá položka v CAT testu byla skutečně kvalitní, protože jediná nekvalitní položka by mohla narušit celý test. CAT totiž více **závisí na jednotlivých položkách**, než klasické testy, i proto, že jsou kratší. Jedna problematická položka může zvrátit celý test nebo alespoň snížit jeho přesnost a efektivitu.

#### 4.3.1.1 Postup vytvoření nové položkové banky

Na začátku je nutné samozřejmě **vytvořit znění jednotlivých položek**. Je třeba ověřit unidimenzionalitu testu a **zvolit vhodný IRT model** pro kalibraci položek (Wainer & Dorans, 2000). Principy kalibrace položek byly v podstatě popsány v kapitole o IRT. **Kalibrace položek** je vlastně zjištění hodnot parametrů položek (Baker, 2001). Proces kalibrace položek byl navržen Alanem Birnbaumem v roce 1968.

Pro kalibraci položek je potřeba **velký počet respondentů**, obvykle několik stovek, jako například ve studii Olea a kol. (2011), kde provedli kalibraci na 1 576 respondentech, nebo studii Vogels a kol. (2011), kde prováděli kalibraci dokonce na 2 041 respondentech. Nové položky je třeba **pretestovat**, což je možné provést prostřednictvím **simulací** (Wainer & Dorans, 2000). U těchto simulací je třeba zohlednit všechny skupiny theta, aby mohly být nalezeny možné problémy. Teprve po provedení simulací je vhodné provést **testování na respondentech**.

#### 4.3.1.2 Úpravy položkové banky - přidávání a odebírání položek

Jak již bylo zmíněno, je možné provádět úpravy položkové banky - **přidávání, odebírání a úpravy položek** (Wainer & Dorans, 2000). Položková banka v CAT tak prakticky **nikdy nemá definitivní podobu**. Při přidávání položek do testu jsou obvykle tyto položky zobrazovány při běžném testování vedle již kalibrovaných položek (Ozaki & Toyoda, 2009). Během **linkovací procedury** jsou pak nové položky kalibrovány do již existující škály (Kingsbury & Wise, 2000). Existuje více podob linkovací procedury. Mohou být znovu kalibrovány všechny položky a je srovnáván rozdíl v kalibraci původních a nových položek. Vhodnější je ale metoda, při které jsou kalibrovány pouze nové položky za použití respondentova odhadu schopnosti, získaného pomocí původních položek. Tato metoda je přesnější a jednodušší.

Pokud daná položka z nějakého důvodu dobře **neodpovídá modelu**, je lepší ji vyřadit, protože může způsobovat problémy a nepřesnosti (Kingsbury & Wise, 2000). Pro porovnání vhodnosti položek může být výhodné graficky zobrazit jejich křivky a srovnat je s těmi teoretickými. Pokud je pak k vyřazení určeno větší procento položek, je vhodné zvážit unidimenzionalitu testu a vhodnost použitého modelu.

#### 4.3.1.3 Bezpečnost a rovnoměrnost využití položkové banky

Pro kvalitní položkovou banku jsou důležité otázky její bezpečnosti a rovnoměrnosti využití položek. Pokud jsou **položky z testu známé veřejnosti**, mohou se respondenti na test připravit předem a dochází tak často ke zkreslování výsledků (Wainer & Dorans, 2000). Od respondenta, který položky zná, je již obtížné získat nějaké relevantní údaje o jeho úrovni schopnosti. Problémem jsou také **organizované krádeže** většího množství položek například konkurenční společností.

Na bezpečnost testu má vliv také použití nevhodného modelu pro výběr položek, který neumožňuje **kontrolovat využití položkové banky** a některé položky jsou tak používány velmi často, jiné zase jen výjimečně. Velmi často používané položky jsou pak méně odolné proti krádežím a zveřejňování. Podle Guo, Tay a Drasgow (2009) jsou ale celkově CAT testy poměrně odolné vůči zcizení menších počtů položek za předpokladu použití dostatečně rezistentních metod výběru položek.

Jednou z méně vhodných metod pro výběr položek je právě často používaná ***Fisher maximum item selection,*** **která se ukázala jako nejzranitelnější** z hlediska závažnosti narušení testové bezpečnosti u *high-stake tests* (rozhodných testů) (Qing Yi, Jinming Zhang, & Chang, 2008). Je však možné upravit tuto metodu podle Barrada a kol. (2008) a tím zlepšit její bezpečnost při minimální nebo žádné ztrátě přesnosti. Tato úprava spočívá v zahrnutí náhodnosti při výběru položky na základě její informační hodnoty.

Pro zabezpečení položkové banky se také často používají metody ***item exposure control*** (kontrola expozice položek)**.** Pro tyto metody je však potřeba ještě mnohem větší položková banka, než pro ostatní CAT metody (Parshall, 2002). Taková položkové banka může čítat i **několik tisíc položek**. Zvláštním, jednodušším, případem kontroly expozice položek je použití více položkových bank (Kingsbury & Wise, 2000). Tyto položkové banky pak rotují: nějakou dobu je používána jedna, pak je vystřídána a později použita znovu. Rozšiřování položkové banky ale může vést k méně přesnému hodnocení úrovně respondenta a nevyváženosti v obsahu (Barrada, Veldkamp, & Olea, 2008).

Dalšími metodami je například ***Sympson-Hetter method***, *The restricted method a The item-eligibility method* (Barrada, Abad, & Veldkamp, 2009). *Sympson-Hetter* procedura využívá parametru míry expozice položky $r_{max}$, který nese informaci o tom, jak často smí být

položka použita (Chen, Lei, & Liao, 2008). Velikost $r_{max}$ je obvykle zjišťována na základě řady simulací. Pokud je její hodnota např. $r_{max}$ = 0,2, znamená to, že bude položka prezentována každému pátému respondentovi. Vylepšením této procedury je metoda vícenásobné maximální expozice položek (Barrada, Veldkamp, & Olea, 2008). Tato metoda by měla přinášet vyváženější použití položkové banky a menší zkreslení odhadu rysu. Zvláštním případem je **kontrola maximální expozice položek u multidimenzionálních CAT** (Finkelman, Nering, & Roussos, 2009). Zde je používána například *The generalized Stocking-Lexis method*.

### 4.3.2 Převod klasických testů do CAT podoby

Převádění klasických testů do CAT podoby může být poměrně složité (Thompson & Weiss, 2011). Nejdříve je potřeba zajistit **vhodný software a dostatek prostředků pro vývoj nových položek**, protože počty položek v klasických testech zpravidla zdaleka nedostačují pro vytvoření položkové banky CAT testu. Je také vhodné provést **simulace** zjišťující, jestli skutečně převedení testu bude mít významný dopad na jeho délku a přesnost měření. Je na místě zvážit, jestli se takovýto nákladný převod testu vyplatí. Pro simulace je možné využít například programů CATSim nebo FireStar. Pro generování dat k simulacím pak slouží programy WINGEN nebo PARDSIM.

### 4.3.3 Vytváření nových CAT testů

Vytvoření nového CAT testu má obvykle několik fází (Jelínek, Květon, & Vobořil, 2011b). Jako u klasických testů jsou samozřejmě nejprve **vytvořeny samotné položky a je provedeno jejich psychometrické hodnocení**. V druhé fázi je již třeba **zvolit vhodný IRT model a provést kalibraci položek**. Pro nalezení vhodného modelu je třeba provést faktorovou analýzu pro **ověření unidimenzionality** (Thompson & Weiss, 2011). Parametry položek jsou pak převedeny na společnou škálu (Jelínek, Květon, & Vobořil, 2011b). Následně jsou většinou provedeny **simulace** a na základě nich úpravy testu.

Simulace mohou být několika forem. Často se používá **post-hoc (real-data) simulace**, která je obdobou **Monte Carlo simulace**, ale je prováděna na skutečných datech (Thompson

& Weiss, 2011). Zde může nastat problém, že ne všichni respondenti zodpovědí všechny položky. Lepší volbou může být **hybridní simulace**, kde jsou použita skutečná data jako u post-hoc, ale pokud nejsou některé odpovědi k dispozici, jsou automaticky generovány jako v Monte Carlo, na základě celkového theta respondenta.

V další fázi je třeba provést **Live-CAT testing,** tedy **testovat položky přímo na respondentech**. To se většinou provádí tak, že jsou vytvořeny **kotvící položky**, které jsou prezentovány všem respondentům a spolu s nimi je pro každou skupinu respondentů prezentována další část položek zbývajících. Je to hlavně proto, že pokud je kalibrována položková banka čítající 1 000 položek, každý respondent by musel podstoupit test o 1 000 položkách, což by bylo velmi časově náročné a odhad by mohl být nepřesný kvůli únavě apod.

Ani touto fází ale vývoj počítačového adaptivního testu nekončí. Test je vhodné i po jeho publikování dále **pozorovat, zdokonalovat a aktualizovat** (Thompson & Weiss, 2011). Obvykle se také provádí srovnávání skutečných výsledků testů se simulacemi a **hodnotí se shoda s původními předpoklady**. Porovnává se například počet položek, který byl potřebný pro dostatečně přesné hodnocení během simulací a ve skutečnosti. Pokud se výrazně liší, je vhodné provést dodatečné analýzy a simulace pro zjištění příčiny tohoto rozdílu.

## 4.4 Administrace

Zatímco u P&P testů musíme dbát na takové věci, jako je vhodné osvětlení, stejný čas testování pro každého, podvádění a opisování, u adaptivních testů musíme kontrolovat vhodnost počítačového vybavení, zabarvení a rozlišení monitoru, externí zařízení apod. (Wainer & Dorans, 2000). **Testovací algoritmus** je set pravidel specifikujících položky, které mají být zodpovězeny respondentem a jejich pořadí prezentace. Zahrnuje

- **Zahájení testu** - výběr první položky nebo několika položek
- **Pokračování testu** - metody výběru následujících položek
    - Může zahrnovat pravidla pro hádání, přeskakování položek, kontrolu zobrazení položek *(item exposure control)* apod.
- **Ukončení testu** - podle různých kritérií

Většinou jsou rozlišovány počítačové adaptivní testy s fixní délkou *(fixed-length)*, které jsou jednodušší na přípravu, nebo testy s variabilní délkou *(variable-length)* (Wainer & Dorans, 2000). Testy s **fixní délkou** obvykle bývají delší a jsou používány hlavně při simulacích, ve kterých se snažíme odhadnout vhodnou úroveň chyby pro testy s variabilní délkou. Pokud v takovém testu nemají být použity všechny položky, je nutné dobře zvolit počet použitých položek. Respondenti zde dosáhnou různých úrovní chyby. Testy s **variabilní délkou** mají různý počet položek a o ukončení testu obvykle rozhoduje dosažení nastavené přesnosti měření nebo standardní chyby. Respondentům s úrovní schopnosti, pro kterou máme dostatek vhodných položek, bude prezentováno menší množství položek pro dosažení požadované úrovně chyby. Při simulaci testy s variabilní délkou přinášejí informaci o tom, kolik položek bude přibližně potřeba pro dosažení určitých úrovní chyby.

V zahraničí se pro **simulaci CAT** testů používají různé počítačové programy. Je možné použít například balíček pro CAT testy *catR*, který je založený na statistickém programovacím jazyce R (Magis & Raiche, 2011). Mezi open source patří také *Firestar-D* pro dichotomní položky (Choi, Podrabsky, & McKinney, 2012), a *Firestar* pro polytomní položky (Choi, 2009). Dalším příkladem je také *SimulCAT* (Han, 2012) a mnoho dalších.

V ČR vzniká software **CATO *(computerized adaptive testing optimized)*** (Květoň, Jelínek, Denglerová, & Vobořil, 2008). Obsahuje editor pro nastavení testovací procedury a modul pro administraci testu klientovi. Je v něm možné formátovat položky pomocí HTML, včetně použití obrázků a zvuků. Je samozřejmě potřeba zadat parametry položek. Lze vytvořit i více testů a položkových bank, přičemž položky z různých položkových bank mohou být administrovány zvlášť nebo současně. Samozřejmostí je i volba způsobu zahájení testu, výběru položek a ukončení testu.

K administraci CAT testu je možné kromě počítače použít i **další mobilní zařízení**, jako mobilní telefony, PDA apod. (Triantafillou, Georgiadou, & Economides, 2008). Počítačové adaptivní testy v této formě pak nesou označení ***Computerized adaptive tests for mobile devices CAT-MD*.** U těchto zařízení je třeba zohlednit jejich různou velikost, která navíc bývá obvykle menší než u klasického počítače. Délku položek, případné použití multimediálního obsahu i celé uživatelské prostředí je tak třeba přizpůsobit. Vhodné je také provedení validačních studií.

### 4.4.1 Zahájení testu

Zahájení testu je první fází administrace počítačového adaptivního testu. Tato fáze je velmi důležitá, protože může mít **dopad na další pokračování testu, jeho délku i přesnost**. Pokud například použijeme na úvod testu vždy ty samé položky, jejich bezpečnost a tím i přesnost odhadu se velmi sníží, protože budou brzy známé široké veřejnosti (Wainer & Dorans, 2000). Podobná situace nastane i v případě, kdy zvolíme jako počáteční úroveň theta tu samou hodnotu, například 0. Pokud tedy **nemáme k dispozici úroveň theta** respondenta (například z předchozích testů nebo podobných metod), je možné vybrat položky náhodně ze skupiny položek se střední obtížností (třeba s obtížností mezi -0,5 a 0,5) (Thompson & Weiss, 2011). Alespoň nějakou informaci o pravděpodobné úrovni theta jedince nám může poskytnout i populační průměr, nebo využití několika zcela náhodně vybraných položek z položkové banky.

Probíhaly studie zjišťující, jaký v**liv má způsob zahájení testu na jeho pokračování a výsledek**. Podle Kingsbury a Wise (2000) pravděpodobně způsob zahájení testu jeho celkový výsledek příliš neovlivní. To platí spíše pro delší testy, u kratších testů může být toto tvrzení sporné. Rulison a Loken (2008) provedli studii, která se zabývá případy, kdy je odhad schopnosti po prvních položkách výrazně **nízký**, přestože skutečná schopnost respondenta je mnohem vyšší. V těchto případech obvykle schopnost v průběhu testu postupně stoupala. Podle nich mohou chyby v úvodních položkách u vysoce nadaných studentů vést k velkému podcenění výsledné úrovně schopnosti, a to i u delších testů. V opačném případě, kdy respondenti s nízkou úrovní schopnosti díky hádání získali v počátečních položkách mnohem vyšší úroveň schopnosti, docházelo k mnohem menší chybě. Zmiňují zde Bartonův a Lordův 4 parametrový model, který by měl být schopný tyto nepřesnosti výrazně omezit.

#### 4.4.1.1 Odhad latentního rysu

Základem pro výběr položek je tedy odhad úrovně schopnosti (rysu) **theta.** Jde o škálu schopnosti náležející modelu IRT, která má obvykle střední hodnotu 0 a odchylku 1 (Wainer & Dorans, 2000). Většinou se její hodnota pohybuje mezi -3 a +3, ale teoreticky může dosahovat hodnot od minus nekonečna po nekonečno. Každý test složený z položek, které odpovídají použitému modelu, může produkovat skóre na této škále schopnosti.

Samotný odhad schopnosti je pak obvykle prováděn pomocí **metody maximální věrohodnosti MLE** nebo **Bayesovských metod** *expected a posteriori a maximum a posteriori* (Jelínek, Květon, & Vobořil, 2011b). Tyto metody byly zmíněny v kapitole o IRT. V případě MLE může nastat problém s odhadem schopnosti u respondentů, kteří zodpověděli všechny položky správně, nebo všechny špatně. Pro hodnocení úrovně theta na základě prvních položek v testu je tedy lepší použití Bayesovských metod, kde je ale odhad rysu zatížen hypoteticky stanoveným rozložením v populaci. Pro finální odhad je tedy lepší použít MLE.

### 4.4.2 Volba položek

Po úvodním vyhodnocení úrovně latentního rysu se nám nabízí více možností, jak volit další položky. Dříve byly používány velmi **jednoduché větvící modely**, kde byla při správné odpovědi vybrána obtížnější položka, a při odpovědi nesprávné jednodušší (Wainer & Dorans, 2000). Následovaly složitější způsoby, ve kterých bylo ale stále potřeba používat jistých triků pro zrychlení celého procesu, aby respondent nemusel dlouho čekat, protože počítače byly pomalejší. Dnes, kdy jsou počítače mnohonásobně rychlejší, jsou položky většinou vybírány ihned po zodpovězení dané položky na **základě zjištěné úrovně respondenta**, která je přepočítávána po každé odpovědi. Pokud je přesto přepočítávání z nějakého důvodu problém, je možné využít ***info table***, což je tabulka obsahující seznam položek seřazených podle úrovně informačního přínosu, který přinášejí ve vztahu k různým úrovním theta. Není pak nutné znovu a znovu přepočítávat informační přínos všech položek v celé položkové bance.

Jednou z nejčastěji používaných metod je ***Fisher information selection method FIS***, což je metoda vybírající následující položku podle nejvyššího informačního přínosu pro dané theta. Obvyklým postupem po zodpovězení jedné položky je provedení odhadu úrovně theta, vyhodnocení informačního přínosu každé položky pro toto nové theta a následně výběr a administrace dosud nezodpovězené položky s nejvyšším informačním přínosem. Pokud je v jednom testu použito více položkových bank, neumožňuje tato metoda vyvažování obsahu *(content balancing)* (Zheng, Chang, & Chang, 2013). Může tak vést k nevyváženosti ve využití položkové banky a tím i k nepřesnosti odhadu.

Další možností jsou **metody maximální odhadované přesnosti**, např. Owenova Bayesovská procedura (Wainer & Dorans, 2000). Ta dokáže odhadnout přesnost odhadu, které dosáhneme použitím určité položky. Pracuje na principu minimalizace variability předchozí distribuce (Kingsbury & Wise, 2000). Na začátku testu umožňuje větší variabilitu obtížnosti, kterou postupně snižuje. Je výpočetně méně náročná než FIS, protože **nevyžaduje iterativní odhad** (Wainer & Dorans, 2000). Přesto se používá méně. Obě zmíněné metody ale mohou snadno vést k časté expozici určitých položek, což je třeba ošetřit metodami k tomu určenými.

Metodami pro volbu položek se zabývalo velké množství autorů. Deng a kol. (2010) **porovnával tři procedury výběru položek** a ty ještě navíc srovnával s naprosto náhodným výběrem. U metod porovnávali úrovně chyby, reliabilitu, odhad schopnosti a využití jednotlivých položek. Fischerova metoda se ukázala přesnější než metody *Stratified multistage computer adaptive testing STR* a *Refined stratification procedure that allows more items to be selected from the high strata and fewer items from the low strata USTR*, ale se špatným využitím položek v položkové bance. USTR metoda dobře redukovala velikost chyby a měla lepší využití položkové banky než Fischerova metoda.

Šest metod volby položek srovnával Barrada a kol. (2010). Studie byla provedena na CAT testech s fixní délkou a byla zdůrazněna potřeba zohlednění přiměřené expozice položek. Nejlepší se ukázaly metody *Kullback-Leibler weighted by likelihood, The proportional method a The maximum information stratification method with blocking*. V Cheng a Morgan studii (2013) se ukázala jako nejlepší metoda *The maximum priority index MPI*, a to hlavně v oblasti udržování nízké úrovně expozice položek.

U **polytomních položek** může být užitečná metoda *The maximum posterior weighted information MPWI,* která je poměrně výkonná i ve srovnání se složitějšími a sofistikovanějšími metody z této skupiny (Choi & Swartz, 2009). Pro **Likertovu škálu** jsou vhodné například metody *Current estimate/ability confidence interval method* a *Cut score/sequential probability ratio test method*, které se ukázaly jako velmi přesné a efektivní (Wang & Liu, 2011).

### 4.4.3 Ukončení testu

Ukončovací pravidla v CAT obvykle spadají do dvou hlavních kategorií (Reckase, 2009):

- Testy s pevnou délkou
- Testy s variabilní délkou.

**U testů s pevnou délkou** je test ukončen po administraci daného počtu položek. Může jít o stanovenou část položek z položkové banky nebo o všechny dostupné položky (Wainer & Dorans, 2000). Tento způsob je využíván hlavně při hromadném testování (Jelínek, Květon, & Vobořil, 2011b). Vhodný počet položek je obvykle zjišťován pomocí simulací.

**Testy s variabilní délkou** jsou trochu komplikovanější. Test bývá obvykle ukončen na základě dosažené chyby měření nebo na základě dosažení časového limitu (Reckase, 2009). Obvyklejší je kritérium zahrnující **chybu měření**. Počet položek zde pak obvykle závisí na umístění respondenta na theta škále. Mohou nastat situace, kdy není požadované chyby měření dosaženo ani po vyčerpání všech položek, nebo stanoveného maximálního počtu položek (Wainer & Dorans, 2000). Úroveň chyby je obvykle nastavována na 0,3162, což je hodnota odpovídající informačnímu přínosu s úrovní 10 a tedy reliabilitě přibližně 0,9. Její úroveň ale může být určena opět i na základě simulací.

Velmi sporné je použití **časového limitu** v CAT. Podle Wainer a Dorans (2000) může být výhodný zvláště u **testů schopností**. Někdy ale při jeho použití dochází ke zkreslování výsledků. S časovým omezením se položky mohou stát obtížnějšími (Kingsbury & Wise, 2000). Již existují speciální **CAT modely pro časově omezené testy**, ale i tak je těžké nastavit časový limit tak, aby byl spravedlivý. Respondent, který dostane 40 obtížných otázek, může mít jen těžko stejný časový limit jako respondent, který dostane 40 jednoduchých. Možností by mohlo být zavést časový limit pro každou položku zvlášť, podle její obtížnosti, ale to by bylo velice komplikované. Přítomnost časového limitu, podobně jako nemožnost kontroly položek, výrazně zvyšuje úzkost respondentů při testování (Kingsbury & Wise, 2000).

Použití časového limitu v CAT bylo simulováno Schmitt a kol. (2010). Při zkrácení limitu došlo ke **zhoršení přesnosti odhadu schopnosti**. Respondenti s vyšší úrovní

schopnosti byli ovlivnění více, než ti ostatní. Přesto ale autoři hodnotí **CAT jako relativně odolné proti použití časového limitu**, pokud je ho použito s rozvahou.

Weiss (2004) upozorňuje, že pokud je test ukončen na základě dosažení stanoveného počtu položek nebo časového limitu, aniž by bylo dosaženo požadované chyby, nemusí být výsledek dostatečně spolehlivý a přesný. Podle Wainer a Dorans (2000) je pak nejlepší kombinace pravidel. Při určování vhodného způsobu ukončení testu je opět vhodné využít možnosti simulace různých situací a scénářů testování - porovnání potřebného množství položek na různé úrovni chyby apod. (Thompson & Weiss, 2011).

Postupně se vyvíjejí **nová kritéria pro ukončení testu**. Jedním z nich je i *Predicted standard error reduction PSER* (redukce předpovídané standardní chyby), které představují Choi, Grady a Dodd (2011). Ve srovnání s jinými kritérii PSER zefektivňuje využití položkové banky, takže je potřeba administrovat méně položek a konečná přesnost testu je vyšší.

## 4.5 Další využití CAT

Počítačové adaptivní testování je stále častěji využíváno také pro ***e-learning*** (Youngseok, Jungwon, Sugjae, & Byung-Uk, 2010). Hlavním cílem v těchto testech tedy není získání úrovně schopnosti respondenta, ale vlastně **výuka na úrovni obtížnosti vyhovující danému jedinci**. Ve studii testu pro výuku angličtiny se studenti, v porovnání s obvyklejšími způsoby výuky, učili **efektivněji, projevovali větší zájem o učební látku a vyšší motivaci**. Mezi takovéto e-learning CAT patří například *UZWEBMAT*, který slouží pro výuku matematiky (Özyurt, Özyurt, Baki, & Güven, 2012). Test je založen na 3PL modelu a zahrnuje 752 položek, přičemž v každém testu je jich prezentováno 30. K výuce matematiky slouží také *Maths Garden*, což je webový monitorovací systém zahrnující prostředí pro procvičování aritmetiky u dětí (Klinkenberg, Straatemeier, & van der Maas, 2011). Metoda výběru položek zde byla neobvykle upravena tak, aby byly voleny položky, u kterých má respondent pravděpodobnost správné odpovědi 0,25 místo běžně používané pravděpodobnosti 0,5. Je to proto, aby byly položky skutečnou výzvou. Tento test též prokázal větší efektivitu učení a lepší motivaci dětí (až třetinu úloh děti řešily ve svém volném čase).

CAT může být použito také jako **klasifikační test**, kde není rozhodující přesná úroveň theta, ale pouze hranice, která rozhodne, zda respondent uspěl či neuspěl v testu (Gnambs & Batinic, 2011). Klasifikační testy mohou rozdělovat respondenty i do několika málo skupin, třeba pro rozhodování do které úrovně nějakého kurzu bude nejvhodnější respondenta zařadit - začátečník, mírně pokročilý, pokročilý - jako je tomu například u *Programming Adaptive Testing* PAT (Chatzopoulou & Economides, 2010).

### 4.5.1 Současné psychologické testy založené na CAT

Ve světě je již k dispozici mnoho CAT testů. Jedním z nejstarších je již zmiňovaný ***Computerized adaptive version of the armed sciences vocational aptitude battery CAT-ASVAB*** (Wainer & Dorans, 2000). Dále pak například TOEFL, SAT, GRE, AMT, GMAT a další. Počítačové adaptivní testy jsou již také součástí **Vienna test system,** např. *Adaptive spatial ability test* měřící inteligenci (Schuhfried, 2013).

Do CAT podoby jsou také převáděny některé **klinické škály**. Jednou z nich je například škála deprese *Center of Epidemiological Studies-Depression scale CES-D*, kde se CAT verze, testovaná na 1 392 respondentech, ukázala jako velmi efektivní (Smits, Cuijpers, & van Straten, 2011). Dalším testem pro hodnocení deprese v CAT verzi je také *D-CAT*, který byl navíc porovnáván s již zmíněnou škálou CES-D, BDI a HADS s korelací 0,68 - 0,77. Opět bylo prokázáno výrazné zefektivnění testu, z původních 64 položek bylo průměrně použito 10 (Fliege et al., 2009). U škály pro měření úzkosti A-CAT bylo při srovnávání s klasickými testy dosaženo korelace 0,56 - 0,66, při průměrné prezentaci 6 otázek z původních 50 (Becker et al., 2008).

# 5 Shrnutí

Přestože je u **nás stále IRT opomíjena**, má své nesporné výhody oproti klasické testové teorii. Je také nenahraditelným matematickým aparátem, na kterém jsou postaveny **počítačové adaptivní testy**. Existují **unidimenzionální a multidimenzionální IRT modely**, které se podle druhů položek dělí dále také na **dichotomní a polytomní**. Pro použití unidimenzionálních modelů je nutné splnit několik **základních předpokladů**: unidimenzionalitu, lokální nezávislost a shodu modelu s daty. Pomocí těchto modelů pak

bývá provedena **kalibrace testu**, tedy odhady parametrů položek a **odhad úrovní respondentů**. Pro oba tyto odhady je možné získat hodnoty **informačního přínosu a standardní chyby**.

Počítačové adaptivní testování je **zvláštní formou počítačových testů**. Lze vytvořit zcela nový CAT test, ale je také možné převést stávající testy, založené na klasické testové teorii, do CAT podoby. CAT test je tvořen souborem položek, nazývaným **položková banka**. Administrace CAT obvykle začíná **volbou prvních položek, odhadem úrovně respondenta** a pokračuje střídavou **volbou další položky** a opětovného hodnocení úrovně respondenta. Test může být **ukončen na základě různých pravidel**, například dosažení maximálního počtu položek, požadované úrovně přesnosti nebo úrovně chyby případně na základě časového limitu.

Tato práce se bude dále zabývat **empirickým posouzením splnění předpokladů IRT modelů, volbou vhodného modelu a kalibrací položek testu eEOD**. Zároveň bude provedena **simulace CAT** testu na datech z reálné studie. V další fázi bude následovat **reálná online** administrace CAT eEOD, testu eEPQ a krátkého sebehodnotícího dotazníčku. Budou porovnávány výsledky hrubého skóru eEOD, simulací a výsledků reálné administrace CAT eEOD. Zároveň bude provedena **jednoduchá validizační studie**. V poslední fázi bude provedeno srovnání CAT verze eEOD s jeho **klasickou papírovou obdobou**. Předpokládám poměrně **výraznou shodu výsledků v jednotlivých typech administrací** i v jednotlivých testech mezi sebou. U CAT verze eEOD pak předpokládám **možnost výrazného zkrácení testu při zachování dostatečné přesnosti měření.**

# II. Praktická část

# 6 Metodologie

**Cílem této diplomové práce je srovnání různých forem administrace dotazníku eEOD.** Srovnávána bude klasická administrace tužka-a-papír, počítačová online administrace, počítačová adaptivní administrace, provedená též online, a administrace simulovaná z reálných dat. Bude provedeno také srovnání CAT eEOD s podobnými testy měřícími extraverzi (EPQ, sebehodnocení).

Cílem první fáze výzkumu je vytvoření počítačové adaptivní verze škály extraverze Eysenckova osobnostního dotazníku (dále nazývaného eEOD), následuje testování respondentů touto metodou a porovnání výsledků s výsledky klasických způsobů testování. Práce se bude zaměřovat na **výhody a nevýhody plynoucí z CAT administrace** ve srovnání s počítačovou a tužka-a-papír administrací.

Předchozí výzkumy, které jsou uvedeny v teoretické části, naznačují velký zájem o tuto oblast, ale produkují **odlišné výsledky**. Zdá se, že **u některých testů je CAT vhodnou variantou, jinde však nepřináší dostatečně výrazné zlepšen**. Obvykle je lepších výsledků dosahováno v oblasti výkonových testů. Zároveň je většina těchto výzkumů založena pouze na výsledcích simulací CAT (např. Žitný, 2011), navíc jen v některých případech z reálných dat.

Právě z důvodu nedostatku výzkumů s **reálnou administrací CAT** bude v této práci taková administrace provedena, aby mohlo být posouzeno, zda se nějak liší od počítačových simulací. Stejně tak není dostatek výzkumů hodnotících **online počítačové administrace** v porovnání s klasickou počítačovou administrací případně administrací tužka-a-papír. Právě tento typ administrace bude v mé práci stěžejní.

Výzkum bude rozdělen do 3 fází. V **první fázi** je hlavním cílem **sběr dat pro kalibraci položek eEOD**, ověření vhodnosti modelů IRT, převedení eEOD do CAT podoby a provedení simulací pomocí autorkou naprogramovaného software. V **druhé fázi** bude nejdříve naprogramováno online prostředí pro sběr dat. Následně budou **sbírána data prostřednictvím metod CAT eEOD, eEPQ a sebehodnotícím dotazníkem**. Výsledky jednotlivých testů a způsobů administrací budou porovnány. **Třetí fáze** bude zahrnovat pouze **administraci tužka-a-papír verze eEOD** dotazníku a porovnání výsledků s předchozími výsledky.

## 6.1 Výzkumné otázky

**Jsou jednoduché IRT modely vhodné pro Eysenckův osobnostní dotazník? Který model je nejvhodnější?**

Tato otázka se zabývá hlavně ověřením vhodnosti dostupných unidimenzionálních IRT modelů, shodě modelu s daty a splnění podmínek pro použití těchto modelů.

## 6.2 Hypotézy

Ze zmíněných výzkumných cílů byly vytvořeny následující hypotézy, které budou ověřovány prostřednictvím statistických metod:

**Shoda výsledků klasického eEOD a simulované CAT eEOD**

$H_0$1 - Výsledky simulovaného CAT eEOD, při použití všech položek a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky klasického eEOD z první fáze výzkumu.

$H_0$2 - Výsledky simulovaného CAT eEOD, při použití různého počtu položek a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky eEOD z první fáze výzkumu.

$H_0$3 - Výsledky simulovaného CAT eEOD, při použití různých úrovní chyby pro ukončení testu a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky eEOD z první fáze výzkumu.

**Shoda výsledků CAT eEOD při reálné administraci a simulaci**

$H_0$4 - Průměry výsledků reálné administrace CAT eEOD, při použití všech položek, se signifikantně neliší od průměrů výsledků simulovaného CAT eEOD.

$H_0$5 – CAT eEOD test bude stejně efektivní, jako klasický eEOD test (tzn. bude potřeba použít všechny položky k dosažení dostatečné přesnosti testu).

$H_0$6 - Výsledky reálné administrace CAT eEOD, při použití sníženého počtu položek, signifikantně nekorelují s výsledky simulovaného CAT eEOD.

**Shoda výsledků CAT eEOD s výsledky dalších testů**

$H_0$7 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky eEPQ.

$H_0$8 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky sebehodnocení.

**Shoda výsledků počítačové a počítačové adaptivní verze eEOD s výsledky klasické tužka-a-papír administrace**

$H_0$9 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky klasické tužka-a-papír administrace.

## 6.3 Použité metody

### 6.3.1 EOD (eEOD)

Eysenckův osobnostní dotazník patří mezi osobnostní dotazníky konstruované na základě faktorové analýzy. Byl vytvořen na základě teorie podobně jako 16PF nebo KUD a je jednou z 9 metod vytvořených Eysenckem v letech 1947 - 1975 (Stančák, 1996). **Eysenck personality inventory EPI pochází z roku 1964 a patří k nejrozšířenějším osobnostním dotazníkům** po celém světě (Svoboda, 1999). **U nás je známý jako EOD** (Stančák, 1996). Dotazník **přeložil Smékal a Stančák** a na naši populaci byl upraven a restandardizován Miglliernim a Vonkomerem. Příručku přeložila Krállová a spol. roku 1968. EPI je zdokonalený Maudsley personality inventory (vydaný v roce 1959) (Eysenck & Eysenck, 1968).

Měří hlavně **extraverzi a neuroticismus**. Skládá se ze **dvou paralelních forem**, z nichž každá zahrnuje 57 otázek - 24 otázek měří extraverzi, 24 neuroticismus a 9 položek tvoří lži škálu (Svoboda, 1999). Respondent na položky odpovídá **ano nebo ne**. Vyplnění dotazníku trvá většinou 5-15 minut. Věková hranice není jasně stanovena, většinou je test doporučován od 14 let. Administrátor test vyhodnocuje podle šablony. Pokud respondent odpoví na položku klíčově, značí to extraverzi a respondent získává jeden bod. Výsledný skór tvoří **součet skórů jednotlivých položek**. Skóry jsou pak **porovnávány s normami**. EOD

test restandardizovaný na naši populaci vydalo Bratislavské vydavatelství roku 1979. Je používán **ve výzkumu, pro klinické a poradenské účely** a často také v personalistice.

Byla ověřována test-retest a split-half reliabilita testu. **Reliabilita test-retest** byla měřena na dvou skupinách - skupina X 92 respondentů a skupina Y 27 respondentů. Skupina X absolvovala opakování testu po jednom roce, skupina Y po 9 měsících. Reliabilita u skupiny X dosahovala u škály extraverze celkově 0,88 a u skupiny Y 0,94. Reliabilita u jednotlivých forem testu byla o něco nižší. S**plit-half reliabilita** byla provedena na vzorku 1 655 normálních respondentů, 210 neurotických respondentů a 90 psychotických respondentů a dosahovala 0,85 - 0,95. Při **srovnání s jinými testy** škála extraverze korelovala s extraverzí v KUD r = 0,664 a s Bellovou sociální přizpůsobivostí r = 0,560 (Svoboda, 1999).

V tomto výzkumu bude používána **pouze škála extraverze** z důvodu využití jednoduchých IRT modelů a tedy nutnosti splnit předpoklad unidimenzionality. IRT má samozřejmě prostředky, jako multidimenzionální modely, díky kterým by bylo možné převést test jako celek, to ale není v možnostech této práce. Zároveň budou použity **položky z obou forem testu**, aby byl k dispozici dostatečný počet položek pro CAT administraci. Tato mutace testu, zahrnující položky z obou forem a pouze škály extraverze, bude dále nazývána **eEOD**. Test eEOD zahrnuje celkem 48 položek.

#### 6.3.1.1 CAT eEOD

CAT eEOD je test **založený na teorii odpovědi na položku**. Jednotlivé položky testu byly na základě dat získaných v první fázi výzkumu (n=124) **kalibrovány pomocí programu PARAM** - Calibration Software for the 1 & 3 Parameter Logistic IRT Models.

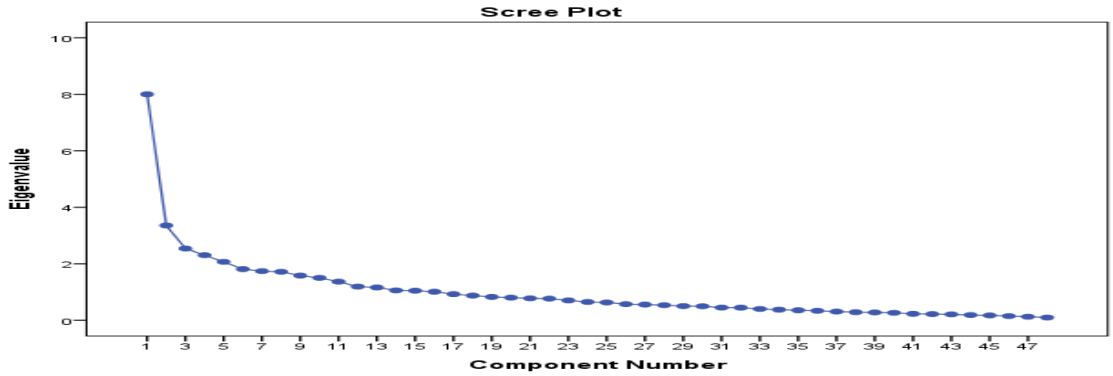
Při kalibraci testu bylo nutné ověřit splnění podmínek pro použití IRT modelů. První podmínkou je **předpoklad unidimenzionality**. Lze očekávat, že je tento předpoklad splněn, vzhledem k použití pouze jedné dimenze EOD, a to dimenze extraverze. Podle Jelínka a kol. (2011a) znamená splnění předpokladu unidimenzionality zároveň i to, že test je **lokálně nezávislý**, tedy splňuje i druhý předpoklad.

**Unidimenzionalita je ověřována pomocí faktorové analýzy.** Pokud první faktor vysvětluje alespoň 20% rozptylu, je možné test považovat za unidimenzionální (Templin, 2013). Mezi prvním a druhým faktorem by měl navíc být co největší rozdíl a ostatní faktory by měly být menší nebo rovny jedné. **Přestože lze předpokládat unidimenzionalitu, v eEOD testu nebyla bohužel prokázána.** První faktor vysvětluje **pouze 16,671%** rozptylu, což je příliš málo. I **rozdíl mezi prvním a druhým faktorem je malý** a ostatní faktory dosahují hodnot vyšších než 1. Protože nebyl jasně ověřen předpoklad unidimenzionality testu, nelze tedy automaticky předpokládat jeho lokální nezávislost.

**Tabulka 1 Ověření unidimenzionality testu**

| Faktor | Celkem | % rozptylu | Kumulativní % |
|---|---|---|---|
| 1 | 8,002 | 16,671 | 16,671 |
| 2 | 3,357 | 6,994 | 23,665 |
| 3 | 2,541 | 5,294 | 28,960 |
| 4 | 2,305 | 4,802 | 33,762 |
| 5 | 2,069 | 4,311 | 38,073 |
| 6 | 1,812 | 3,774 | 41,848 |
| 7 | 1,741 | 3,626 | 45,474 |
| 8 | 1,718 | 3,580 | 49,054 |
| 9 | 1,583 | 3,298 | 52,352 |
| 10 | 1,502 | 3,129 | 55,480 |
| 11 | 1,368 | 2,850 | 58,330 |
| 12 | 1,193 | 2,485 | 60,815 |
| 13 | 1,162 | 2,421 | 63,236 |
| 14 | 1,058 | 2,205 | 65,441 |
| 15 | 1,049 | 2,185 | 67,626 |
| 16 | 1,012 | 2,109 | 69,735 |

**Graf 1 Ověření unidimenzionality testu**



Pokud tedy nemůžeme předpokládat splnění prvních dvou podmínek pro použití jednoduchých IRT modelů, **nelze předpokládat příliš spolehlivé výsledky při kalibraci testu, shodě modelu s daty a následně ani při měření schopnosti respondentů**. Přesto se test pokusím pro svůj výzkum použít.

Předpoklad **shody modelu s daty** se testuje srovnáváním skutečných a očekávaných pravděpodobností správné odpovědi pro každou položku na různé úrovni theta pomocí upraveného $X^2$-testu dobré shody. Používá se následující vzorec:

$$X^2 = \sum_{j=1}^{J} m_j \frac{\left[p(\theta_j) - P(\theta_j)\right]^2}{P(\theta_j)\, Q(\theta_j)}$$

Vzorec pro výpočet shody modelu s daty

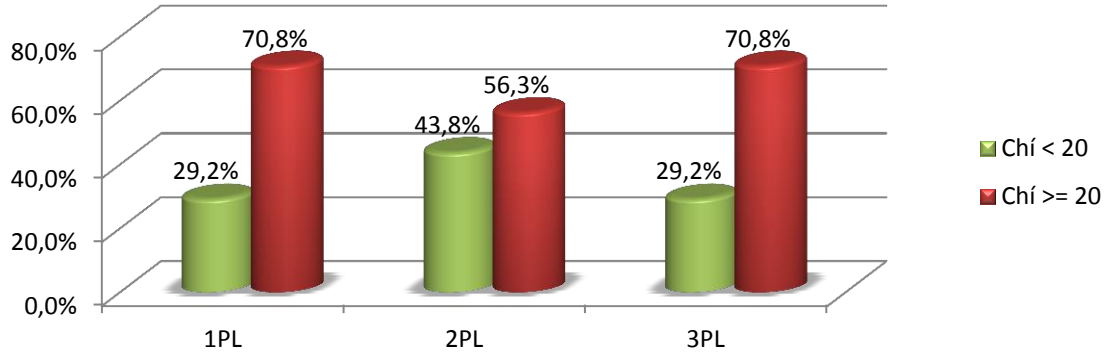
Shoda modelu s daty byla **nejvyšší u 2PL modelu, kde byla shoda nalezena u 21 položek (43,8%), což je poměrně málo**. Požadovaná minimální shoda modelu s daty byla stanovena na 75%. U 3PL a 1PL modelu byla shoda pouze u 14 položek (29,2%). Zde tedy bohužel **nebyla shoda modelu s daty prokázána**.

Příčinou může být to, že skutečné proporce správných odpovědí jsou tak **široce rozptýlené**, že dobrá shoda dat nemůže být dosažena s pomocí žádného modelu (Baker, 2001). Pravděpodobnou příčinou zde může být také **příliš malý vzorek**, který neumožňuje vytvoření dostatečného množství skupin pro spolehlivější porovnání skutečných dat s teoretickými.

Tabulka 2 Shoda modelu s daty pro 1P, 2PL a 3PL model

| | Shoda nalezena (Chí < 23,213) | | Shoda nenalezena (Chí >= 23,213) | |
|---|---|---|---|---|
| 1PL | 14 | 29,2% | 34 | 70,8% |
| 2PL | 21 | 43,8% | 27 | 56,3% |
| 3PL | 14 | 29,2% | 34 | 70,8% |

Graf 2 Shoda modelu s daty pro 1PL, 2PL a 3PL model




Tabulka 3 Deskriptivní statistika odhadnutých hodnot parametrů a a b

| | N | Min | Max | Průměr | Std. Deviation |
|---|---|---|---|---|---|
| **Parametr a** | 48 | 0,090 | 2,000 | 1,099 | 0,683 |
| **Parametr b** | 48 | -5,000 | 4,210 | 0,295 | 1,515 |

V reálné administraci CAT testu bylo jako **ukončovací kritérium nastaveno dosažení maximálního počtu položek**. Pro účely simulací, srovnávání s dalšími fázemi výzkumu a s dalšími testy ve druhé fázi tedy **respondenti zodpověděli všech 48 položek**.

Pro výběr položek byla zvolena metoda ***Maximum Fisher Information MFI***, což je nejčastěji používaná metoda (Cheng & Morgan, 2008). Jednoduše vybírá položky, které maximalizují Fisherovu informaci hodnocenou pro danou úroveň schopnosti. Zároveň se snaží minimalizovat standardní chybu. Nevýhodou této metody je nedostatečné řešení problematiky **vyrovnávání obsahu** a nadměrné **expozice položek**. Je tedy zranitelnější proti narušení bezpečnosti testu a krádežím položek (Qing Yi, Jinming Zhang, & Chang, 2008). Při využití kvalitnějších a tím i bezpečnějších metod je ale třeba mít k dispozici mnohem větší položkové banky.

### 6.3.2 EPQ (eEPQ)

Eysenck personality questionnaire EPQ byl **vydán v roce 1975** a opírá se o dotazníky MMQ a EPI. Stejně jako EPI měří **extraverzi, neuroticismus a zahrnuje lži škálu** (Eysenck & Eysenck, 1993). Navíc však zavádí **novou proměnnou psychoticismu**, která je měřena 25

položkami. K dispozici je také revidovaná verze EPQ-R. Příručka k tomuto dotazníku u nás byla vydána v roce 1993, do češtiny ji **přeložila Emilie Smékalová.** Jsou k dispozici **normy na slovenské populaci** ve věku 16-70 let. Hodnocení je podobné jako u EOD, jeden bod získává respondent za klíčovou dopověď, tedy odpověď směřující k extraverzi.

Zde bude EPQ použit podobně jako EOD pouze **ve formě mutace o jediné škále** - škále extraverze. Celkem bude tedy eEPQ zahrnovat 23 položek. Bude použit pouze pro **jednoduché ověření validity** počítačové adaptivní verze eEOD.

### 6.3.3 Sebehodnocení

Sebehodnotící dotazník zahrnoval **10 protikladných přídavných jmen**, popisujících na jedné straně **introverzi** a na druhé straně **extraverzi**. Respondenti měli u každé dvojice zvolit, ke kterému pólu se blíží na **7 bodové škále.** Nejnižší možný počet bodů, který bylo možno získat, byl 10 bodů, což je hodnota odpovídající naprosté introverzi. Nejvyšší počet bodů byl 70, což je hodnota odpovídající naprosté extraverzi. Tento dotazník bude sloužit pro **velmi jednoduché hodnocení validity** počítačové adaptivní verze eEOD spolu s dotazníkem eEPQ.

## 6.4 Sběr dat

Sběr dat v **první fázi** výzkumu probíhal **v lednu 2013** prostřednictvím webového formuláře google forms. Respondenti vyplňovali **dotazník eEOD** a pouze několik **demografických otázek** týkajících se věku a pohlaví. Respondenti byli zároveň požádáni o souhlas s účastí v druhé fázi výzkumu. Zúčastnilo se celkem **124 respondentů.**

**Druhá fáze** výzkumu zahrnovala vyplnění **krátkého sebehodnocení, počítačové adaptivní verze eEOD a nakonec dotazník eEPQ**. Celkem respondentům vyplnění dotazníků trvalo přibližně 15-20 minut. Tato fáze výzkumu probíhala v **srpnu a první polovině září 2013**, tedy přibližně **7-8 měsíců po první fázi**. Respondenti, kteří se zúčastnili již první fáze výzkumu nedostávali žádné doplňující položky, ostatní respondenti byli vyzváni k uvedení věku a pohlaví. **Zúčastnilo se celkem 137 respondentů**, z nichž 69 byli respondenti z první fáze výzkumu a 68 byli noví respondenti.

Třetí a poslední fáze výzkumu se zúčastnilo pouze **10 respondentů**. Tato fáze probíhala v **říjnu 2013**, tedy **pouhý měsíc po ukončení druhé fáze**. Bylo by vhodnější tuto fázi zařadit přibližně se stejným odstupem, jako byl mezi první a druhou fází, to ale bohužel nebylo z časových důvodů možné. Tato fáze výzkumu je pouze doplňující a respondenti, vybraní náhodně z nových respondentů, vyplňovali **klasickou tužka-a-papír verzi dotazníku eEOD**. Výsledky této fáze výzkumu budou sloužit k porovnání klasických tužka-a-papír testů s testy počítačovými.

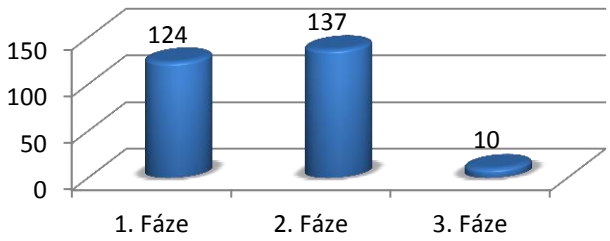
# 7 Popis vzorku

Jak již bylo popsáno v kapitole o metodologii, výzkum probíhal ve třech fázích. První fáze výzkumu se zúčastnilo **124 respondentů**. Druhé části výzkumu se zúčastnilo **jen 69 (55,6%)** respondentů, kteří absolvovali test v první fázi. Původních respondentů tak bylo v druhé části výzkumu 69 (50,4%). Znovu se tedy zúčastnilo 55,6% respondentů z první fáze výzkumu. K původním respondentům v druhé fázi **přibylo 68 (49,6%) nových respondentů**. Ve 3. fázi výzkumu vyplnilo klasický papírový test 10 respondentů náhodně vybraných z respondentů, kteří se zúčastnili pouze druhé fáze výzkumu.

Tabulka 4 Počet respondentů v jednotlivých fázích výzkumu

| | Četnost |
|---|---|
| 1. fáze | 124 |
| 2. fáze | 137 |
| 3. fáze | 10 |

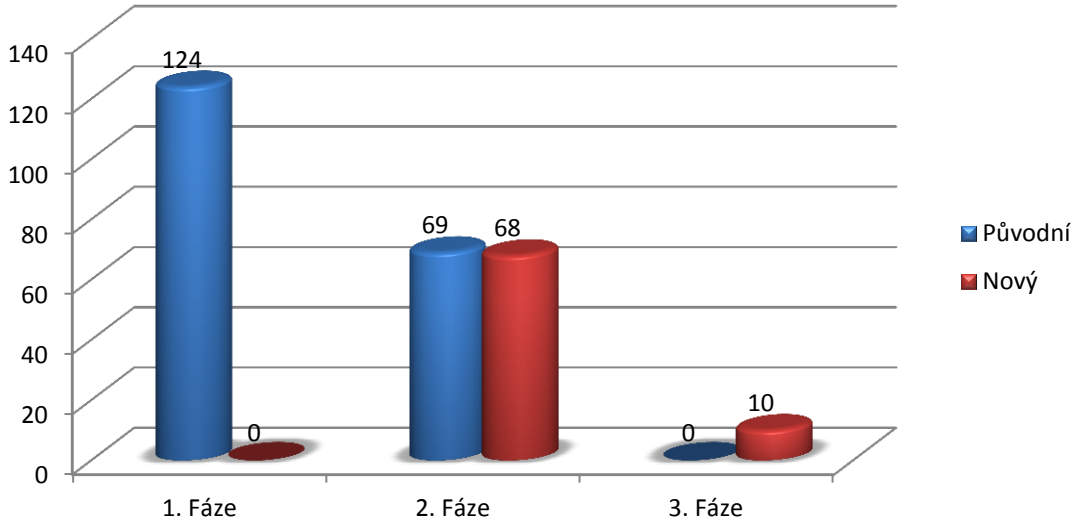
Graf 3 Počet respondentů v jednotlivých fázích výzkumu




Tabulka 5 Rozdělení původních a nových respondentů ve vzorku druhé fáze výzkumu

| | Četnost | % |
|---|---|---|
| Původní | 69 | 50,4 |
| Nový | 68 | 49,6 |
| Celkem | 137 | 100,0 |

Graf 4 Rozdělení původních a nových respondentů ve vzorku




## 7.1 Pohlaví

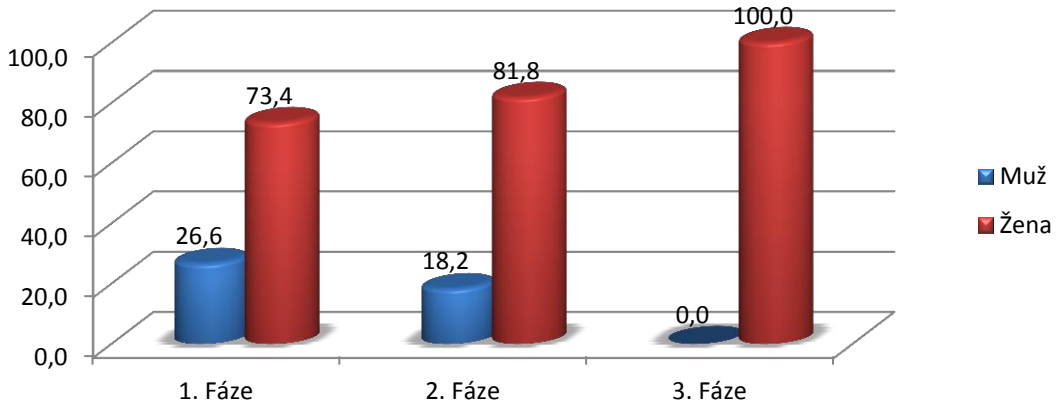
Ve vzorku **není vyrovnaný počet mužů a žen**. První fáze výzkumu se zúčastnily téměř **tři čtvrtiny žen (91; 73,4%)** a jen jedna **čtvrtina mužů (33; 26,6%)**. Nevyrovnanost v druhé fázi výzkumu se ještě prohloubila. Zúčastnilo se jen **25 mužů (18,2%), tedy necelá jedna pětina, a 112 žen (81,8%)**.

Při podrobnějším pohledu na druhou fázi výzkumu vidíme, že **poměr mužů a žen je podobný ve skupině nových a původních respondentů** ve druhé fázi výzkumu. Mužů je mezi respondenty z původního vzorku ve druhé fázi celkem 14 (10,2%), mezi novými respondenty pouze 11 (8,0%). Žen je mezi respondenty z původního vzorku ve druhé fázi celkem 55 (40,1%), mezi novými respondenty 57 (41,6%).

Tabulka 6 Pohlaví

| | 1. fáze | | 2. fáze | | 3. fáze | |
|---|---|---|---|---|---|---|
| | Četnost | % | Četnost | % | Četnost | % |
| Muž | 33 | 26,6 | 25 | 18,2 | 0 | 0,0 |
| Žena | 91 | 73,4 | 112 | 81,8 | 10 | 100,0 |
| **Celkem** | **124** | **100,0** | **137** | **100,0** | **10** | **100,0** |

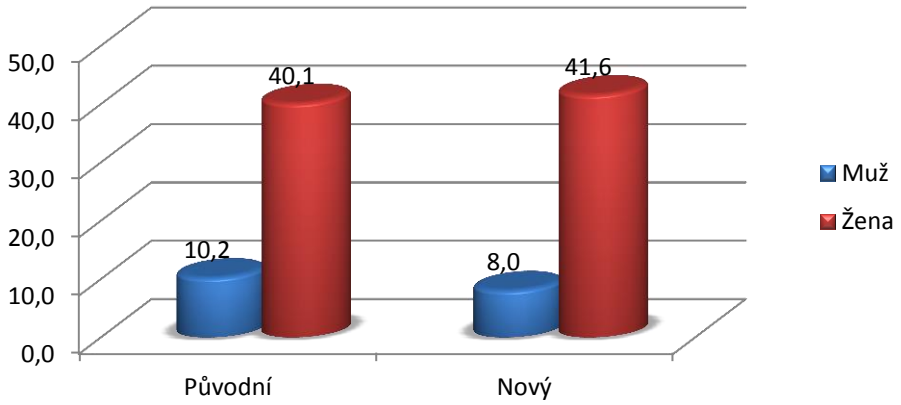
Graf 5 Pohlaví




Tabulka 7 Rozdělení pohlaví u původních a nových respondentů v 2. fázi výzkumu

| | | Četnost | % | % Celkem |
|---|---|---|---|---|
| Původní | Muž | 14 | 10,2 | 50,4 |
| | Žena | 55 | 40,1 | |
| Nový | Muž | 11 | 8,0 | 49,6 |
| | Žena | 57 | 41,6 | |
| Celkem | | 137 | 100,0 | |

Graf 6 Rozdělení pohlaví u původních a nových respondentů v 2. fázi výzkumu



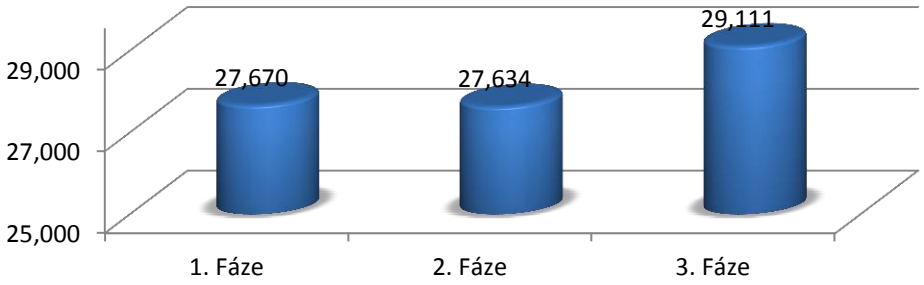
## 7.2 Věk

Nejmladšímu respondentovi v první fázi výzkumu bylo 16 let, v druhé fázi 17 let a ve třetí fázi 22. Nejstaršímu respondentovi bylo v první fázi výzkumu 63 let, v druhé fázi 62 let a ve třetí fázi 43 let. **Věkový průměr v prvních dvou fázích byl velmi podobný**, v první fázi 27,670, v druhé fázi 27,634 a **ve třetí fázi o něco vyšší** 29,111.

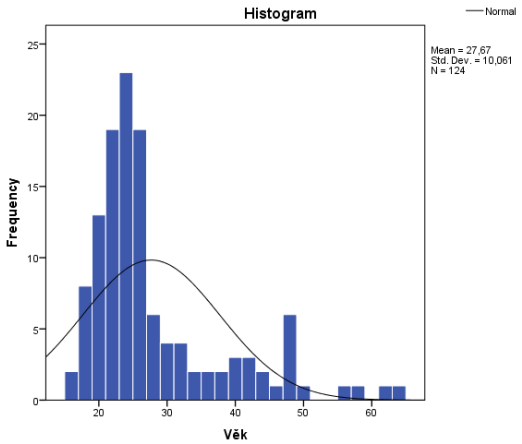
Tabulka 8 Deskriptivní statistika věku respondentů v jednotlivých fázích výzkumu

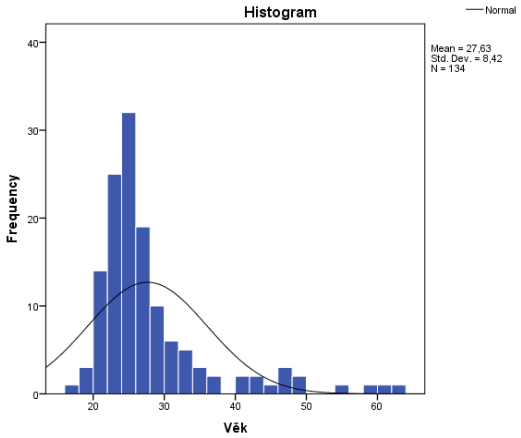
| | N | Min | Max | Průměr | Std. odchylka | Šikmost | Špičatost |
|---|---|---|---|---|---|---|---|
| 1. fáze | 124 | 16 | 63 | 27,670 | 10,061 | 1,597 | 2,093 |
| 2. fáze | 137 | 17 | 62 | 27,634 | 8,420 | 2,126 | 4,727 |
| 3. fáze | 10 | 22 | 43 | 29,111 | 7,167 | 1,030 | 0,203 |

Graf 7 Průměr věku respondentů v různých fázích výzkumu




Graf 8 Rozložení věku v první a druhé fázi výzkumu



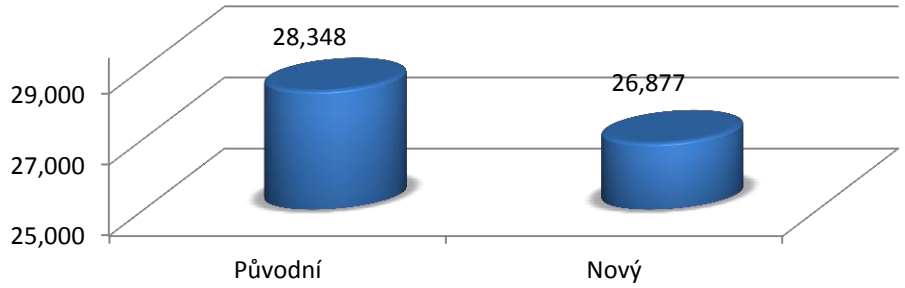



Věkový průměr původních respondentů ve 2. fázi výzkumu je o něco vyšší (28,348), než věkový průměr nových respondentů (26,877).

Tabulka 9 Deskriptivní statistika věku ve 2. fázi výzkumu - porovnání původních a nových respondentů

| | N | Min | Max | Průměr | Std. odchylka | Šikmost | Špičatost |
|---|---|---|---|---|---|---|---|
| Původní | 69 | 19 | 62 | 28,348 | 10,029 | 1,710 | 2,244 |
| Nový | 68 | 17 | 60 | 26,877 | 6,271 | 2,840 | 12,049 |

Graf 9 Srovnání průměru věku u původních a nových respondentů v 2. fázi výzkumu

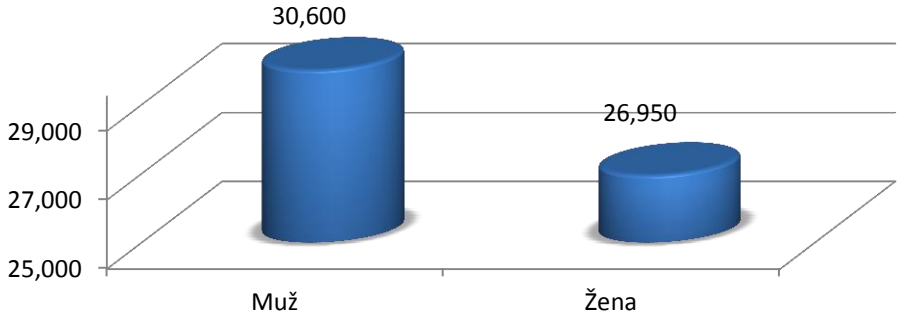



Věkový průměr mužů (30,600) ve druhé fázi výzkumu je vyšší, než věkový průměr žen (26,950).

Tabulka 10 Deskriptivní statistika věku ve 2. fázi výzkumu - porovnání mužů a žen

| | N | Min | Max | Průměr | Std. odchylka | Šikmost | Špičatost |
|---|---|---|---|---|---|---|---|
| Muž | 25 | 17 | 62 | 30,600 | 12,169 | 1,504 | 1,493 |
| Žena | 109 | 19 | 58 | 26,950 | 7,204 | 2,160 | 5,233 |

Graf 10 Srovnání průměru věku mužů a žen v 2. fázi výzkumu



# 8 Výsledky

Shrnutí výsledků je rozděleno podle jednotlivých fází výzkumu:

1. Fáze - vytvoření adaptivního eEOD testu
2. Fáze - srovnání adaptivního eEOD testu s jeho klasickou verzí
3. Fáze - srovnání s tužka-a-papír verzí testu.

## 8.1 1. fáze - vytvoření adaptivního eEOD testu

V **první fázi** bylo hlavním úkolem **získat dostatek dat** prostřednictvím online počítačové verze eEOD (škály extraverze EOD) testu. Byla získána data od **124 respondentů**. Tato data posloužila ke kalibraci položek a vytvoření počítačové adaptivní verze eEOD (popsáno v kapitole Metodologie). Byly provedeny **simulace adaptivní administrace eEOD** a jejich výsledky byly srovnávány s výsledky reálné online počítačové administrace tohoto testu.

### 8.1.1 Výsledky eEOD

V testu eEOD, zahrnujícím 48 otázek, byli respondenti hodnoceni za každou položku 1 bodem při klíčové odpovědi (značí extraverzi) nebo 0 body (značí introverzi). V **průměru respondenti získali 24,782 bodu. N**ejmenší získaný počet bodů byl 4, nejvyšší 42.

**Tabulka 11 Deskriptivní statistika HS eEOD**

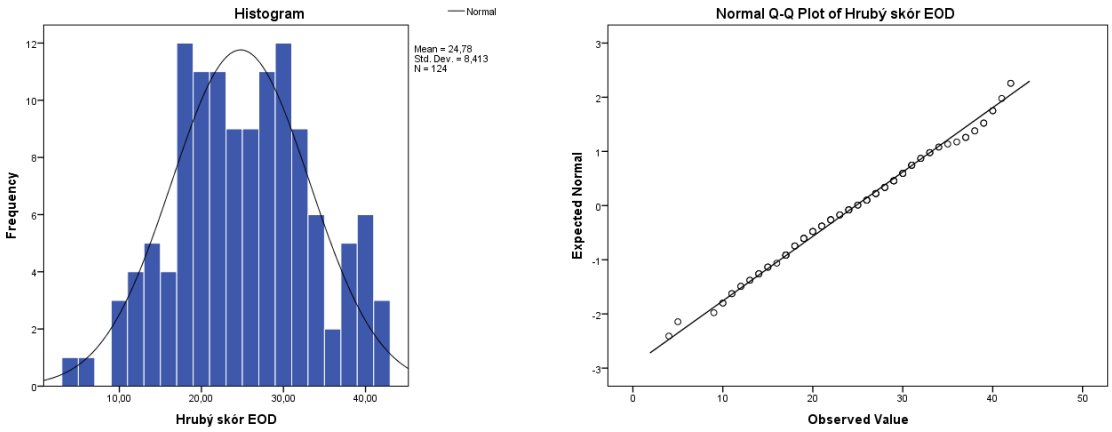
| | N | Min | Max | Průměr | Std. odchylka | Šikmost | Špičatost |
|---|---|---|---|---|---|---|---|
| **HS** | 124 | 4 | 42 | 24,782 | 8,413 | -0,002 | -0,477 |

Kolmogorov-Smirovovým i Shapiro-Wilkovým testem byla **prokázáno normální rozložení hrubého skóru eEOD** v mém vzorku. Normalita je patrná i z histogramu a Q-Q grafu. Díky tomu bude možné použít statistické testy vyžadující normalitu vzorku, jako je t-test.

Tabulka 12 Normalita HS eEOD

| | Kolmogorov-Smirnov[a] | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| HS | 0,052 | 124 | 0,200* | 0,989 | 124 | 0,401 |

Graf 11 Normalita HS eEOD




### 8.1.2 Porovnání výsledků klasického testu a simulovaného adaptivního testu

Simulace CAT eEOD byla provedena na reálných datech 124 respondentů, kteří vyplnili online eEOD. **První část simulace zjišťovala výsledky pomocí fixed-length CAT**. Test byl ukončen po použití 8, 16, 24, 32, 40 a 48 položek. Výsledky byly srovnávány s hrubým skórem eEOD.

Při srovnávání HS eEOD a simulovaného CAT při použití všech 48 položek bylo dosaženo signifikantní korelace ($r = 0,970$; $p < 0,001$).

**Zamítám $H_0$1 - Výsledky simulovaného CAT eEOD, při použití všech položek a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky eEOD z první fáze výzkumu.**

Zajímavější jsou však výsledky srovnání HS eEOD a **simulovaného CAT při sníženém počtu položek**, které demonstrují efektivitu testu, hlavní výhodu CAT. Při použití 40 položek (83,3%) bylo dosaženo korelace ($r = 0,946$; $p < 0,001$). Ještě při použití 32

položek (66,7%) dosahuje korelace nad 0,9 (r = 0,929; p < 0,001). **Při administraci pouhé poloviny testu bylo dosaženo též signifikantní korelace (r = 0,882; p < 0,001)**.

Vzhledem k tomu, že všechny korelace simulovaného CAT a HS eEOD jsou poměrně silné, i při použití minimálního počtu položek,

**zamítám $H_0$2 - Výsledky simulovaného CAT eEOD, při použití různého počtu položek a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky eEOD z první fáze výzkumu.**

Pokud uvážíme **nesplnění požadavku unidimenzionality** testu a **velmi nízkou shodu modelu s daty**, **jsou takto vysoké korelace poměrně neočekávaným výsledkem**. Pokud by se podařilo získat lepší kalibrační vzorek, byly by korelace pravděpodobně ještě vyšší.

**Tabulka 13 Korelace mezi hrubým skórem klasického eEOD a simulovaným CAT eEOD s použitím různého počtu položek**

| | | HS eEOD | 8 položek | 16 položek | 24 položek | 32 položek | 40 položek | 48 položek |
|---|---|---|---|---|---|---|---|---|
| **HS eEOD** | **r** | | **,583**** | **,649**** | **,882**** | **,929**** | **,946**** | **,970**** |
| | **Sig.** | | **,000** | **,000** | **,000** | **,000** | **,000** | **,000** |
| **8 položek** | r | ,583** | | ,429** | ,665** | ,647** | ,626** | ,566** |
| | Sig. | ,000 | | ,000 | ,000 | ,000 | ,000 | ,000 |
| **16 položek** | r | ,649** | ,429** | | ,810** | ,752** | ,716** | ,604** |
| | Sig. | ,000 | ,000 | | ,000 | ,000 | ,000 | ,000 |
| **24 položek** | r | ,882** | ,665** | ,810** | | ,973** | ,940** | ,855** |
| | Sig. | ,000 | ,000 | ,000 | | ,000 | ,000 | ,000 |
| **32 položek** | r | ,929** | ,647** | ,752** | ,973** | | ,979** | ,911** |
| | Sig. | ,000 | ,000 | ,000 | ,000 | | ,000 | ,000 |
| **40 položek** | r | ,946** | ,626** | ,716** | ,940** | ,979** | | ,946** |
| | Sig. | ,000 | ,000 | ,000 | ,000 | ,000 | | ,000 |
| **48 položek** | r | ,970** | ,566** | ,604** | ,855** | ,911** | ,946** | |
| | Sig. | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | |

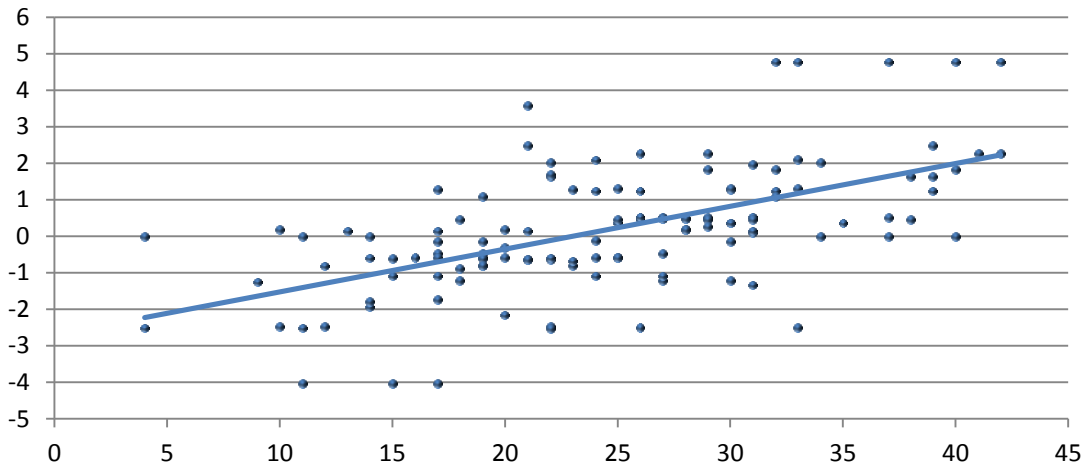
U výsledků je nutné zohlednit také **standardní chybu měření,** která ukazuje přesnost testu. Nejnižší chyby bylo podle očekávání dosaženo při použití všech položek, kde dosahovala hodnot mezi 0,293 a 0,576 s **průměrem 0,318**, což je hodnota běžně používaná jako kritérium pro ukončení CAT testu. **Při použití pouhé poloviny položek bylo průměrně dosaženo standardní chyby 0,506.** Průměrná standardní chyba při použití nejmenšího počtu

položek byla 1,230, což znamená již velkou nepřesnost. Přesto považuji dosažení minimální standardní chyby 0,683, alespoň u některých respondentů, při použití pouhých 8 položek za velmi dobrý výsledek.

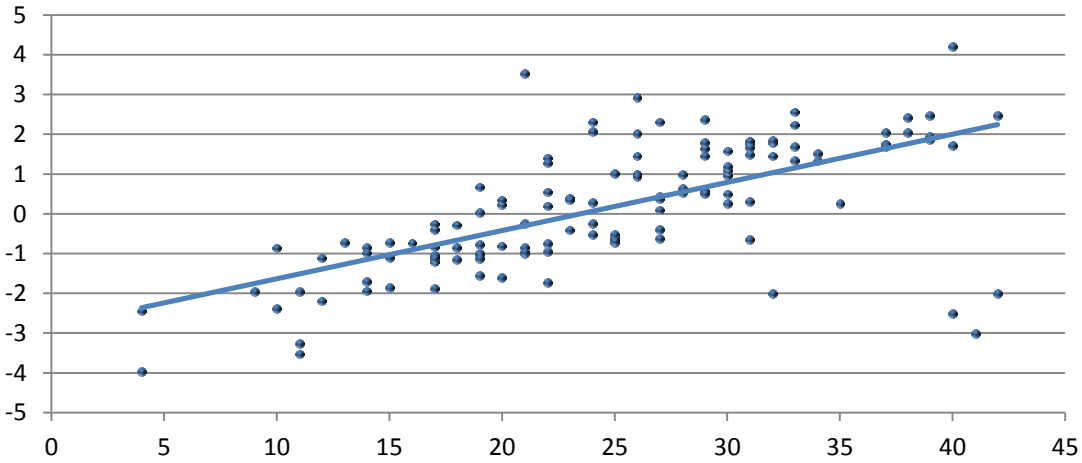
**Tabulka 14 Popisná statistika dosažené úrovně standardní chyby výsledků simulovaného CAT eEOD s použitím různého počtu položek**

| | N | Min | Max | Průměr | Std. Odchylka |
|---|---|---|---|---|---|
| **8 položek** | 124 | 0,683 | 2,812 | 1,230 | 0,493 |
| **16 položek** | 124 | 0,402 | 1,904 | 0,692 | 0,226 |
| **24 položek** | 124 | 0,358 | 1,395 | 0,506 | 0,171 |
| **32 položek** | 124 | 0,335 | 1,370 | 0,420 | 0,134 |
| **40 položek** | 124 | 0,312 | 1,366 | 0,361 | 0,112 |
| **48 položek** | 124 | 0,293 | 0,576 | 0,318 | 0,051 |

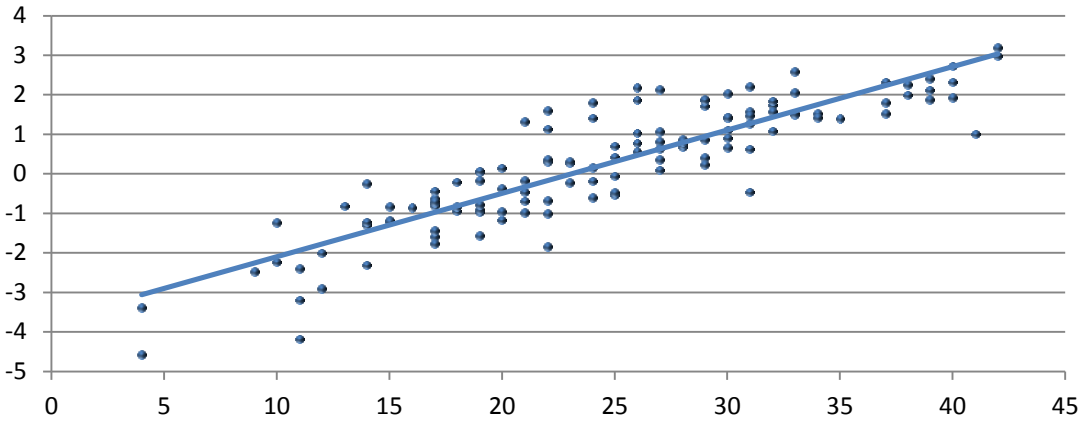
**Graf 12 Korelace HS eEOD a CAT eEOD s použitím 8 položek (r = 0,583)**



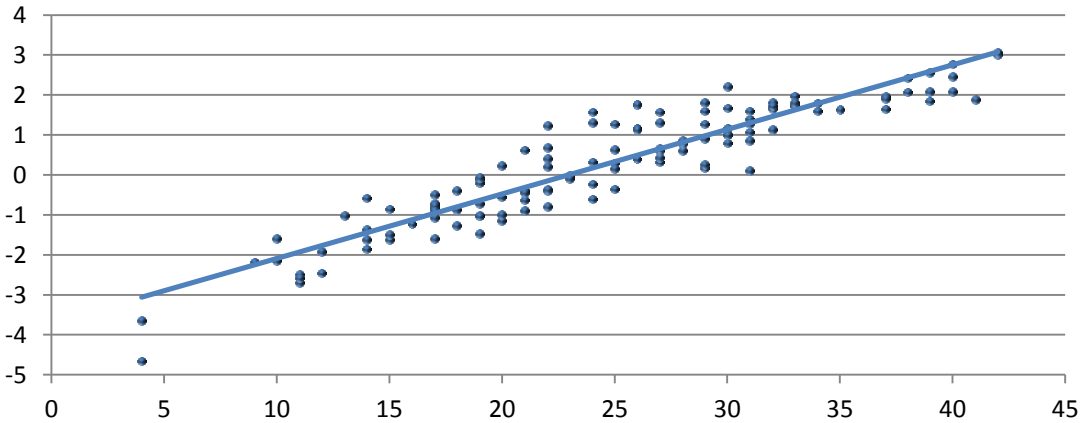
Graf 13 Korelace HS eEOD a CAT eEOD s použitím 16 položek (r = 0,649)



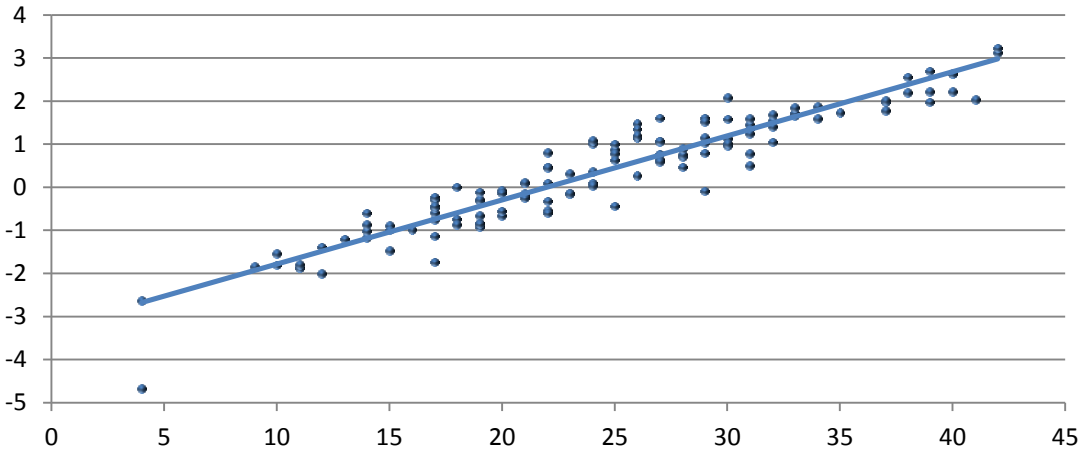

Graf 14 Korelace HS eEOD a CAT eEOD s použitím 24 položek (r = 0,882)



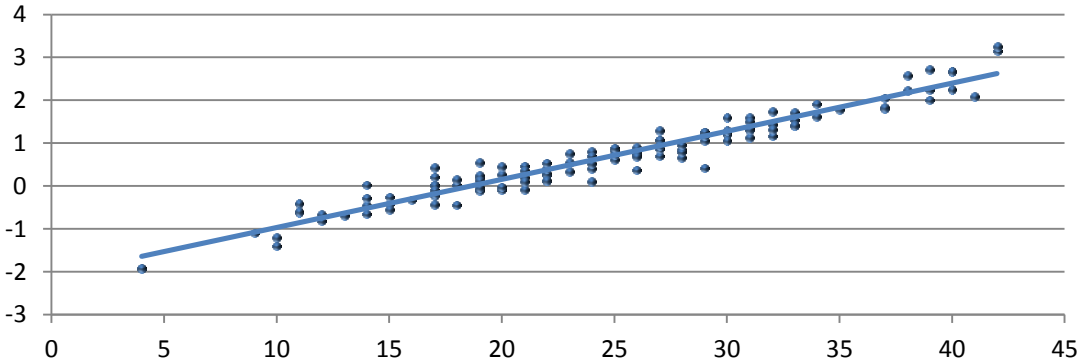

Graf 15 Korelace HS eEOD a CAT eEOD s použitím 32 položek (r = 0,929)



Graf 16 Korelace HS eEOD a CAT eEOD s použitím 40 položek (r = 0,946)




Graf 17 Korelace HS eEOD a CAT eEOD s použitím 48 položek (r = 0,970)



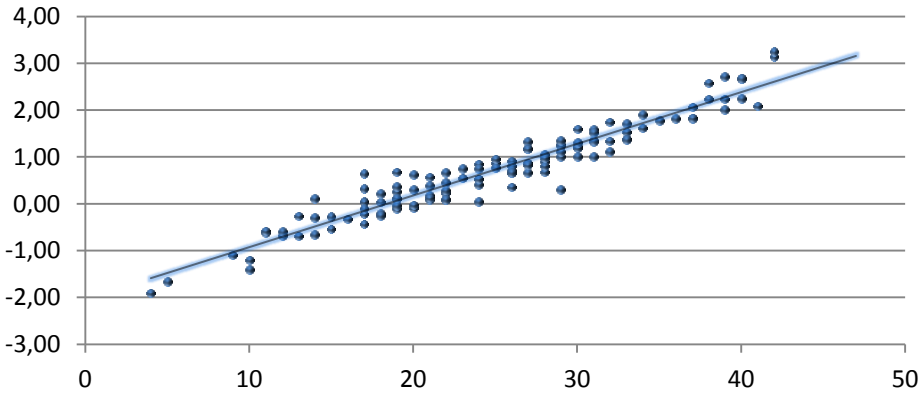

Druhá část simulací se zaměřila na **CAT ukončované na základě dosažení stanovené úrovně standardní chyby**. Hranice standardní chyby byla nastavena nejdříve na 0,3, následně na 0,4 a nakonec na 0,5. Při **srovnání s HS eEOD bylo dosaženo korelace r = 0,962 (p < 0,001).** Protože i korelace HS eEOD s CAT eEOD ukončeným na úrovni chyby 0,4 a 0,5 dosahují silných signifikantních korelací r = 0,800 a r = 0,758,

**zamítám $H_0$3 - Výsledky simulovaného CAT eEOD, při použití různých úrovní chyby pro ukončení testu a dat získaných v první fázi výzkumu, signifikantně nekorelují s výsledky eEOD z první fáze výzkumu.**

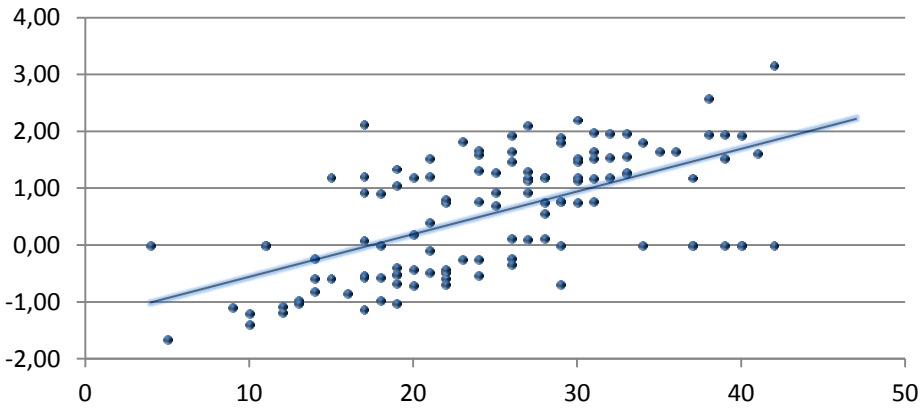
Tabulka 15 Korelace HS eEOD s CAT eEOD s ukončením na různých úrovních chyby

| | | HS eEOD | Všechny položky | SE 0,3 | SE 0,4 | SE 0,5 |
|---|---|---|---|---|---|---|
| **HS eEOD** | r | | ,970** | **,962**** | **,800**** | **,758**** |
| | Sig. | | ,000 | **,000** | **,000** | **,000** |
| **Všechny položky** | r | ,970** | | ,999** | ,847** | ,783** |
| | Sig. | ,000 | | ,000 | ,000 | ,000 |
| **SE 0,3** | r | ,962** | ,999** | | ,853** | ,786** |
| | Sig. | ,000 | ,000 | | ,000 | ,000 |
| **SE 0,4** | r | ,800** | ,847** | ,853** | | ,952** |
| | Sig. | ,000 | ,000 | ,000 | | ,000 |
| **SE 0,5** | r | ,758** | ,783** | ,786** | ,952** | |
| | Sig. | ,000 | ,000 | ,000 | ,000 | |

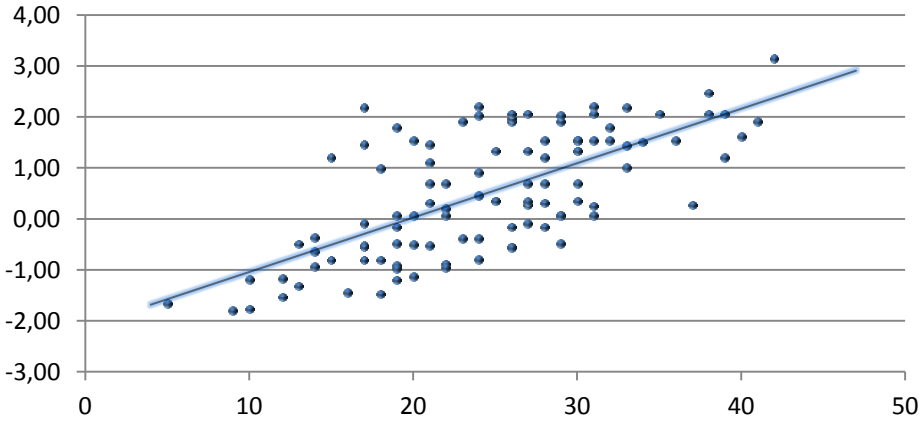
Graf 18 Korelace HS eEOD a CAT eEOD ukončeného na úrovni chyby 0,03 (r = 0,962)




Graf 19 Korelace HS eEOD a CAT eEOD ukončeného na úrovni chyby 0,04 (r = 0,800)



Graf 20 Korelace HS eEOD a CAT eEOD ukončeného na úrovni chyby 0,05 (r = 0,758)




#### 8.1.2.1 Počet použitých položek v závislosti na úrovni rysu respondenta

Při ukončení testu na základě dosažené standardní chyby je výhoda oproti ukončení na základě dosaženého počtu položek v tom, že počet **položek je přizpůsoben individuálním potřebám hodnocení u každého respondenta**. U respondentů pro které je dostatek položek v jejich úrovni obtížnosti je možné použít mnohem menší počet položek pro dosažené potřebné přesnosti měření.

**Počet použitých položek se tedy poměrně zásadně mění v závislosti na úrovni rysu respondenta**. Testy vyvinuté na základě klasické testové teorie jsou obvykle **nastaveny tak, aby měřily nejlépe v oblasti průměru, kde je očekáván nejvyšší počet respondentů**. Klasické testy jsou tvořeny hlavně položkami s obtížností kolem 0,5 (v IRT by této hodnotě odpovídala hodnota 0 parametru b). V **CAT testech je ale pravidlem vytvářet položky s co nejrůznější úrovní obtížnosti**, tedy včetně obou extrémů. Tyto položky pak umožňují mnohem efektivnější měření úrovně respondentů, a to hlavně v úrovni obou extrémů.

Při nastavení velmi nízké standardní chyby na úrovni SE = 0,3 byl **počet administrovaných položek poměrně vysoký, v průměru 84,82%**. Tento počet se samozřejmě s tolerantnějšími hodnotami standardní chyby zmenšoval, u SE = 0,4 stačilo v průměru 33,420 položek a při stále poměrně nízké úrovni chyby SE = 0,5 v průměru pouze 23,056 položek, což je **méně než polovina testu**.

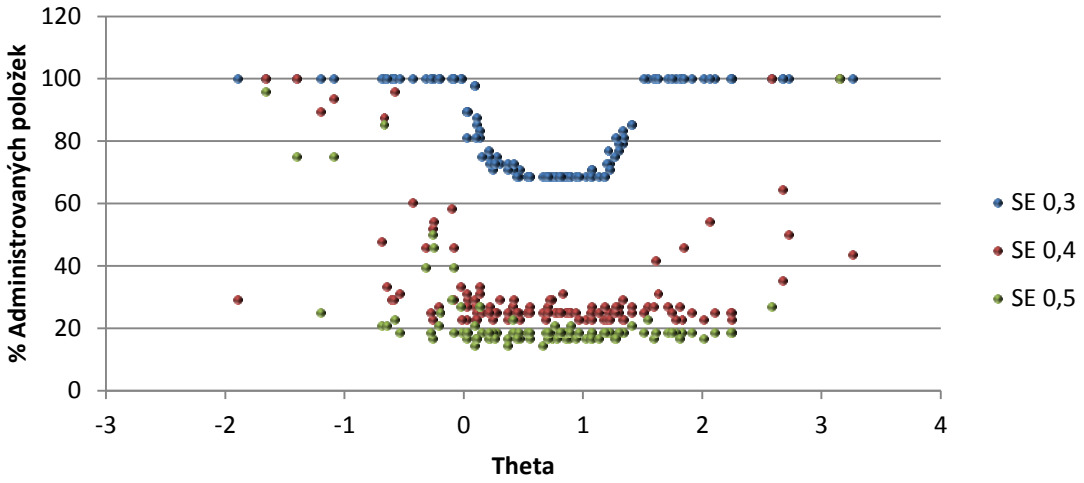
Z grafu je patrné, že hlavně u nejnižší úrovně standardní chyby SE = 0,3 bylo u respondentů **v obou extrémech použito většinou maximálního počtu položek**. Naopak u respondentů se střední úrovní rysu, přibližně od 0 do 1,5, se počet administrovaných položek

snižoval až k 68,75%. U vyšších úrovní standardní chyby není tento trend tak výrazný, přesto je stále patrný.

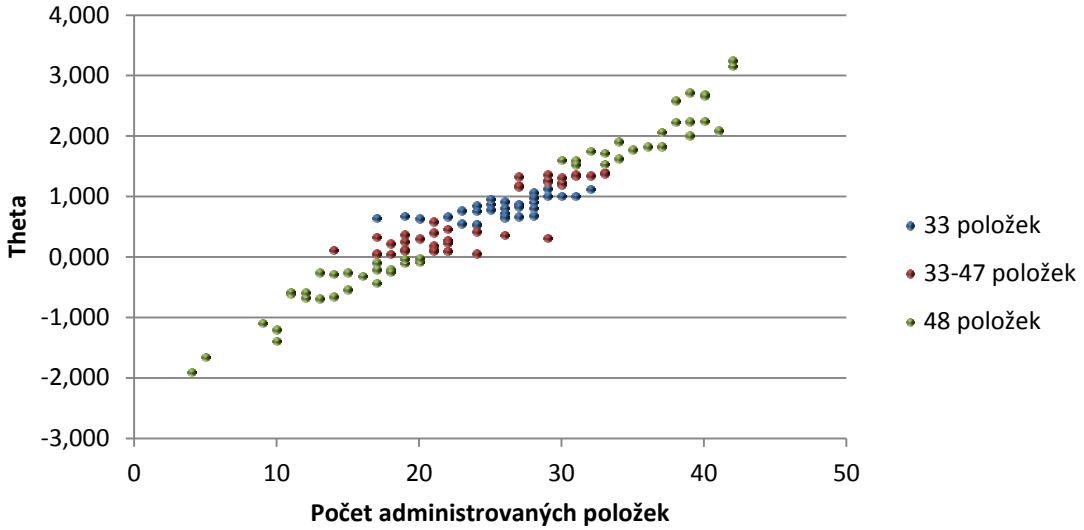
Graf 21 Počet použitých položek podle úrovně maximální chyby

| | N | Min % | Max % | Průměr % | Std. Odchylka |
|---|---|---|---|---|---|
| SE 0,3 | 120 | 68,75 | 100 | 84,82 | 13,893 |
| SE 0,4 | 120 | 22,92 | 100 | 33,42 | 18,816 |
| SE 0,5 | 105 | 14,58 | 100 | 23,05 | 15,400 |

Graf 22 Procento administrovaných položek v závislosti na úrovni rysu respondenta a úrovni maximální chyby




Graf 23 Skupiny počtu použitých položek v závislosti na úrovni rysu respondenta SE = 0,3



## 8.2 2. fáze - srovnání reálné administrace CAT eEOD testu a klasického eEOD

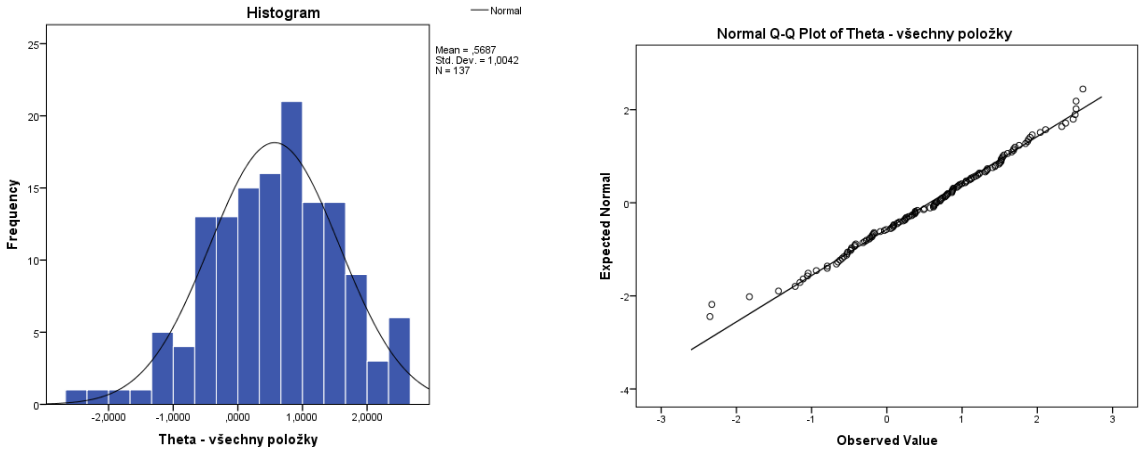
Ve 2. fázi výzkumu byla provedena **reálná administrace adaptivního eEOD testu** a to zčásti respondentům z první fáze výzkumu a zčásti novým respondentům. Spolu s tímto testem byl administrován **krátký sebehodnotící dotazník zaměřený na otevřenost a škála extraverze EPQ testu**. Jsou srovnávány výsledky těchto testů mezi sebou a zároveň s výsledky testů v první fázi výzkumu.

Výsledky skóru CAT eEDO vykazují **normální rozložení**, což je předpoklad pro použití některých statistických testů.

**Tabulka 16 Normalita CAT eEOD**

| | Kolmogorov-Smirnova | | | Shapiro-Wilk | | |
|---|---|---|---|---|---|---|
| | Statistic | df | Sig. | Statistic | df | Sig. |
| **CAT eEOD** | 0,060 | 137 | 0,200* | 0,988 | 137 | 0,289 |

**Graf 24 Normalita CAT eEOD**




### 8.2.1 Porovnání simulace s výsledky reálného CAT testování všech položek

**V simulované administraci CAT eEOD jsou výsledné skóry o něco vyšší, než u reálné administrace**. U simulované administrace bylo minimum -1,930, zatímco u reálné administrace -2,352. Skóry v simulované administraci dosahovaly až 3,260, ale u reálné administrace jen 2,606. **Odlišnosti však mohou být připisovány odlišným úrovním rysu u respondentů v první a druhé fázi výzkumu**. Respondenti v druhé fázi výzkumu tak byli

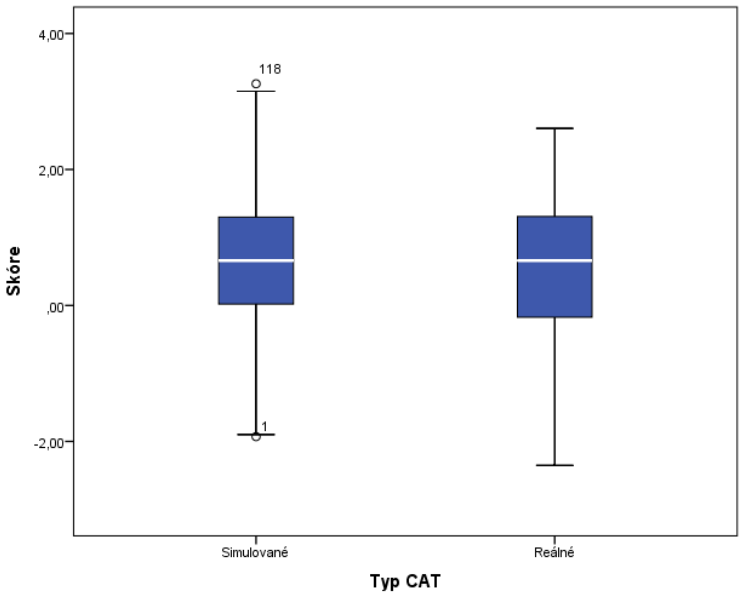
pravděpodobně více introvertní. **Průměr výsledných skórů simulované a reálné administrace CAT eEOD se ale statisticky významně neliší (t = 0,779; df = 260; p = 0,437)**, takže

**zamítám $H_0$4 - Průměry výsledků reálné administrace CAT eEOD, při použití všech položek, se signifikantně neliší od průměrů výsledků simulovaného CAT eEOD.**

Tabulka 17 Porovnání výsledků simulovaného a reálného CAT eEOD testu

| | N | Min | Max | Průměr | Std. Odchylka |
|---|---|---|---|---|---|
| Simulované | 125 | -1,930 | 3,260 | 0,664 | 0,981 |
| Reálné | 137 | -2,352 | 2,606 | 0,569 | 1,004 |

Graf 25 Boxplot porovnávající výsledky simulovaného a reálného CAT eEOD testu




### 8.2.2 Hodnocení efektivity CAT eEOD

Nejvýraznější výhodou CAT je jeho efektivita, tedy **schopnost zachovat přesnost měření i při sníženém počtu položek**. Tato přesnost je obvykle určována informačním přínosem testu nebo také standardní chybou. Při použití všech položek CAT eEOD test bylo

dosaženo v průměru chyby 0,319, což lze považovat za velmi dobrý výsledek. Nejnižší dosažená chyba byla 0,293 a nejvyšší 0,623. Se snižujícím se počtem položek velikost chyby pravděpodobně vzroste. Při použití **pouhé poloviny položek** se stále průměr standardní chyby drží velmi nízko, na 0,391. Až při použití pouze **8 položek** (tedy asi 17% položek) chyba vzrostla na 0,911, což značí, že test je již poměrně nepřesný. Přesto u některých respondentů bylo i při takto **malém počtu položek dosaženo standardní chyby 0,502**, tedy poměrně malé chyby a tedy dostatečné přesnosti testu. Protože zkrácení testu o celou polovinu při zachování poměrně dobré přesnosti testu, a to s ohledem na to, že nebylo dosaženo podmínek pro použití IRT modelů,

**zamítám $H_0$5 – CAT eEOD test bude stejně efektivní, jako klasický eEOD test (tzn. bude potřeba použít všechny nebo většinu položek k dosažení dostatečné přesnosti testu)**.

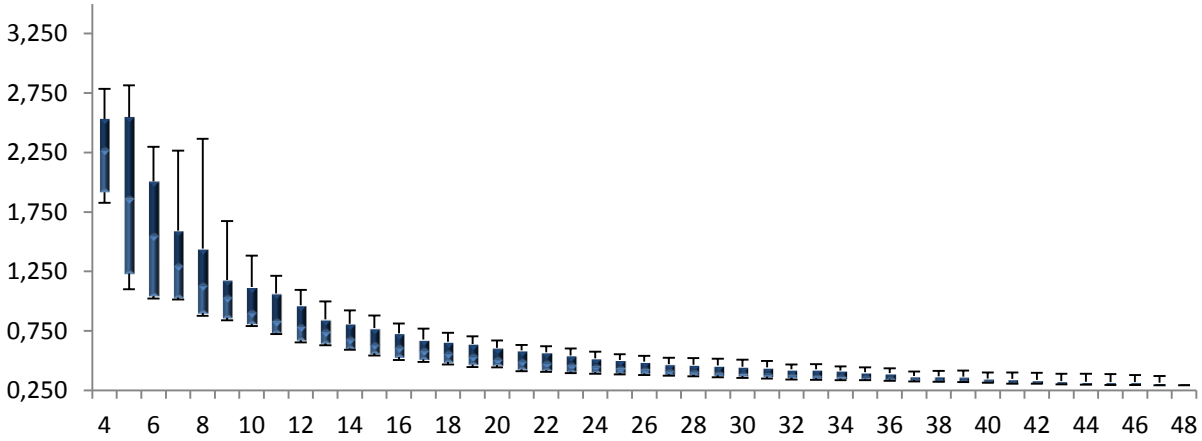
Tabulka 18 Deskriptivní statistika CAT eEOD - Theta

| | N | Min | Max | Průměr | Std. Odchylka |
|---|---|---|---|---|---|
| Simulace CAT eEOD | 125 | -1,930 | 3,260 | 0,664 | 0,981 |
| Theta - 48 položek | 137 | -2,352 | 2,606 | 0,569 | 1,004 |
| Theta - 8 položek | 81 | -3,136 | 3,238 | 0,236 | 1,610 |
| Theta - 16 položek | 127 | -2,635 | 3,678 | 0,529 | 1,032 |
| Theta - 24 položek | 137 | -2,317 | 2,883 | 0,554 | 1,028 |
| Theta - 32 položek | 137 | -2,501 | 2,618 | 0,484 | 1,044 |
| Theta - 40 položek | 137 | -2,516 | 2,504 | 0,509 | 1,043 |

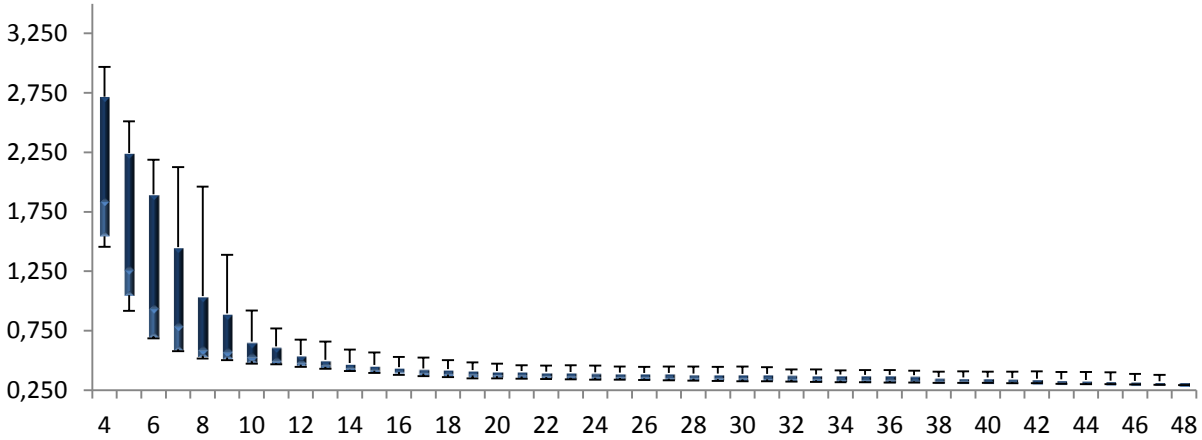
Tabulka 19 Deskriptivní statistika CAT eEOD - SE

| | N | Min | Max | Průměr | Std. odchylka |
|---|---|---|---|---|---|
| SE - 48 položek | 137 | 0,293 | 0,632 | 0,319 | 0,051 |
| SE - 8 položek | 81 | 0,502 | 2,529 | 0,911 | 0,576 |
| SE - 16 položek | 127 | 0,349 | 1,329 | 0,444 | 0,115 |
| SE - 24 položek | 137 | 0,322 | 0,852 | 0,391 | 0,072 |
| SE - 32 položek | 137 | 0,316 | 0,834 | 0,373 | 0,071 |
| SE - 40 položek | 137 | 0,301 | 0,806 | 0,351 | 0,069 |

Graf 26 Vývoj standardní chyby u simulace CAT eEOD




Graf 27 Vývoj standardní chyby u reálné administrace CAT eEOD




### 8.2.3 Porovnání simulace s výsledky reálného testování části položek

Při porovnání výsledků simulace a reálné administrace **bylo dosaženo signifikantních korelací**, které jsou uvedeny v tabulce níže. Simulovaný test využívající k hodnocení všech dostupných položek **koreluje se zkráceným CAT testem o osmi položkách r = 0,598**. Korelace testu CAT eEOD při použití všech položek koreluje s jeho zkrácenou formou o 8 položkách r = 0,718. **Při použití poloviny položek koreluje simulovaný test r = 0,786 a celý CAT eEOD test r = 0,940**. Vzhledem k těmto výsledkům

**zamítám $H_0$6 - Výsledky reálné administrace CAT eEOD, při použití sníženého počtu položek, signifikantně nekorelují s výsledky simulovaného CAT eEOD.**

Tabulka 20 Korelace mezi simulací CAT eEOD, CAT eEOD s použitím všech položek a CAT eEOD s použitím různého počtu položek

| | | Simulace CAT eEOD | CAT eEOD 48 položek | CAT eEOD 8 položek | CAT eEOD 16 položek | CAT eEOD 24 položek | CAT eEOD 32 položek | CAT eEOD 40 položek |
|---|---|---|---|---|---|---|---|---|
| Simulace CAT eEOD | r | | ,844** | ,598** | ,715** | ,786** | ,801** | ,813** |
| | Sig. | | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 |
| CAT eEOD 48 položek | r | ,844** | | ,718** | ,887** | ,940** | ,962** | ,983** |
| | Sig. | ,000 | | ,000 | ,000 | ,000 | ,000 | ,000 |
| CAT eEOD 8 položek | r | ,598** | ,718** | | ,813** | ,772** | ,735** | ,733** |
| | Sig. | ,000 | ,000 | | ,000 | ,000 | ,000 | ,000 |
| CAT eEOD 16 položek | r | ,715** | ,887** | ,813** | | ,957** | ,933** | ,912** |
| | Sig. | ,000 | ,000 | ,000 | | ,000 | ,000 | ,000 |
| CAT eEOD 24 položek | r | ,786** | ,940** | ,772** | ,957** | | ,987** | ,968** |
| | Sig. | ,000 | ,000 | ,000 | ,000 | | ,000 | ,000 |
| CAT eEOD 32 položek | r | ,801** | ,962** | ,735** | ,933** | ,987** | | ,985** |
| | Sig. | ,000 | ,000 | ,000 | ,000 | ,000 | | ,000 |
| CAT eEOD 40 položek | r | ,813** | ,983** | ,733** | ,912** | ,968** | ,985** | |
| | Sig. | ,000 | ,000 | ,000 | ,000 | ,000 | ,000 | |

### 8.2.4 Porovnání HS eEOD, simulovaného CAT eEOD a reálného CAT eEOD

Tabulka 21 Deskriptivní statistika skórů HS eEOD, simulovaného CAT eEOD a reálného CAT eEOD

| | N | Minimum | Maximum | Mean | Std. Deviation |
|---|---|---|---|---|---|
| Z-skór eEOD | 69 | 2,083 | 1,875 | 0,045 | 0,868 |
| Simulace CAT eEOD | 69 | -1,900 | 3,260 | 0,625 | 0,924 |
| Reálné CAT eEOD | 69 | -2,352 | 2,606 | 0,615 | 1,073 |

Tabulka 22 8.2.4 Porovnání HS eEOD, simulovaného CAT eEOD a reálného CAT eEOD

| | | HS eEOD | Simulace CAT eEOD | Reálné CAT eEOD |
|---|---|---|---|---|
| Z-skór eEOD | r | | ,964** | ,866** |
| | Sig. | | ,000 | ,000 |

| | | | | |
|---|---|---|---|---|
| **Simulace CAT eEOD** | r | ,964** | | ,844** |
| | Sig. | 0,000 | | ,000 |
| **Reálné CAT eEOD** | r | ,866** | ,844** | |
| | Sig. | ,000 | ,000 | |

### 8.2.5 Porovnání reálného testování CAT eEOD s výsledky eEPQ a sebehodnocení

Spolu s počítačovou adaptivní verzi eEOD byla administrována také **škála extraverze testu EPQ a krátký sebehodnotící dotazník** zaměřený na hodnocení otevřenosti. Účelem bylo **srovnání výsledků těchto testů s úrovní extraverze zjištěné CAT eEOD** jako jednoduché porovnání validity CAT administrace.

**Škála extraverze dotazníku eEPQ** obsahuje 23 položek s možnosti odpovědět ano nebo ne. Položky jsou hodnoceny 1 bodem při klíčové odpovědi, tedy odpovědi naznačující extraverzi. Respondenti dosáhli průměrně 12,070 bodů, přičemž nejnižší skóre bylo 0 bodů a nejvyšší 23. Výsledek testu eEPQ silně **koreluje s výsledkem CAT eEOD** ($r = 0{,}875$; $p < 0{,}001$), takže

**zamítám $H_0$7 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky EPQ.**

**Sebehodnocení** zahrnovalo 10 dvojic přídavných jmen popisujících na jedné straně extraverzi a na straně druhé introverzi. Respondent měl na 7 bodové škále určit, které tvrzení ho vystihuje lépe. Minimálně mohl respondent získat celkem 10 bodů (introverze), maximálně 70 (extraverze). Respondenti dosáhli **průměrně 42,420 bodů**, nejmenší dosažený počet bodů byl 18, nejvyšší byl 64. Protože výsledek sebehodnocení středně silně pozitivně koreluje s výsledkem CAT eEOD ($r = 0{,}802$; $p < 0{,}001$),

**zamítám $H_0$8 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky sebehodnocení.**

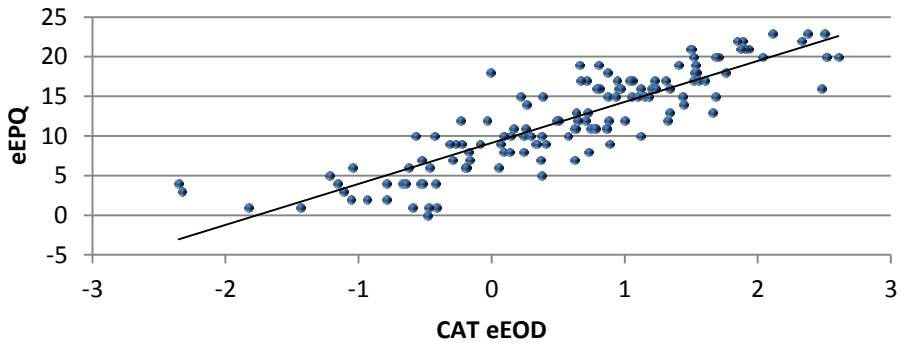
Tabulka 23 Deskriptivní statistika výsledků CAT eEOD a dalších testů eEPQ a sebehodnocení

| | N | Min | Max | Průměr | Std. Odchylka |
|---|---|---|---|---|---|
| CAT eEOD | 137 | -2,352 | 2,606 | 0,569 | 1,004 |
| eEPQ | 137 | 0 | 23 | 12,070 | 5,938 |
| Sebehodnocení | 137 | 18 | 64 | 42,420 | 9,645 |

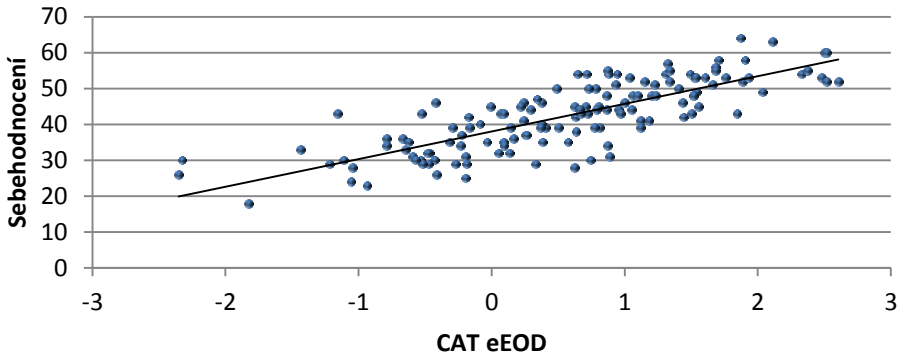
Tabulka 24 Korelace výsledků CAT eEOD s výsledky dalších testů (eEPQ a sebehodnocení)

| | | CAT eEOD | HS eEOD | eEPQ | Sebehodnocení |
|---|---|---|---|---|---|
| CAT eEOD | r | | ,866** | ,875** | ,802** |
| | Sig. | | 0,000 | 0,000 | 0,000 |
| HS eEOD | r | ,866** | | ,876** | ,759** |
| | Sig. | 0,000 | | 0,000 | 0,000 |
| eEPQ | r | ,875** | ,876** | | ,784** |
| | Sig. | 0,000 | 0,000 | | 0,000 |
| Sebehodnocení | r | ,802** | ,759** | ,784** | |
| | Sig. | 0,000 | 0,000 | 0,000 | |

Graf 28 Korelace výsledků CAT eEOD s výsledky eEPQ




Graf 29 Korelace výsledků CAT eEOD s výsledky sebehodnocení



## 8.3 3. fáze - srovnání s tužka-a-papír verzí testu

Třetí fáze výzkumu pro úplnost srovnává výsledky testů z druhé fáze výzkumu **s administrací klasického tužka-a-papír eEOD**. Hlavním cílem je zde **zjišťování možného vlivu počítačového prostředí v administraci** eEOD v mé studii. Byl použit **pouze malý vzorek 10 respondentů**, kteří podstoupili klasický tužka-a-papír eEOD test. V tomto testu získali respondenti průměrně 27,900 bodů s minimem 10 bodů a maximem 46 bodů z celkových 48 možných. Zároveň tito respondenti získali v druhé fázi výzkumu v adaptivním testu průměrně skóre 0,512 s minimem -0,793 a maximem 2,515. Zdá se, že **korelace těchto testů je poměrně silná**, se signifikancí na hladině významnosti 0,01. S vědomím, že zde může mít jistý vliv malý počet respondentů,

**zamítám $H_0$9 - Výsledky reálné administrace CAT eEOD, při použití všech položek, signifikantně nekorelují s výsledky klasické tužka-a-papír administrace.**

Tabulka 25 Deskriptivní statistika výsledků CAT eEOD testu a P&P eEOD testu a výsledky korelace

| | N | Min | Max | Průměr | Std. odchylka | r | Sig. |
|---|---|---|---|---|---|---|---|
| CAT eEOD | 10 | -0,793 | 2,515 | 0,512 | 0,948 | 0,876* | 0,001 |
| P&P eEOD | 10 | 10 | 46 | 27,900 | 11,714 | | |

# 9 Diskuse

Cílem této diplomové práce bylo **srovnání různých forem administrace dotazníku eEOD a jednoduché ověření validity CAT eEOD.** Bylo provedeno porovnání počítačové a tužka-a-papír administrace, CAT simulací a real-life CAT administrací. Pro ověření validity byl použit test eEPQ a krátké sebehodnocení.

U testu eEOD nebyla potvrzena unidimenzionalita ani dostatečná shoda modelu s daty. Přesto bylo **dosaženo silných signifikantních korelací při srovnávání klasického eEOD (hrubé skóry) a simulovaného CAT eEOD (theta)**, a to při simulaci celého CAT testu i při použití menšího počtu položek nebo různých úrovní chyby měření. Z hlediska efektivity testu bylo **hlavně u respondentů z oblasti průměru dosaženo výrazného zkrácení testu** až na polovinu, při zachování vysoké přesnosti měření. U respondentů s extrémními úrovněmi rysu byly dosažené výsledky horší, což je dáno psychometrickými charakteristikami CTT testů. Jako výhodnější bylo vyhodnoceno ukončovací kritérium na základě dosažené úrovně přesnosti.

**Silných korelací bylo dosaženo také při srovnávání klasického a simulovaného CAT testu s real-life administrovaným CAT testem**. Tento reálně administrovaný CAT test také středně silně koreloval s výsledky dalších testů eEPQ a sebehodnocení. Lepších výsledků bylo ale dosaženo při korelaci eEPQ a sebehodnocení s klasickým a simulovaným CAT.

Byla zjištěna **silná korelace mezi real-life CAT eEOD testem a tužka-a-papír formou** klasického eEOD testu. Pro toto srovnání byl použit velice malý vzorek čítající pouze 10 respondentů.

## 9.1 Metodologie a použité metody

Faktorovou analýzou **nebyla prokázána unidimenzionalita** testu eEOD a nebylo dosaženo ani uspokojivé shody modelu s daty při použití 1PL, 2PL ani 3PL modelu. **Nejlepší shody bylo dosaženo při použití 2PL modelu.** Příčinou může být příliš malý vzorek nebo velký rozptyl dat, u kterého nemůže být shoda dosažena pomocí žádného modelu (Baker, 2001). U malých vzorků se také někdy objevuje falešná negativita (Jelínek, Květon, & Vobořil, 2011a). V tomto případě tedy lze konstatovat, že **jednoduché IRT modely nejsou pravděpodobně zcela vhodné pro eEOD,** s přihlédnutím k možné falešné negativitě způsobené velikostí vzorku.

Výzkum byl rozdělen do **tří fází.** V první fázi byly provedeny simulace nově vytvořeného CAT eEOD testu, v druhé fázi byla provedena reálná administrace CAT eEOD testu spolu s testy eEPQ a sebehodnocením a ve třetí fázi proběhla administrace tužka-a-papír verze eEOD. **Časový odstup** druhé fáze byl 7-8 měsíců, u třetí fáze pouze 1 měsíc. Bylo by vhodnější, aby časový odstup byl stejný.

Tento výzkum by bylo jistě lepší **provést formou experimentu**, kdy by byli respondenti náhodně přiřazeni do skupin a podle toho jim byl test administrován v klasické počítačové, papírové nebo CAT počítačové verzi. I v tomto případě by ale bylo nutné provést nejdříve kalibraci testu, a to ideálně na úplně jiném a dostatečně velkém vzorku respondentů. Možností by bylo také provést u všech respondentů všechny typy administrace. Zde by ale mohlo dojít k efektu učení, vlivu únavy a dalších nežádoucích proměnných.

Při CAT testování by také mohlo být přínosné rozdělit respondenty do několika skupin a zadat jim **testy s odlišnými ukončovacími kritérii**: odlišný počet položek u fixed-length testu, nastavení různých úrovní chyby či využití časového limitu. V tomto výzkumu byl, ne zcela vhodně, **použit fixed-length test s využitím všech položek**, takže nelze zcela jednoznačně odhadnout vliv zkrácení testu na respondenta.

Více zohledněno by mělo být také **důsledné zajištění stejných podmínek,** a to hlavně u počítačových testů. V tomto výzkumu nebyly stejné podmínky nijak zajišťovány. Respondenti vyplňovali testy online na svých počítačích. Při administraci mohlo docházet například ke **zpožďování zobrazení položek** z důvodu vytíženosti serveru. Podle časových záznamů také respondenti poměrně **často vyplňování testu přerušili**, přesto že byli žádáni, aby to nedělali. Mohlo se lišit zobrazení zadání, a to například z důvodu **různého rozlišení a nastavení barevnosti** obrazovek. Respondenti také pravděpodobně používali různý hardware.

**U osobnostních testů zřejmě odlišné podmínky zobrazení nemají až tak zásadní dopad**, takže lze očekávat, že nedošlo k žádnému zásadnímu zkreslení. Přesto by tato problematika měla být oblastí dalšího zkoumání a ověřování. U výkonnostních testů již byl vliv těchto testovacích podmínek prokázán např. (Chen, White, McCloskey, Soroui, & Chun, 2011), (Květon & Klimusová, 2002), u testu IST (Květon, Martin, Vobořil, & Klimusová, 2003) a také u Bourdonova testu (Květon, Jelínek, Vobořil, & Klimusová, 2007).

## 9.2 Vzorek

V první fázi výzkumu tvořilo vzorek **124 respondentů**, jejichž data sloužila hlavně ke **kalibraci testu eEOD**. Všichni tito respondenti souhlasili s účastí na druhé fázi výzkumu, v rámci níž měli podstoupit nově vzniklý CAT test pro porovnání. Bohužel, **druhé části výzkumu se zúčastnila jen polovina těchto původních respondentů**. Spolu s těmito 69 původními respondenty se druhé fáze výzkumu zúčastnilo ještě 68 nových respondentů. Celkový počet respondentů v druhé fázi byl tedy 137. Vzorek ve třetí fázi tvořilo pouze 10 respondentů náhodně vybraných z nových respondentů, kteří se zúčastnili druhé fáze.

**Ve vzorku není vyrovnaný počet mužů a žen**. V prvních dvou fázích tvořily většinu respondentů ženy, ve třetí fázi dokonce celý vzorek. **Průměrný věk respondentů** v první fázi byl 27,670 a podobně v druhé fázi 27,634. Průměrný věk respondentů ve třetí fázi byl o něco vyšší 29,111. V druhé fázi výzkumu byl věkový průměr vyšší u původních respondentů ve srovnání s novými (28,348) a také vyšší u mužů (30,600). Věk má v první a druhé fázi normální rozložení.

Ačkoli kalibraci testu není nezbytně nutné provádět na reprezentativním souboru, bylo by **vhodné získat mnohem větší vzorek respondentů,** než je tento, aby bylo možné spolehlivě určit hodnoty jednotlivých parametrů. V podobných výzkumech se mluví obvykle o stovkách až tisících respondentů. Například Olea a kol. (2011) provedli kalibraci na 1 576 respondentech a Vogels a kol. (2011) dokonce na 2 041 respondentech. I pro další analýzy by bylo přínosné získání většího počtu respondentů. **Problém s mortalitou respondentů** by mohlo vyřešit okamžité testování na místě bez časového rozestupu a další úpravy v designu výzkumu.

## 9.3 Výsledky

Při porovnání klasického eEOD a simulovaného CAT eEOD bylo dosaženo uspokojivých výsledků. Při použití všech položek spolu **testy korelovaly až na úrovni r = 0,970, přičemž dosažená průměrná chyba byla SE = 0,318**. Zde by bylo možné namítat, že testy by při použití všech položek měly korelovat r = 1,000, to je ale nepravděpodobné, protože CAT používá odlišné postupy a v tomto případě navíc **nebylo dosaženo dostatečné shody modelu s daty ani potvrzení unidimenzionality** testu. Při možnosti kalibrovat položky s pomocí většího vzorku respondentů by bylo pravděpodobně

dosaženo vyšší shody. I **výsledky simulovaného CAT eEOD při použití menšího počtu položek nebo nastavení požadované úrovně chyby silně korelovaly s výsledky klasického eEOD**. Při použití pouhé poloviny testu bylo dosaženo korelace 0,882. Zároveň při nastavení velmi nízké úrovně chyby 0,3 stačilo k dosažení korelace r = 0,962 při použití 85% položek.

**Větší efektivity testu bylo dosaženo při nastavení ukončovacího kritéria ve formě stanovené přesnosti testu, tedy standardní chyby**. Počet administrovaných položek se totiž výrazně zvyšoval u respondentů s extrémní úrovní rysu, a naopak u průměrných respondentů stačilo velmi malé procento položek. CAT eEOD je totiž test, který vznikl pouze z položek konvenčního eEOD testu, a konvenční testy většinou výrazně ztrácí na přesnosti v extrémních pólech distribuce skóru (Parshall, 2002). Je to proto, že **většina položek v těchto testech se kvůli normálnímu rozložení pohybuje kolem průměru**, kde se předpokládá nejvyšší podíl respondentů (Wainer & Dorans, 2000). V klasických testech není možné zařadit větší množství obtížnějších a jednodušších položek, protože by to neúměrně prodloužilo test. Tento problém ale u počítačových adaptivních testů nehrozí, protože jsou použity právě jen ty položky, které nejlépe odpovídají úrovni respondenta. Pokud by tedy měl být eEOD test (nebo třeba i celý EOD test) převeden do CAT podoby, bylo by potřeba **dovyvinout dostatečné množství položek hlavně v oblasti nadprůměru a podprůměru**.

Výsledky této práce jsou podobné výsledkům dalších studií, zabývajících se efektivitou CAT testů. V podobném výzkumu používajícím též EOD test, pouze se škálou neuroticismu místo extraverze, při standardní chybě SE = 0,5 téměř polovina respondentů zodpověděla jen polovinu položek (Květoň, Jelínek, Denglerová, & Vobořil, 2008). S**tejně jako v mé práci zde byla skupina respondentů s extrémními úrovněmi rysu, u kterých nebylo dostatečné přesnosti dosaženo ani při zodpovězení většiny nebo dokonce všech položek.**

Dále například v testu MMPI pro adolescenty bylo ušetřeno přibližně 10,7-26,4% položek (Forbey, Handel, & Ben-Porath, 2000), u testu MMPI-2 dokonce až 20-35% položek (Forbey & Ben-Porath, 2007) a u testu NEO PI-R až 50% při korelaci r = 0,92 (Reise & Henson, 2000). Z Eysenckových dotazníků pak například MPQ, kde bylo **dosaženo redukce až o 50% při silnější korelaci (r = 0,970)** (Waller & Reise, 1989) **než v této práci (r = 0,882)**. Efektivnějších a přesnějších výsledků v porovnání s klasickými testy dosáhli také Vispoel a kol. (1994). Zvláště tedy u testování časově náročnými bateriemi testů (úspora

obvykle až 40-60%) (Linden & Glas, 2010) a při testování v omezeném čase je CAT dobrou volbou (Parshall, 2002).

U reálně administrovaného CAT byly výsledné skóry o něco vyšší, než u simulací. Pravděpodobně je to ale jen důsledek odlišných vzorků. Přesto se **průměry výsledků statisticky významně neliší**. I zde byla prokázána větší efektivita CAT testu oproti testu klasickému. Po administraci první poloviny testu bylo obvykle dosaženo průměrné chyby 0,391, což je stále přijatelná hodnota. Výsledky jsou zde podobné jako u simulované administrace CAT a korelují spolu r = 0,844. S klasickým testem pak lépe koreluje simulovaný test (r = 0,968) než reálně administrovaný test (r = 0,866). Při použití všech položek bylo u simulovaného a reálného testu dosaženo v průměru prakticky stejné standardní chyby (simulovaný 0,318; reálný 0,319). Bohužel zde **nebylo provedeno další podrobnější srovnání počtu položek při dosažení různých úrovní chyb a u respondentů s různými úrovněmi rysu mezi simulovaným a reálně administrovaným testem**. Tímto srovnáním by mohly být nalezeny další zajímavé odlišnosti těchto administrací.

**Výsledky CAT eEOD pouze středně silně korelují s výsledky eEPQ (r = 0,547) a sebehodnocením (r = 0,511)**. Nakonec proběhlo srovnání výsledků CAT eEOD s výsledky klasické tužka-a-papír administrace. Zde bylo dosaženo signifikantní korelace r = 0,876. I podle Chen a kol. (2011) může počítačové testování sice ovlivnit výsledky testů, ale to se týká hlavně testů výkonových, u osobnostních testů není tento vliv tak patrný, jak vyplývá i z mého výzkumu.

## 9.4 CAT postupy

Ačkoli byla efektivita testu prokázána i v mém výzkumu, **mnohem lepších výsledků by bylo pravděpodobně dosaženo při vytvoření kvalitnějšího CAT testu**. Příkladem takového zlepšení by bylo například vhodnější kritérium pro výběr položek, které **zvyšuje bezpečnost a tím i přesnost testu** (Wainer & Dorans, 2000). To ale zřejmě nemá vliv u fixed-length testů s použitím celé položkové banky. V tomto výzkumu byla použita metoda Maximum Fisher information hlavně pro její jednoduchost. Tato metoda má ale **špatné využití položek v položkové bance a nezohledňuje maximální expozici položek**. Přestože v tomto výzkumu nebylo hlavně z důvodu velmi malé položkové banky (obvykle je potřeba až několik tisíc položek) možné využít kontrolu expozice položek, v dalším výzkumu by bylo vhodné tuto kontrolu zařadit. Alespoň malá úprava by mohla být provedena podle Barrada a

kol. (2008), který používá též Maximum Fisher information, ale **zahrnuje při výběru položek prvky náhodnosti**. První položky by bylo vhodnější volit z položek s úrovní obtížnosti kolem nuly, například od -0,5 do 0,5, aby nebyly vybírány stále ty samé položky.

**Odhad úrovně rysu by měl být hlavně u prvních položek nahrazen Bayesovskými metodami** MAP nebo EAP. Hlavním důvodem je neschopnost metody MLE hodnotit úroveň rysu v případě zodpovězení všech položek správně, nebo špatně, což se především u prvních několika položek může snadno stát. Další výběr položek na základě úrovně theta je v tomto případě nemožný a musí být zvolena například položka na průměrné úrovni. To může hlavně u velmi krátkých testů, kterým eEOD je, způsobit snížení přesnosti a zvýšit počet administrovaných položek.

## 9.5 Prostor pro zlepšení a další výzkum

Jak již bylo několikrát řečeno, nejzásadnějším místem zlepšení by měla být **velikost vzorku**. Počty respondentů u kalibrací CAT testů pro běžné použití se pohybují v tisících, takže 124 respondentů je skutečně velmi malý vzorek. I z toho důvodu se možná nepodařilo splnit podmínku shody modelu s daty. Je možné, že by při větším počtu respondentů byl výraznější jeden faktor testu a byla by tím splněna podmínka unidimenzionality. Vhodné by také bylo ověření třetí podmínky lokální nezávislosti, která v tomto výzkumu ověřována nebyla. Lze předpokládat, že při splnění těchto podmínek by byly korelace mnohem silnější.

Pro dosažení zobecnitelných výsledků by také bylo nezbytné provést kalibraci a následně **administraci celého EOD testu** a ne jen jedné škály. Zde by již bylo nutné využít multidimenzionální IRT modely. Využití těchto modelů by samozřejmě mělo vliv na shodu modelu s daty. Zvláště zajímavé by mohly být výsledky u lži škály, která dosud zkoumána nebyla. Pro zlepšení výsledků by také bylo nezbytné **dovyvinout dostatek dalších položek a to hlavně v oblasti obou extrémů**. Příliš malá položková banka je totiž limitem pro využití plného potenciálu IRT (Fan, 1998). S nedostatkem položek v extrémních úrovních obtížnosti se setkal i výzkum Stage (1996). Počty položek v klasických testech tedy zpravidla zdaleka nedostačují pro vytvoření položkové banky CAT testů (Thompson & Weiss, 2011).

Jisté výhody by také mohlo mít **experimentální provedení**, kde by respondenti byli náhodně rozřazeni do skupin a následně testováni formou administrace podle své skupiny.

Tím by byl neutralizován přinejmenším vliv časového odstupu a případně i další intervenující proměnné.

Kromě osobnostních testů by bylo vhodné zkoumat také testy výkonnostní. Mohlo by být velmi přínosné provést **srovnání efektivity u výkonnostních a osobnostních testů**. U výkonnostních testů by bylo mnohem důležitější zajistit stejné počítačové vybavení a celkově stejné podmínky testování. Vhodné by bylo i zjištění úrovně schopností v práci s počítačem a získání zpětné vazby od respondentů. Možností by mohlo být i **srovnání skutečné časové úspory u CAT**. To bohužel v mém výzkumu nebylo možné, protože u klasických forem časový limit měřen nebyl a u počítačových online forem byly časové údaje zkresleny ze strany respondentů přerušováním testu ale i ze strany serveru, kde docházelo ke zpožďování zobrazení položek v řádech sekund. Měření času by bylo možné hlavně u klasických počítačových testů, kde zpožďování zobrazení položek prakticky nehrozí.

Celkově tato práce přináší zajímavé výsledky hlavně v oblasti **online CAT**, které zatím není dostatečně prozkoumáno. Přináší také podrobnější zkoumání **reálné administrace CAT** v porovnání se simulacemi, které bývají použity mnohem častěji, a jejichž ekvivalence k reálnému testování nebyla nijak ověřována. Přínosem je i srovnání zmíněných forem administrace s **tužka-a-papír administrací**, která se může od počítačových administrací výrazně lišit, což není vždy zohledňováno.

# Závěr

Cílem této práce bylo přiblížit problematiku **teorie odpovědi na položku a počítačového adaptivního testování**, která u nás zatím není příliš známá. CAT má mnoho výhod mezi které patří hlavně **efektivita testování**, která se zvyšuje spolu s využitím sofistikovanějších IRT modelů. Aby mohl být využit skutečný potenciál CAT, je nutné mít k dispozici dostatečně velkou položkovou banku, čítající stovky až tisíce položek. Tyto položky by také měly být kalibrovány na dostatečně velkém počtu respondentů, aby bylo dosaženo co nejlepší shody modelu s daty. Přes všechny výhody které CAT přináší jsou testy tohoto typu stále **mnohem dražší než testy klasické**, což je nejspíš hlavní překážkou pro jejich masové rozšíření spolu s výpočetní náročností IRT.

V práci byla shrnuta teorie a některé výzkumy zabývající se IRT a CAT. V praktické části pak byly **demonstrovány zmíněné výhody CAT**. Přestože vysoká efektivita CAT v mém výzkumu je prokazatelná, výsledky by byly mnohem lepší při použití výrazně většího souboru respondentů a splnění podmínek pro využití jednoduchých IRT modelů (případně využití modelů multidimenzionálních). Zároveň lze očekávat lepší výsledky při ověřování efektivity u výkonnostních testů.

Teorie odpovědi na položku a počítačové adaptivní testování jsou oblasti psychodiagnostiky, kde **lze očekávat velký rozvoj do budoucnosti**. Vývoj těchto oblastí bude velmi souviset také s vývojem počítačů, které stále přinášejí nové možnosti. Zatím si však nelze představit, že by papírové testy byly zcela nahrazeny těmi počítačovými.

## Bibliografie


1. Amarnani, R. (2009). Two theories, One theta: A gentle introduction to item response theory as an alternative to classical test theory. The International Journal of Educational and Psychological Assessment, (3), pp. 104-109.

2. Baker, F.B. (2001). The basics of item response theory. (2. vydání). College Park, Md.: ERIC Clearinghouse on Assessment and Evaluation.

3. Barrada, J., Veldkamp, B.P., & Olea, J. (2008). Multiple Maximum Exposure Rates in Computerized Adaptive Testing. Applied Psychological Measurement, 33(1), pp. 58-73.

4. Barrada, J.R., Abad, F.J., & Veldkamp, B.P. (2009). Comparison of methods for controlling maximum exposure rates in computerized adaptive testing. Psicothema, 21(2), pp. 313-320.

5. Barrada, J.R., Olea, J., Ponsoda, V., & Abad, F.J. (2008). Incorporating randomness in the Fisher information for improving item-exposure control in CATs. British Journal of Mathematical and Statistical Psychology, 61(2), pp. 493-513.

6. Barrada, J.R., Olea, J., Ponsoda, V., & Abad, F.J. (2010). A Method for the Comparison of Item Selection Rules in Computerized Adaptive Testing. Applied Psychological Measurement, 34(6), pp. 438-452.

7. Becker, J., Fliege, H., Kocalevent, R.D., Bjorner, J., Rose, M., Walter, O., & Klapp, B.F. (2008). Functioning and validity of A Computerized Adaptive Test to measure anxiety (A-CAT). Depression and Anxiety, 25(12), pp. E182-E194.

8. Bock, R.D. (1997). A Brief History of Item Theory Response. Educational Measurement: Issues and Practice, 16(4), pp. 21-33.

9. De Ayala, R.J., Dodd, B.G., & Koch, W.R. (1992). A Comparison of the Partial Credit and Graded Response Models in Computerized Adaptive Testing.. Applied Measurement in Education, 5(1).

10. Deng, H., Ansley, T., & Chang, H. (2010). Stratified and Maximum Information Item Selection Procedures in Computer Adaptive Testing. Journal of Educational Measurement, 47(2), pp. 202-226.

11. Eggen, T.J.H.M. (2004). Contributions to the theory and practice of computerized adaptive testing. [S.l: s.n.].

12. Embretson, S.E. (1992). Computerized Adaptive Testing: Its Potential Substantive Contributions to Psychological Research and Assessment. Current Directions in Psychological Science, 1(4), pp. 129-131.

13. Embretson, S.E., & Reise, S. (2000). Item response theory for psychologists. Mahwah: Lawrence Erlbaum Associates.

14. Eysenck, H., & Eysenck, S.B.G. (1993). Eysenckovy osobnostní dotazníky pro dospělé: příručka. Bratislava: Psychodiagnostika.

15. Eysenck, H.J., & Eysenck, S.B.G. (1968). Eysenckov osobnostný dotazník - EOD: příručka pro administraci a interpretaci testu. (1. vyd.). Bratislava: Psychodiagnostické a didaktické testy.

16. Fan, X. (1998). Item response theory and classical test theory: an empirical comparison of their item/person statistics. Educational and Psychological Measurement: a quarterly journal devoted to the development and application of measures of individual differences, 58(3), pp. 1-17.

17. Filípková, Z., & Byčkovský, P. (2008). Studie proveditelnosti počítačem adaptovaného testování v prostředí českých škol. Systémový projekt Kvalita I: CERMAT.

18. Finkelman, M.D., Nering, M., & Roussos, L. (2009). A Conditional Exposure Control Method for Multidimensional Adaptive Testing. Journal of Educational Measurement, 46(1), pp. 84-103.

19. Fliege, H., Becker, J., Walter, O.B., Rose, M., Bjorner, J.B., & Klapp, B.F. (2009). Evaluation of a computer-adaptive test for the assessment of depression (D-CAT) in clinical application. International Journal of Methods in Psychiatric Research, 18(1), pp. 23-36.

20. Forbey, J.D., & Ben-Porath, Y.S. (2007). Computerized adaptive personality testing: A review and illustration with the MMPI-2 computerized adaptive version. Psychological Assessment, 19(1), pp. 14-24.

21. Forbey, J.D., Handel, R.W., & Ben-Porath, Y.S. (2000). A real data simulation of computerized adaptive administration of the MMPI-A. Computers in Human Behavior, 16(1), pp. 83-96.

22. Fox, J. (2004). Modelling response error in school effectiveness research. Statistica Neerlandica, 58(2), pp. 138-160.

23. Gnambs, T., & Batinic, B. (2011). Polytomous Adaptive Classification Testing: Effects of Item Pool Size, Test Termination Criterion, and Number of Cutscores. Educational and Psychological Measurement, 71(6), pp. 1006-1022.

24. Guo, J., Tay, L., & Drasgow, F. (2009). Conspiracies and Test Compromise: An Evaluation of the Resistance of Test Systems to Small-scale Cheating. International Journal of Testing, 9(4), pp. 283-309.

25. Halama, P. (2005). Adaptívne testovanie pomocou počítača: Aplikácia teórie odpovede na položku v diagnostike inteligencie. Psychológia a patopsychológia dieťaťa, 40(3), pp. 252-266.

26. Halama, P., & Biescad, M. (2011). Meranie psychoterapeutickej zmeny: porovnanie klasického testového skóre a skóre odvodeného prostredníctvom teórie odpovede na polozku. Československá psychologie, 55(5), pp. 400-411.

27. Halama, P., & Matús, B. (2006). Psychometrická analýza Rosenbergovej škály sebahotnotenia s použitím teród klasickej teórie testov (CTT) a teórie odpovede na položku (IRT). Československá psychologie, 50(6), pp. 569-583.

28. Han, K.T. (2012). SimulCAT: Windows Software for Simulating Computerized Adaptive Test Administration. Applied Psychological Measurement, 36(1), pp. 64-66.

29. Han, K.T. (2013). Item Pocket Method to Allow Response Review and Change in Computerized Adaptive Testing. Applied Psychological Measurement, 37(4), pp. 259-275.

30. Hendl, J. (2009). Přehled statistických metod: analýza a metaanalýza dat. (3., přeprac. vyd.). Praha: Portál.

31. Hornke, L.F. (2000). Item Response Times in Computerized Adaptive Testing. Psicológica: Revista de metodología y psicología experimental, (21), pp. 175-190.

32. Hula, W.D., Fergadiotis, G., & Martin, N. (2012). Model Choice and Sample Size in Item Response Theory Analysis of Aphasia Tests. American Journal of Speech-Language Pathology, 21(2), pp. S38-S50.

33. Chang, S.R., Plake, B.S., Kramer, G.A., & Lien, S.M. (2011). Development and Application of Detection Indices for Measuring Guessing Behaviors and Test-Taking Effort in Computerized Adaptive Testing. Educational and Psychological Measurement, 71(3), pp. 437-459.

34. Chatzopoulou, D.I., & Economides, A.A. (2010). Adaptive assessment of student's knowledge in programming courses. Journal of Computer Assisted Learning, 26(4), pp. 258-269.

35. Chen, J., White, S., McCloskey, M., Soroui, J., & Chun, Y. (2011). Effects of computer versus paper administration of an adult functional writing assessment. Assessing Writing, 16(1), pp. 49-71.

36. Chen, S., Lei, P., & Liao, W. (2008). Controlling item exposure and test overlap on the fly in computerized adaptive testing. British Journal of Mathematical and Statistical Psychology, 61(2), pp. 471-492.

37. Cheng, Y., & Morgan, D.L. (2008). Comparison of Methods for Constrained CAT Item Selection in Classification Accuracy and Consistency. In: http://research.collegeboard.org. Retrieved from: http://research.collegeboard.org/sites/default/files/publications/2012/7/presentation-2008-19a-item-selection-computerized-adaptive-testing.pdf

38. Cheng, Y., & Morgan, D.L. (2013). Classification accuracy and consistency of computerized adaptive testing. Behavior Research Methods, 45(1), pp. 132-142.

39. Choi, S.W. (2009). Firestar: Computerized Adaptive Testing Simulation Program for Polytomous Item Response Theory Models. Applied Psychological Measurement, 33(8), pp. 644-645.

40. Choi, S.W., & Swartz, R.J. (2009). Comparison of CAT Item Selection Criteria for Polytomous Items. Applied Psychological Measurement, 33(6), pp. 419-440.

41. Choi, S.W., Grady, M., & Dodd, B.G. (2011). A New Stopping Rule for Computerized Adaptive Testing. Educational and Psychological Measurement, 71(1), pp. 37-53.

42. Choi, S.W., Podrabsky, T., & McKinney, N. (2012). Firestar-D: Computerized Adaptive Testing Simulation Program for Dichotomous Item Response Theory Models. Applied Psychological Measurement, 36(1), pp. 67-68.

43. Jelínek, M., Květoň, P., & Denglerová, D. (2006). Adaptivní testování: základní pojmy a principy. Československá psychologie, 50(2), pp. 163-173.

44. Jelínek, M., Květon, P., & Vobořil, D. (2011a). Adaptivní administrace NEO PI-R: výhody a omezení. Československá psychologie, 55(1), pp. 69-81.

45. Jelínek, M., Květon, P., & Vobořil, D. (2011b). Testování v psychologii: teorie odpovědi na položku a počítačové adaptivní testování. Praha: Grada.

46. Katz, L., & Dalby, J. (1981). Computer-assisted and traditional psychological assessment of elementary-school-aged children. Contemporary Educational Psychology, 6(4), pp. 314-322. DOI: 10.1016/0361-476X(81)90014-X.

47. Kingsbury, G.G., & Wise, S.L. (2000). Practical Issues in Developing and Maintaining a Computerized Adaptive Testing Program. Psicológica: Revista de metodología y psicología experimental, (21), pp. 135-156.

48. Klinkenberg, S., Straatemeier, M., & van der Maas, H.L.J. (2011). Computer adaptive practice of Maths ability using a new item response model for on the fly ability and difficulty estimation. Computers, 57(2), pp. 1813-1824.

49. Koch, W.L., & Dodd, B.G. (1989). An Investigation of Procedures for Computerized Adaptive Testing Using Partial Credit Scoring. Applied Measurement in Education, 2(4).

50. Kujal, P. (2008). Aplikace teorie odpovědi na položku: Odlišné fungování položek Eysenckova osobnostního dotazníku podle pohlaví. Brno.

51. Květon, P., & Klimusová, H. (2002). Metodologické aspekty počítačové administrace psychodiagnostických metod. Československá psychologie, 46(3), pp. 251-264.

52. Květon, P., Jelínek, M., Denglerová, D., & Vobořil, D. (2008). Software pro adaptivní testování: CAT v praxi. Československá psychologie, 52(2), pp. 145-154.

53. Květon, P., Jelínek, M., Vobořil, D., & Klimusová, H. (2007). Computer-based tests: the impact of test design and problem of equivalency. Computers in Human Behavior, 23(1), pp. 32-51.

54. Květoň, P., Jelínek, M., Vobořil, D., & Klimusová, H. (2012). Rozbor volby odpověďových kategorií v testu prostorové představivosti s využitím teorie odpovědi na položku. Československá psychologie, 56(1), pp. 31-40.

55. Květon, P., Martin, J., Vobořil, D., & Klimusová, H. (2003). Ekvivalence tradiční a počítačové formy testu IST-70. Československá psychologie, 47(6), pp. 562-572.

56. Linden, W.J., & Glas, C.A.W. (2010). Elements of adaptive testing. New York: Springer.

57. Magis, D., & Raiche, G. (2011). catR: An R Package for Computerized Adaptive Testing. Applied Psychological Measurement, 35(7), pp. 576-577.

58. Makransky, G., Mortensen, E., & Glas, C.A.W. (2013). Improving Personality Facet Scores With Multidimensional Computer Adaptive Testing: An Illustration With the Neo Pi-R. Assessment, 20(1), pp. 3-13.

59. McKay, D. (2008). Handbook of research methods in abnormal and clinical psychology. Los Angeles: Sage Publications.

60. Mills, C.N., & Stocking, M.L. (1996). Practical Issues in Large-Scale Computerized Adaptive Testing. Applied Measurement in Education, 9(4), pp. 287-305.

61. Moreland, K. (1985). Validation of computer-based test interpretations: Problems and prospects. Journal of Consulting and Clinical Psychology, 53(6), pp. 816-825. DOI: 10.1037/0022-006X.53.6.816.

62. Olea, J., Abad, F.J., Ponsoda, V., Barrada, J.R., & Aguado, D. (2011). eCAT-Listening: Design and psychometric properties of a computerized adaptive test on English Listening. Psicothema, 23(4), pp. 802-807.

63. Olea, J., Revuelta, J., Ximénez, M.C., & Abad, F.J. (2000). Psychometric and psychological effects of review on computerized fixed and adaptive tests. Psicológica, (21), pp. 157-173.

64. Ozaki, K., & Toyoda, H. (2009). Item difficulty parameter estimation using the idea of the graded response model and computerized adaptive testing. Japanese Psychological Research, 51(1), pp. 1-12.

65. Özyurt, H., Özyurt, Ö., Baki, A., & Güven, B. (2012). Integrating computerized adaptive testing into UZWEBMAT: Implementation of individualized assessment module in an e-learning system. Expert Systems with Applications, 39(10), pp. 9837-9847.

66. Parshall, C.G. (2002). Practical considerations in computer-based testing. New York: Springer.

67. Qing Yi, Jinming Zhang, & Chang, H.H. (2008). Severity of Organized Item Theft in Computerized Adaptive Testing: A Simulation Study. Applied Psychological Measurement, 32(7), pp. 543-558.

68. Reckase, M. (2009). Multidimensional item response theory. New York: Springer.

69. Reeve, B.B., & Fayers, P. (2005). applying item response theory modelling for evaluating questionnaire item and scale properties. In Assessing quality of life in clinical trials: methods and practice. (2nd ed.). New York: Oxford University Press.

70. Reise, S.P., & Henson, J. (2000). Computerization and Adaptive Administration of the NEO PI-R. Assessment, 7(4), pp. 347-364.

71. Reise, S.P., Widaman, K.F., & Pugh, R.H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. Psychological Bulletin, 114(3), pp. 552-566.

72. Rulison, K.L., & Loken, E. (2008). I've Fallen and I Can't Get Up: Can High-Ability Students Recover From Early Mistakes in CAT?. Applied Psychological Measurement, 33(2), pp. 83-101.

73. Schermis, M.D., & Stemmer, P.M. (1996). Computerized adaptive skill assessment in a statewide testing. Journal of Research on Computing in Education, 29(1), pp. 49-68.

74. Schmettow, M., & Vietze, W. (2008). Introducing item response theory for measuring usability inspection processes. Retrieved from: http://portal.acm.org/citation.cfm?doid=1357054.1357196

75. Schmitt, T.A., Sass, D.A., Sullivan, J.R., & Walker, C.M. (2010). A Monte Carlo Simulation Investigating the Validity and Reliability of Ability Estimation in Item Response Theory with Speeded Computer Adaptive Tests. International Journal of Testing, 10(3), pp. 230-261.

76. Schuhfried. (2013) Vienna Test System: Schuhfried GmbH. Retrieved from: http://www.schuhfried.com

77. Smits, N., Cuijpers, P., & van Straten, A. (2011). Applying computerized adaptive testing to the CES-D scale: A simulation study. Psychiatry Research, 188(1), pp. 147-155.

78. Stage, C. (1996). An attempt to fit IRT models to the DS subtest in the SweSAT. Educational measurement, (19).

79. Stage, C. (1997a). The applicability of item response models to the SweSAT. A Study of the DTM Subtest. Educational measurement(21). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59545_enr2197sec.pdf

80. Stage, C. (1997b). The applicability of item response models to the SweSAT. A Study of the ERC Subtest. Educational measurement(24). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59547_enr2497sec.pdf

81. Stage, C. (1997c). The applicability of item response models to the SweSAT. A Study of the READ Subtest. Educational measurement(25). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59548_enr2597sec.pdf

82. Stage, C. (1997d). The applicability of item response models to the SweSAT. A Study of the WORD Subtest. Educational measurement(26). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59549_enr2697sec.pdf

83. Stage, C. (1998a). A Comparison Between Item Analysis Based on Item Response Theory and on Classical Test Theory. A Study of the SweSAT Subtest WORD. Educational measurement(29). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59551_enr2998sec.pdf

84. Stage, C. (1998b). A Comparison Between Item Analysis Based on Item Response Theory and on Classical Test Theory. A Study of the SweSAT Subtest ERC. Educational measurement(30). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59552_enr3098sec.pdf

85. Stage, C. (1999a). A Comparison Between Item Analysis Based on Item Response Theory and on Classical Test Theory. A Study of the SweSAT Subtest READ. Educational measurement(33). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59554_enr3399sec.pdf

86. Stage, C. (1999b). Predicting Gender Differences in WORD Items. A Comparison of item Response Theory and Classical Test Theory. Educational measurement(34). Retrieved from: http://www.edusci.umu.se/digitalAssets/59/59555_enr3499sec.pdf

87. Stage, C. (2003). Classical test theory or item response theory: the swedish experience. Educational measurement(42). Retrieved from: http://www.edusci.umu.se/digitalAssets/60/60581_em-no-42.pdf

88. Stančák, A. (1996). Klinická psychodiagnostika dospelých. Nové zámky: Psychoprof.

89. Svoboda, M. (1999). Psychologická diagnostika dospělých. (vyd. 2.). Praha: Portál.

90. Templin, J. (2013) Jonathan Templin's Website: Psychometrics and Statistics. Retrieved from: http://jonathantemplin.com

91. The International Association for Computerized and Adaptive Testing (IACAT). (2010) Retrieved from: http://iacat.org

92. Thompson, N.A., & Weiss, D.J. (2011). A Framework for the Development of Computerized Adaptive Tests. Practical Assessment, Research & Evaluation, 16(1), pp. 1-9.

93. Triantafillou, E., Georgiadou, E., & Economides, A.A. (2008). The design and evaluation of a computerized adaptive test on mobile devices. Computers, 50(4), pp. 1319-1330.

94. Urbánek, T., Denglerová, D., & Širůček, J. (2011). Psychometrika: měření v psychologii. Praha: Portál.

95. van der Linden, W.J., & Glas, C.A.W. (2000). Computerized Adaptive Testing: Theory and Practice. New York: Kluwer Academic Publishers.

96. Vispoel, W.P., Rocklin, T.R., & Tianyou, W. (1994). Individual Differences and Test Administration Procedures: A Comparison of Fixed-Item, Computerized-Adaptive, and Self-Adapted Testing. Applied Measurement in Education, 7(1), pp. 53-80.

97. Vogels, A.G.C., Jacobusse, G.W., & Reijneveld, S.A. (2011). An accurate and efficient identification of children with psychosocial problems by means of computerized adaptive testing. BMC Medical Research Methodology, 11(1), pp. 111-.

98. Vos, H.J. (2000). A Bayesian Procedure in the Context of Sequential Mastery Testing. Psicológica, (21), pp. 191-211.

99. Wainer, H., & Dorans, N.J. (2000). Computerized adaptive testing: a primer. (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum Associates.

100. Waller, N.G., & Reise, S.P. (1989). Computerized adaptive personality assessment: An illustration with the Absorption scale. Journal of Personality and Social Psychology, 57(6), pp. 1051-1058.

101. Wang, W.C., & Liu, C.W. (2011). Computerized Classification Testing Under the Generalized Graded Unfolding Model. Educational and Psychological Measurement, 71(1), pp. 114-128.

102. Weiss, D.J. (2004). Computerized Adaptive Testing for Effective and Efficient Measurement in Counseling and Education. Measurement & Evaluation in Counseling & Development (American Counseling Association), 37(2), pp. 70-84.

103. Williams, J.E., & McCord, D.M. (2006). Equivalence of standard and computerized versions of the Raven Progressive Matrices Test. Computers in Human Behavior, 22(5), pp. 791-800.

104. Yen, Y.C., Ho, R.G., Laio, W.W., Chen, L.J., & Kuo, C.C. (2012). An Empirical Evaluation of the Slip Correction in the Four Parameter Logistic Models With Computerized Adaptive Testing. Applied Psychological Measurement, 36(2), pp. 75-87.

105. Youngseok L., Jungwon, Ch., Sugjae, H., & Byung-Uk, Ch. (2010). A Personalized Assessment System Based on item response Theory. pp. 381-386.

106. Zheng, Y., Chang, C., & Chang, H. (2013). Content-balancing strategy in bifactor computerized adaptive patient-reported outcome measurement. Quality of Life Research, 22(3), pp. 491-499.

107. Žitný, P. (2011). Presnosť, validita a efektívnosť počítačového adaptívneho testovania. Československá psychologie, 55(2), pp. 167-179.

108. Žitný, P., Halama, P., Jelínek, M., & Květon, P. (2012). Validity of cognitive ability tests - comparison of computerized adaptive testing with paper and pencil and computer-based forms of administrations. Studia Psychologica, 54(3), pp. 181-194.

# Přílohy

## Odhadnuté parametry

Tabulka 26 Výsledný odhad parametrů položek

| Položka | a | b | Položka | a | b |
|---|---|---|---|---|---|
| 1 | 1,28 | -1,32 | 25 | 0,42 | 0,82 |
| 2 | 0,37 | -0,19 | 26 | 2,00 | -1,37 |
| 3 | 0,44 | 2,81 | 27 | 0,09 | -5,00 |
| 4 | 0,63 | -0,75 | 28 | 2,00 | 1,91 |
| 5 | 2,00 | 1,00 | 29 | 2,00 | 0,14 |
| 6 | 1,41 | -0,12 | 30 | 2,00 | 2,81 |
| 7 | 2,00 | -0,35 | 31 | 0,21 | 1,33 |
| 8 | 1,13 | 0,14 | 32 | 0,39 | -1,04 |
| 9 | 0,77 | -0,09 | 33 | 0,38 | -1,63 |
| 10 | 0,33 | 0,03 | 34 | 0,39 | -0,80 |
| 11 | 0,32 | 4,21 | 35 | 2,00 | 1,71 |
| 12 | 0,46 | 1,83 | 36 | 1,51 | -0,31 |
| 13 | 0,85 | 0,24 | 37 | 1,23 | -0,09 |
| 14 | 2,00 | -0,02 | 38 | 2,00 | 0,63 |
| 15 | 1,10 | 1,65 | 39 | 0,49 | 0,67 |
| 16 | 2,00 | 1,39 | 40 | 1,42 | 1,17 |
| 17 | 0,99 | 0,36 | 41 | 0,50 | -0,78 |
| 18 | 0,66 | 0,26 | 42 | 1,12 | 0,61 |
| 19 | 0,31 | -0,33 | 43 | 2,00 | 1,12 |
| 20 | 0,77 | -3,23 | 44 | 0,65 | -0,66 |
| 21 | 0,26 | 1,62 | 45 | 2,00 | 1,91 |
| 22 | 1,78 | -0,57 | 46 | 0,65 | -0,57 |
| 23 | 0,55 | 0,47 | 47 | 2,00 | 1,64 |
| 24 | 1,91 | -0,15 | 48 | 0,99 | 1,05 |

## Dotazník sebehodnocení

**Instrukce:** „Vyberte prosím na škále bod, který nejvíce odpovídá vašemu umístění mezi dvěma vlastnostmi. (Např. Pokud jste extravert, označíte políčko nejblíže slovu extravert.)“

Tabulka 27 Sebehodnotící dotazník použitý ve druhé fázi výzkumu

| | 7 | 6 | 5 | 4 | 3 | 2 | 1 | |
|---|---|---|---|---|---|---|---|---|
| Otevřený | | | | | | | | Uzavřený |
| Společenský | | | | | | | | Samotářský |
| Upovídaný | | | | | | | | Tichý |
| Optimistický | | | | | | | | Pesimistický |
| Impulzivní | | | | | | | | Rozvážný |
| Bezstarostný | | | | | | | | Úzkostný |
| Aktivní | | | | | | | | Pasivní |
| Riskující | | | | | | | | Opatrný |
| Sebevědomý | | | | | | | | Nejistý |
| Konající | | | | | | | | Plánující |