# Automatické generování položek

[TODO: kompletně přepracovat]

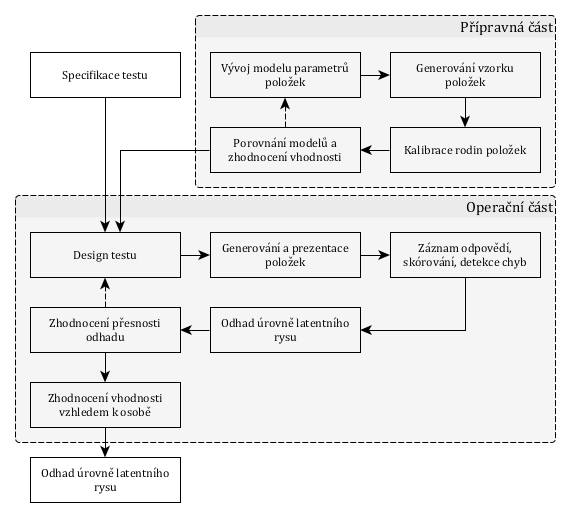
K měření rozličných atributů jedinců, jako jsou např. dovednosti, kognitivní schopnosti, osobnostní rysy, postoje atd., jsou používány testy, které se obvykle skládají z jednotlivých položek. Těch test obvykle obsahuje větší množství, aby se zvýšila přesnost měření. Navíc spolu s tím, jak se míra využívání testů, zejména ve vzdělávacím sektoru, zvyšuje, jsou kladeny stále větší nároky na jejich zabezpečení proti podvádění a trénování – což vede k nutnosti vytváření větších položkových bank (Arendasy & Sommer, 2012). Tvorba a kalibrace položek proto tvoří nejnáročnější a nejnákladnější část procesu vzniku každého testu.

Klasický způsob vytváření položek má řadu nevýhod: jejich tvůrci se ne vždy drží daných pokynů a směrnic, a nejsou schopni vymýšlet položky přesahující jejich vlastní schopnosti (srov. Hornke & Habon, 1986); tvorba zabírá velké množství času (Wainer, 2002); a velká část položek musí být v průběhu kalibrace vyřazena kvůli nedostatečným psychometrickým parametrům (Arendasy & Sommer, 2012). Automatické generování položek (*automatic item generation*, AIG) představuje způsob, jak se s pomocí kognitivních teorií, počítačové techniky a psychometriky s těmito problémy vyrovnat. V kombinaci s počítačovým adaptivním testováním (*computerized adaptive testing*, CAT) pak dovoluje značnou část diagnostického procesu částečně či plně zautomatizovat. Tím se potenciálně šetří čas a zdroje, snižuje prozrazování položek (protože každá osoba dostane unikátní test), ale také zvyšuje validitu testu – neboť automaticky vytvářené položky musí být generovány podle standardizovaných pravidel (Geerlings, 2012).

AIG můžeme v ideální podobě popsat jako proces, kdy je z jediné šablony generováno více kalibrovaných položek (Gierl, Lai & Breithaupt, 2012). To vyžaduje popsat testovací úkol z hlediska jeho vlastností anebo elementů, které mohou být manipulovány a tyto element variovat tak, aby se algoritmicky vygenerovaly nové položky; psychometrické parametry položek jsou pak, díky využití statistických metod, založeny na dané kombinaci použitých elementů. Aby bylo zajištěno, že kvalita výsledného testu se neliší od těch „ručně“ vytvářených, je nutné použít metody, které do tohoto proces generování zahrnou i další náležitosti, jako je cíl a účel diagnostiky, a obsah a obtížnost položek.

V dalších částech tohoto textu podrobněji rozvedu, jak AIG zapadá do konceptu automatizované diagnostiky.

## Automatizovaná diagnostika

Příprava automatizovaného diagnostického systému se obecně sestává z vývoje možných způsobů, jak modelovat položky (a zejména jejich obtížnost), odhad parametrů těchto modelů, zhodnocení vhodnosti modelů a volba některého z nich. Vybraný model a specifikace testu pak slouží jako vstup operační části diagnostického systému, který se stará o návrh testu, jeho vygenerování, a skórování odpovědí respondenta.

Design celého systému znázorňuje diagram; každý z kroků by měl fungovat bez, či pouze s minimálním zapojením administrátora testu. Proto je nutné vytvořit nové, případně upravit a přizpůsobit existující diagnostické postupy. Podrobnost k vybraným krokům přípravné i operační části budou rozebrány v dalších kapitolách. Lze říci, že AIG zhruba pokrývá právě přípravnou část, a tudíž tvoří primární zájem první fáze mého dizertačního projektu.

## Vývoj modelu parametrů položek

AIG ve svém základu vyžaduje identifikaci vlastností, rysů či elementů položek, jejichž kombinací vznikají položky relevantní k měřenému konstruktu. Elementy můžeme rozdělit na dvě skupiny. Do první patří ty, které maximalizují rozptyl obtížnosti položky – tyto elementy jsou nazývány radikály (*radicals*). Ty elementy položek, které vytvářejí variace položek bez systematického účinku na jejich obtížnost, a mohou tak být využity ke zvýšení rozmanitosti položek se stejnou obtížností, označujeme jako *incidentals* (Geerlings, 2012). Položky, které se neliší v radikálech, ale pouze v incidentals, patří do jedné rodiny položek.

Jakým způsobem jsou tyto druhy elementu identifikovány a využívány k odhadu parametrů položek závisí na konkrétním přístupu AIG. Ty můžeme rozdělit do dvou širších kategorií: přístupy dovolující předpovědět psychometrické parametry položek jsou nazývány jako ty, které vycházejí ze silné teorie (*strong theory*); pokud ovšem žádná taková teorie není k dispozici (což se většinou týká měření jiných než kognitivních schopností), hovoří se o generování ze slabé teorie (*weak theory*), a cílem pak není předpovědět měnící se parametry položek, ale udržovat je u každé rodiny položek stejné (Lai, Alves & Gierl, 2009, Haladyna, 2013).

Protože AIG je vhodné především pro výkonové testy, budu se dále zaměřovat právě na ně.

##### přístup založený na šabloně položky

Tento přístup je v podstatě základním; proces generování začíná výběrem položek, které mají podobné psychometrické kvality, a jejichž určité prvky – irelevantní vzhledem k povaze úkolu, na který se položka dotazuje – lze nahradit jednou či více proměnnými a vytvořit tak šablonu položky. Změnou hodnot jednotlivých proměnných se z modelu generují konkrétní položky (Embretson & Yang, 2007). Nevýhodou tohoto přístupu je limitovaný počet odlišných položek, a naopak relativně velké množství položek, které musí být po fázi kalibrace eliminovány kvůli nevyhovujícím psychometrickým parametrům. Položky založené na jednom modelu lze rovněž snadněji trénovat (Arendasy & Sommer, 2012).

##### Přístup založený na designu kognitivního systému

Struktura položek je založena přímo na příslušné kognitivní teorii, která specifikuje proces řešení položky a výsledný výkon jednotlivce a vliv vlastností položky na tento proces – jednotlivé elementy položek se vztahují k obtížnosti tohoto zpracování. Primárním psychometrickým parametrem je tedy obtížnost položky, nicméně ovlivněn může být i diskriminační parametr (Embretson & Yang, 2007, Arendasy & Sommer, 2012).

##### automatický min–max přístup

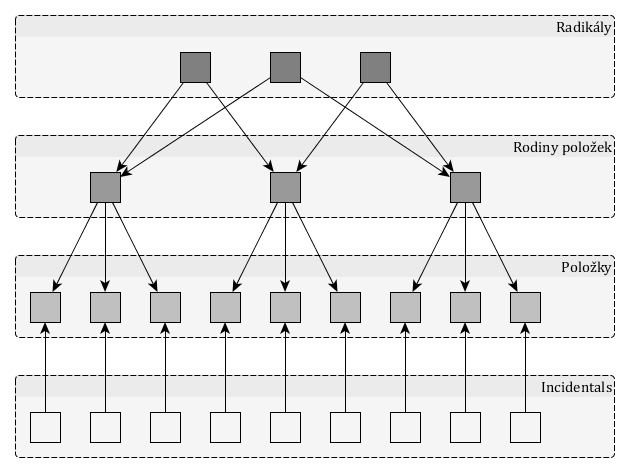
Pro překonání problémů spojených se ztrátou položek po jejich kalibraci vyvinul Arendasy s kolegy automatický min–max přístup, který lze považovat za extenzi přístupu předchozího (viz např. Arendasy & Sommer, 2012): konstrukce generátoru položek začíná definicí měřeného latentního rysu a souvisejícího kognitivního modelu. Poté je upřesněn formát položky, a kognitivní model je specifikován vzhledem k němu; jsou určeny radikály a *incidentals*.

Dále je nutné formálně definovat, které elementy položek je třeba z procesu tvorby vyřadit, aby se snížila interference těch kognitivních procesů, které s měřeným konstruktem nesouvisejí – tyto vlastnosti položky pak tvoří funkční omezení (*functional constraints*). Zatímco radikály tedy plní roli generativní komponenty, funkční omezení implementují komponentu kontroly kvality. Tyto dvě komponenty odlišují automatický min–max přístup od přístupu založeného na designu kognitivního systému (Arendasy & Sommer, 2012).

## Kalibrace rodin položek

Položky jsou vytvářeny kombinací radikálů a *incidentals* podle pravidel vybraného AIG přístupu; pokud vygenerujeme reprezentativní sadu položek a získáme s její pomocí odpovědi od respondentů, lze tato data využít ke kalibraci parametrů psychometrického modelu. Ten je jednak nutný k potvrzení správnosti kategorizace elementů položek na radiály a *incidentals*, a v pozdějších fázích lze ten model, který nejlépe vystihuje data, používat ve výsledném testu k odhadu úrovně latentního rysu osob.

Při použití IRT modelů pro kalibraci položek je parametr osob obvykle pokládán za náhodný, zatímco parametry položek jako fixované (De Boeck, 2008). Pokud však předpokládáme, že položky jsou generovány jako instance šablony z určité rodiny položek, jako je tomu v případě AIG, můžeme jejich parametry rovněž považovat za náhodné proměnné. Takovýto IRT model – tj. s náhodnými parametry osob i položek – se těžko odhaduje při použití frekventistického přístupu – z toho důvodu je vhodnější používat přístup Bayesovský.

Pro automaticky generované položky vzniklo již několik statistických modelů; jako nejužitečnější se v současnosti jeví ten, který vytvořil Geerlings s kolegy (Geerlings, Glas & van der Linden, 2011, Geerlings, van der Linden & Glas, 2013) kombinací LLTM[[1]](#footnote-1) a RSM[[2]](#footnote-2) – *linear item cloning model* (LICM). Snaží se jednak popsat vliv „klonování“ položek – tedy vytváření instancí z jedné rodiny –, ale i vysvětlit předpokládaný vliv radikálů na obtížnost položek (viz diagram).

Model lze definovat následovně: f = 1,… F rodin položek, přičemž rodina f se skládá z položek if = 1,… If. Celkem existuje K položek. Rodiny jsou určeny kombinací radikálů r = 1,… R. Každá z n = 1,… N osob zodpoví podmnožinu z K položek, výsledkem čehož je vektor odpovědí s realizacemi proměnných odpovědi = {0, 1} pro každou osobu n.

#### První úroveň modelu

První úroveň specifikuje pravděpodobnost správné odpovědi osobou na položku:

Jedná se o tříparametrový normální model (3PNO) s parametry , , , jež představují diskriminační parametr, obtížnost a pseudouhádnotelnost, parametrem osoby a kumulativní normální distribuční funkcí .

#### Druhá úroveň modelu

Parametry položky, označené jako , jsou následně transformovány:

Díky této transformaci můžeme u předpokládat vícerozměrné normální rozložení:

kde je vektor průměrů parametrů položky a kovarianční matice parametrů položky pro rodinu položek *f*.

Průměrná obtížnost rodiny je lineární kombinací účinků radikálů použitých ke generování položky:

kde je účinek radikálu *r* na průměrnou obtížnost rodiny položek a je designová proměnná určující kolikrát by měl být radikál *r* v položce použitý pro vygenerování položky z rodiny *f*. Na úrovni položky tedy platí:

kde je druhý diagonální prvek . Z toho plyne, že radikály determinují průměr obtížnosti rodiny , zatímco *incidentals* určují kovarianční matici rodiny .

Dále předpokládáme, že má normální rozložení s průměrem a směrodatnou odchylkou . Pro identifikaci modelu je stanoveno jako 0 a jako 1.

V praktické aplikaci musí být rozdělení radikálů a *incidentals* jako fixovaných a náhodných proměnných ověřeno podle empirických výsledků. Stejně tak je třeba zjistit, zda funkční omezení (dle automatického min–max přístupu) správně plní svoji roli.

Celkově však úspěch AIG závisí na tom, jak dobře radikály předpovídají průměrnou obtížnost rodiny položek – předpokládaný vliv radikálů nemusí být nalezen, stejně tak je třeba brát v úvahu případnou interakci radikálů. Nemusí také platit implicitní předpoklad modelu, že radikály mají stejný vliv pro každého respondenta.

## Design testu

Při použití CAT dostává respondent k zodpovězení otázky, které nejlépe odpovídají jeho odhadovaným schopnostem, tj. obvykle položky mající nejvyšší informaci vzhledem k odhadu úrovně latentního rysu.

Při použití generovaných položek existují tři možnosti, jak sestavovat konkrétní test. Tou první je používat položky, které byly generované a kalibrované předem. Tento způsob nicméně popírá důvody, proč AIG používat – nikterak nezvyšuje praktickou velikost položkové banky.

Druhou možností je generovat položky v průběhu testování za využití předem kalibrovaných rodin položek. To znamená, že se vybírá vždy taková rodina, která má nejvyšší informaci vzhledem k odhadu úrovně latentního rysu, a její kombinace radikálů se doplní náhodně vybranými *incidentals*.

Třetí možností je pak generovat položky v průběhu testování pouze za využití radikálů – to poskytuje ještě větší položkovou banku, neboť nejsme omezeni předem danými kombinacemi radikálů do rodin. Nutné ovšem je nejen mít správně odhadnuté radikály, ale navíc předpokládáme, že diskriminační parametr a pseudouhádnotelnost jsou stejné napříč rodinami, stejně jako kovarianční matice parametrů.

## Využití počítačů

Testování s využitím AIG lze těžko provádět bez využití počítačů. Zároveň je nepraktické, aby test, standardizovaný pro určitý počítač – či obecně zařízení[[3]](#footnote-3) –, byl administrovaný pouze na stejném druhu tohoto zařízení – dá se těžko předpokládat, že psychologové-administrátoři budou mít k dispozici vždy stejné zařízení, na kterém byl test kalibrován. Proto je nutné vzít v úvahu všechna úskalí, která testování s pomocí počítačů obnáší. Konkrétně se jedná o následující problémy (dle Naglieri et al., 2004, Širůček, 2010):

1. Neekvivalence grafických podnětů. „Stejné“ položky mohou na respondenta působit jinak na papíře a na displejích. Rovněž displeje samotné se u různých zařízení liší v řadě parametrů: velikost uhlopříčky, rozlišení, hustota pixelů, kalibrace barev či obnovovací frekvence. To činí velice obtížným, či přímo nemožným plně standardizovat grafickou podobu položek.
2. Specifika prohlížečů. Každé zařízení má jiný software, od operačního systému po webový prohlížeč, které se liší v nesčetném množství detailů.
3. Kompetence respondenta ovládat zařízení, na kterém test probíhá. Pokud uživatel neovládá práci s počítačem, negativně to ovlivní jeho výkon.
4. Odlišnosti ovládání. Různá zařízení lze ovládat mnoha způsoby: klávesnicí, myší, touchpadem, nebo dotykovým displejem; každý vede ke kvalitativně odlišnému zpracovávání – jinak se vybírá odpověď myší a jinak dotykovým displejem, což může hrát roli především v testech měřících čas. Ani způsob ovládání tak nelze standardizovat napříč zařízeními.

Tyto problémy jsou vesměs řešitelné třemi způsoby.

Prvním je zahrnout proměnné, mající případný vliv na přesnost měření, do bayesovského modelu diagnostiky odlišného fungování položek (*differential item functioning*, DIF) – to se týká především kompetencí respondenta, ale i rozličných vlastností zařízení (např. rozlišení a velikosti displeje).

Druhou možností je zahrnout vlastnosti zařízení do procesu generování položek. Tedy rozhodnout, které vlastnosti ovlivňují parametry položek (a fungují tedy podobně jako radikály) a které nikoliv, a lze je tedy považovat za *incidentals*, a dále určit, zda tyto radikály-vlastnosti zařízení statisticky modelovat, nebo je popsat pomocí komponenty kontroly kvality (tj. stanovit funkční omezení používaného zařízení anebo jeho určitých vlastností). Např. u „obyčejného“ dotazníku nejspíš na konkrétní grafické podobě nebude záležet; opačným extrémem naopak může být metoda založená na vnímání barev. Stejně tak zatímco text lze efektivně zpracovat a chápat i na malém displeji (Dillon, Richardson & McKnight, 1990), u grafických podnětů toto platit nemusí.

Třetí variantou je vypořádat se s rozdíly mezi zařízeními po technické stránce. Především by měl být využíván princip transparentní imediace (viz Bolter & Grusin, 2005): práce s testem musí být přirozená a bezprostřední, což klade velké nároky na UX. Inspiraci můžeme najít u tvůrců internetového obsahu (webových stránek, aplikací, apod.). Web developeři vcelku úspěšně řeší problém, jak se vyrovnat s různými vlastnostmi zařízení tak, aby jimi vytvářený obsah byl stále stejně použitelný, pomocí tzv. *responsive design* – rozvržení prvků obsahu se pružně přizpůsobuje použitému zařízení, přičemž tato změna není zcela jednoduchá a automatizovaná, ale spočívá v optimálním a adaptivním využití parametrů použitého zařízení (viz Marcotte, 2010). Podobně je řešitelný i problém s dostatečnou kompetencí uživatele testu – tu lze ověřit přidáním položek, které by kontrolovaly respondentovy dovednosti s ovládáním používaného zařízení. V posledních letech se navíc rozšířilo ovládání pomocí dotykových displejů, které je užitečné díky své intuitivnosti – je vhodné pro rychlou necvičenou interakci (Taveira & Choi, 2009), a tedy na jednoduchou manipulaci s prezentovaným podnětovým materiálem.

Je jisté, že v praxi bude pro zajištění správné standardizace zapotřebí využít kombinaci všech tří možností.

1. *Linear Logistic Test Model* (viz Fischer, 1973) [↑](#footnote-ref-1)
2. *Related Siblings Model* (Sinharay, Johnson & Williamson, 2003) [↑](#footnote-ref-2)
3. Slovo počítač dle mého názoru příliš evokuje klasický desktop – proto budu dále používat termín zařízení, kterým zahrnuji nejen desktopy, ale i laptopy a různé handheld přístroje, jako jsou tablety, telefony či e-inkové čtečky. [↑](#footnote-ref-3)