ePsychology: Psychometrický manuál

Verze manuálu	**0.03**
Verze aplikace	**1.00**
Autoři manuálu	**Hynek Cígler, Michal Jabůrek, Adam Ťápal**
ePsychology online, s.r.o.	**www.epsychology.cz**
Manuál ke stažení	**www.epsychology.cz/files/47/technicky_manual.pdf**

ePsycholog online s.r.o.
Vřesinská 2371/33, 708 00 Ostrava - Poruba, Česká republika

Autoři:	Hynek Cígler, Michal Jabůrek, Adam Ťápal
Název:	ePsycholog: Psychometrický manuál
Verze manuálu:	0.01
Verze dotazníku:	1.00
Datum vydání:	20. 3. 2024
Vydavatel:	ePsycholog online s.r.o.

Tuto publikaci citujte jako: Cígler, H., Jabůrek, M., & Ťápal, A. (2024). *ePsycholog: Psychometrický manuál* (0.03). ePsycholog online. www.epsycholog.cz/files/47/technicky_manual.pdf

Dotazník citujte jako: Cígler, H., Jabůrek, M., & Ťápal, A. (2024). *Screeningový dotazník ePsycholog* (1.00). ePsycholog online. www.epsycholog.cz

Obsah

Seznam obrázků a tabulek

Poděkování

Na vzniku tohoto manuálu a celého screeningového dotazníku ePsychology se podílela celá řada osob. Primárně děkujeme všem rodičům, kteří se zapojili do standardizační studie, a umožnili tvorbu dat nezbytných pro vývoj dotazníku. Díky patří také všem členům týmu ePsychology, a to včetně zaměstnanců společnosti Railsformers, kteří se podíleli na vývoji, programování, testování a dalších činnostech spojených s přípravou dotazníkové aplikace. Konečně pak děkujeme také všem odborníkům, kteří svou radou, pomocí a často i usilovnou prací přispěli k tvorbě dotazníkových otázek, zejména pak: doc. PhDr. Lence Morávkové Krejčové, Ph.D., Mgr. Zuzaně Masopustové, Ph.D., Mgr. Petře Pátkové Daňsové, Ph.D., PhDr. Kateřině Koros Bartošové, Ph.D. a Mgr. Kateřině Janasové. S numerickou optimalizací při odhadu faktorových skórů nám pomohl Ing. Vít Procházka, Ph.D.

Cíl služby

Cílem služby ePsycholog je zlepšit dostupnost psychologické péče pro děti v České republice. Konkrétně se služba snaží rodičům usnadnit rozhodování, zda se svými dětmi navštívit odborníka, a pomoci jim ve výběru konkrétního specialisty. Jádrem celé služby je on-line screeningový dotazník, který na základě výpovědi rodiče odhadne pravděpodobnost, s jakou jeho dítě trpí obtížemi v předem vybraných oblastech. Výsledkem posouzení je doporučení ohledně návštěvy odborníka. Služba přitom nenahrazuje běžnou psychologickou, případně speciálně-pedagogickou diagnostiku; bez individuálního vyšetření dítěte není v žádném případě možné stanovit příslušnou diagnózu. Dotazník však dokáže s přiměřenou mírou spolehlivosti výsledky takové diagnostiky predikovat.

Tento dokument je tzv. „psychometrický manuál“ celého testu, tedy dotazníkové části online služby ePsycholog. Jde o technickou dokumentaci vývoje, standardizace a psychometrických parametrů celé metody. Podrobná dokumentace je klíčová, protože může odborníkům sloužit k posouzení kvality měření a diagnostického nástroje jako takového. V rámci textu se proto čtenář dozví, jak probíhal vývoj dotazníkové metody služby ePsycholog, jak byly vytvářeny položky (dotazníkové otázky), a jakým způsobem bylo výzkumně ověřeno jejich fungování. Popisujeme i hrubý rámec tzv. adaptivního systému, který vybírá vhodné položky na míru rodiči (a jeho dítěti) a přizpůsobuje průběh dotazníku jeho odpovědím, čímž výrazně zrychluje a zpřesňuje měření. Konečně popíšeme také výsledné psychometrické a statistické parametry celé metody.

Úvod

Nedílnou součástí jakékoli psychodiagnostické metody je pečlivá dokumentace jejích základních psychometrických parametrů. Běžný uživatel diagnostického nástroje totiž nemá mnoho možností, jak ověřit, že služba, která je mu nabízena, je skutečně kvalitní a vědecky podložená. V první řadě jde o obecnou vlastnost psychologické diagnostiky, kdy měřené atributy (například podezření na poruchy autistického spektra nebo ADHD, tedy poruchy pozornosti a hyperaktivity) nejsou přímo viditelné, a uživatel (v tomto případě rodič) nemá možnost přímo srovnat výsledky diagnostiky s nějakou očividnou skutečností. Za druhé se do interpretace výsledků zapojuje tzv. Forerův (někdy též Barnumský) efekt[1]. Jde o přirozenou lidskou tendenci věřit výsledkům „psychologického“ měření, které popisuje jejich charakteristiky. Této víře lze vyjít snadno vstříc tím, že výsledky diagnostiky popisují charakteristiky, které jsou velmi obecné a všem lidem společné, případně udávají více možností, z nichž si čtenář vybere tu, která pro něj platí. Příkladem takového nejednoznačného a matoucího výstupu psychologického šetření by mohlo být *„vaše dítě je někdy velmi neklidné a neposedné, jindy se však dokáže soustředěně věnovat zábavné činnosti“*. Toto tvrzení popisuje

[1] Podrobnější popis jevu je k dispozici například zde: https://cs.wikipedia.org/wiki/Forer%C5%AFv_efekt.

téměř každé dítě, a lidé proto mohou mít tendenci tuto metodu hodnotit jako přesnou či spolehlivou. Toho běžně využívají astrologové, kartáři, věštitelé, a rovněž též distributoři nekvalitních, populárních psychodiagnostických metod. Do třetice pak technická dokumentace psychodiagnostických metod slouží odborníkům pro posouzení kvalit měřicího nástroje, a je proto (respektive by měla být) jejich nedílnou součástí.

Každá kvalitní metoda psychologické diagnostiky by měla být výzkumně ověřena z hlediska různých parametrů. Bohužel, nejen v České republice se často setkáváme s existencí nekvalitních, neověřených či přímo nefungujících „psychotestů“, které tak poškozují uživatele, kteří je využívají nebo za ně dokonce platí. Screeningový nástroj ePsycholog oproti tomu vznikl na základě rozsáhlého výzkumu v souladu s běžnými mezinárodními standardy (zejm. ITC a ATP, 2022; ITC, 2012; AERA, APA, a NCME, 2014). Jeho autoři a odborní garanti jsou psychologové, působí ve výzkumu na univerzitách, a jsou českou odbornou komunitou uznáváni jako odborníci na vývoj testů. To samo o sobě však pochopitelně není v žádném případě zárukou kvality. Právě z těchto důvodů dáváme veřejně k dispozici stručnou psychometrickou dokumentaci celého screeningového nástroje ePsycholog. Údaje zde uvedené nemusí být srozumitelné běžnému čtenáři z řad laické veřejnosti, to však není jeho cílem. Je pochopitelné, že pokročilejší statistické postupy nebudou zcela srozumitelné každému. Přesto jsme se celý text pokusili popsat poněkud vstřícnějším jazykem, než bývá zvykem u běžných metod psychologické diagnostiky, aby mohli alespoň rámcovou představu získat i běžní uživatelé našich služeb – tedy zejména rodiče, kteří hledají pomoc pro své děti.

Upozorňujeme čtenáře, že se jedná o první verzi psychometrického manuálu. Dosud jsme se totiž zaměřovali na vývoj samotné diagnostiky a dokumentace našeho postupu v komunikovatelné podobě nebyla prioritou. Publikaci manuálu jsme proto urychlili, aby byl k dispozici souběžně se spuštěním naší služby. Proto je tato první verze dokumentu pojata spíše stručněji. Počítáme nicméně s tím, že v dalších verzích bude manuál rozšiřován, podobně jsou plánovány doplňující publikace dílčích zjištění v odborných psychologických časopisech. Další výzkum bude realizován s využitím průběžně sbíraných dat z ostrého provozu metody ePsycholog. Také o tyto výsledky budoucích analýz budeme manuál průběžně rozšiřovat. Z těchto důvodů je v úvodu uvedena verze manuálu, která by měla být součástí každé jeho citace.

Tento technický manuál je rozdělen do dvou sekcí. V první stručně popisujeme vývoj samotné metody, tvorbu dotazníkových položek a jejich výzkumné ověřování nejdříve v rámci pilotní a posléze také standardizační studie. Popisujeme zde v hrubém rámci konkrétní analytické kroky, které jsme realizovali, a zdůvodňujeme, jakým způsobem přispívají ke kvalitě našeho měření. Následně představujeme vývoj a testování adaptivního algoritmu, který řídí průchod respondenta dotazníkem, tedy návrh jeho designu, ověřování a výsledné vlastnosti. První sekce manuálu je zakončena popisem norem a způsobu interpretace metody.

Druhá sekce tohoto dokumentu popisuje výsledné psychometrické fungování nástroje ePsycholog a jeho psychometrické parametry. Tato sekce bude průběžně doplňována a zpřesňována na základě analýz nových dat z ostrého provozu služby ePsycholog; v současné době je zčásti opřena o simulační studie založené na datech získaných odlišným způsobem.

Terminologická poznámka. Přestože je metoda určena rodičům (kteří odpovídají na položky, tedy otázky, našeho dotazníku), pro zachování jednoduchosti a přehlednosti textu často používáme obraty typu *„položky určené pro mladší děti“* namísto méně obratného, avšak přesnějšího *„položky určené pro rodiče mladších dětí“*.

Vývoj metody ePsycholog

Vývoj metody ePsycholog započal v roce 2017 a trval zhruba sedm let do spuštění první ostré verze dotazníku, která byla otevřena pro testování veřejností v roce 2024. Prvním krokem ve vývoji nástroje ePsycholog bylo vytipování oblastí, na které se metoda zaměří. Požadavkem bylo, aby (1.) šlo o psychologické, psychiatrické, speciálně-pedagogické či obdobné obtíže související buď (1a) s vývojem a zráním mozku, (1b) aktuální životní situací dítěte, (1c) sociálním zázemím rodiny, a případně (1d) spojené se školní neúspěšností. Další (2.) podmínkou bylo, aby tyto obtíže bylo možné posuzovat navrženým způsobem (tedy posouzením rodičem skrze on-line dotazník). Konečně (3.) mělo jít o potíže s dostatečnou prevalencí (výskytem), kterými trpí nezanedbatelné množství populace dětí ve věku od zhruba jednoho roku do patnácti let, a které lze (4.) zmírnit či odstranit vhodnou intervencí, terapií, či alespoň edukací rodiče či dítěte, a které je tedy vhodné a užitečné zavčas diagnostikovat. Konečný výčet vytipovaných oblastí obsahoval:

1. poruchy autistického spektra (PAS);
2. deprese;
3. úzkosti;
4. poruchy pozornosti a aktivity (ADHD a ADD);
5. poruchy čtení a psaní (dyslexie, dysgrafie, případně dysortografie, pokud bychom ji chápali jako samostatnou diagnózu);
6. poruchy počítání (dyskalkulie);
7. logopedické obtíže a poruchy řeči (zejména vývojová dysfázie);
8. opožděný psychomotorický vývoj (indikace potenciálního mentálního postižení).

Následně byla oslovena řada odborníků – psychologů, speciálních pedagogů, logopedů, dětských psychiatrů a dalších specialistů – z nichž posléze vznikla úzce spolupracující skupina odborných garantů jednotlivých oblastí. Aktuální seznam odborníků je k dispozici na webových stránkách služby

ePsychology[2], přehled všech odborníků, zapojených kdykoli během vývoje, je k dispozici v poděkování v úvodu tohoto manuálu. V této fázi vývoje byla nakonec vyřazena diagnostika opožděného psychomotorického vývoje, protože v danou chvíli nebylo pro tuto oblast možné zajistit specializovaného odborného garanta. Vzhledem k tomu, že další zařazené oblasti obsahují položky, které by mohly zachycovat dílčí symptomy opožděného psychomotorického vývoje, rádi bychom se v dalších verzích služby ePsychology této oblasti dále věnovali s možností ji posléze doplnit.

Následujícím krokem byla detailní definice každé z vytipovaných oblastí, opřená o aktuální stav vědeckého poznání a rešerši literatury. Využili jsme tzv. fasetový model (Guttman, 1959; Shye, 1978), kde je každý měřený atribut rozdělen na jednotlivé obsahové složky. Oblasti proto byly rozděleny na dílčí podoblasti a následně užší fasety podle potřeby a způsobu práce každého z odborníků (garantů dané oblasti). Oblasti v tomto kontextu chápeme zpravidla jako širší okruhy různých obtíží (např. úzkostně-depresivní poruchy), podoblasti jsou pak konkrétní diagnostické jednotky (např. generalizovaná úzkostná porucha). Fasety reprezentují dílčí projevy či skupiny symptomů dané potíží (např. potíže se spánkem, afektivní či somatické symptomy apod.).

Jednotlivé oblasti byly rozčleněny do následujících podoblastí a/či faset.

- ADHD a ADD[3]
    - Podoblasti: ADHD, ADD
    - Fasety: hyperaktivita, impulzivita, exekutivní funkce, pozornost, dráždivost, fyziologické koreláty, negativismus, poruchy regulace, rytmicita
- Úzkostně-depresivní poruchy
    - Podoblasti: depresivní fáze, porucha přizpůsobení s depresivní reakcí, generalizovaná úzkostná porucha, nespecifické příznaky deprese, nespecifické příznaky úzkosti, separační úzkostná porucha, situační faktory, fobická úzkostná porucha, obsedantně kompulzivní porucha, panická porucha, sociální úzkostná porucha
    - Fasety: afektivní příznaky, aktivita, externalizující chování, fixace, separační úzkostná porucha, inhibice, internalizující chování, kognitivní příznaky, kompulzivní jednání, nespecifické příznaky, obsedantní myšlenky, opakované ujišťování, poruchy jídla, poruchy spánku a usínání, přidružená patologie, neurotické návyky, somatické stesky, PTSD (post-traumatická stresová porucha), situační faktory, specifické fobie, sociální chování, úroveň aktivity
- PAS[4]
    - Podoblasti: *pro tuto oblast nebyly vymezeny žádné podoblasti*

[2] https://epsychology.cz/nas-tym
[3] ADHD – porucha pozornosti s hyperaktivitou (Attention Deficit and Hyperactivity Disorder). ADD – porucha pozornosti bez hyperaktivity (Attention Deficit Disorder).
[4] PAS – poruchy autistického spektra.

    - Fasety: *pro tuto oblast nebyly vymezeny žádné fasety*
- Dyskalkulie
    - Podoblasti: *pro tuto oblast nebyly vymezeny žádné podoblasti*
    - Fasety: počítání, předčíselné představy, krátkodobá paměť, osobní údaje
- Poruchy čtení a psaní
    - Podoblasti: dyslexie, dysgrafie, dysortografie
    - Fasety: vizuomotorika, grafomotorika, čtení, krátkodobá paměť, fonematické povědomí, pracovní paměť, automatizace, zraková percepce, psaní, osobní údaje
- Logopedické obtíže a poruchy řeči
    - Podoblasti: *podoblasti nejsou sdíleny z důvodu nezahrnutí oblasti do finální verze dotazníku (viz dále)*
    - Fasety: *fasety nejsou sdíleny z důvodu nezahrnutí oblasti do finální verze dotazníku (viz dále)*

Výše uvedené fasety byly využity jako východisko pro generování položek. Některé položky přitom byly sdíleny napříč jednotlivými oblastmi, typicky šlo o symptomy deprese a úzkosti, ale i některé další. Celkem vzniklo více než 470 unikátních dotazníkových položek, které následně prošly několika kroky obsahových revizí. Postup práce byl následující. Položky primárně vytvářeli garanti jednotlivých oblastí, několik málo jich následně doplnil i hlavní výzkumný tým. Členové hlavního týmu dále položky revidovali po obsahové stránce, kdy veškeré revize schvalovali garanti oblastí. V dalším kroku byly položky revidovány i po stránce konkrétních formulací a odpověďových možností tak, aby byly stylisticky podobné napříč jednotlivými oblastmi, aby nevznikaly duplicity, aby měly položky přiměřenou obtížnost (popularitu) a podobně.

Následovala fáze, ve které položky prošly velmi pečlivou jazykovou korekturou. Zároveň byla zcela vynechána oblast logopedických obtíží a poruch řeči, ve které se nepodařilo vytvořit dostatečné množství položek s odpovídající kvalitou. Vývoj této oblasti byl proto dočasně přerušen a není proto součástí stávající verze služby ePsycholog. U každé položky byla dále nastavena věková omezení (tedy věkové rozpětí posuzovaných dětí, pro které je položka určena), případná omezení školní docházkou, a také tzv. „návaznosti“ (samotná administrace některých položek je podmíněna specifickou odpovědí na některou jinou položku – například nemá smysl ptát se na školní prospěch dětí, které do školy nechodí; proto je otázka na školní docházku prerekvizitou značného množství položek). Ke změnám a určitému sjednocení došlo i v případě odpověďových možností jednotlivých položek, stejně jako u fasetové struktury celé položkové banky. Původní verze dotazníku zahrnovala pestrou strukturu odpověďových formátů položek – součástí byly položky typu „vyber vše, co platí“, případně výběr mezi různými tvrzeními (položky s nucenou volbou). Ty se většinou během pilotáže

neosvědčily a byly vyřazeny, u zbylého malého množství položek byl sjednocen formát tak, aby byl pro uživatele služby co nejpřívětivější.

Cílem celého postupu bylo dosažení co nejvyšší obsahové validity, tedy shody položek s cílem dotazníku. Věříme, že díky pečlivému postupu definice jednotlivých oblastí dotazníku ePsycholog a obsahové i formulační korektury položek má vzniklá metoda dostatečnou obsahovou validitu.

Pilotní ověření

Výsledná položková banka byla ověřena v několika na sebe navazujících fázích pilotáží. Jako první jsme realizovali tzv. **kognitivní pilotáž**, tj. všechny položky by administrované velmi malému vzorku několika desítek osob, kterých jsme se dotazovali na porozumění obsahu otázek, jednoznačnosti odpovědí a podobně. Celá kognitivní pilotáž byla vyhodnocována kvalitativně, jejím cílem bylo vylepšit znění dotazníkových položek. Na základě kognitivní pilotáže bylo také několik nejednoznačných položek odstraněno.

Druhým krokem byla **kvantitativní pilotáž**, tedy statistické posouzení všech položek. Za tímto účelem již byly vytvořeny on-line dotazníky. Vzhledem k množství položek nebylo možné administrovat všechny položky všem respondentů, vytvořili jsme proto 5 různých verzí dotazníku, které obsahovaly položky ze všech měřených oblastí a dále se lišily podle věku posuzovaného dítěte. Výběr dotazníku pro posuzovatele nebyl náhodný, respondenti si sami volili jednu z variant. Cílem bylo umožnit respondentům v případě zájmu vyplnit více dotazníků, aniž by došlo k tomu, že budou odpovídat dvakrát na stejné položky. Nábor respondentů probíhal on-line, zejména s využitím sociálních sítí, a byl podpořen reklamní kampaní. Kromě toho jsme respondentům slíbili zařazení do soutěže o drobné ceny.

Celkem byla tímto způsobem nasbírána data od 698 respondentů. Ta byla následně zpracována sérií položkových a faktorových analýz s cílem vybrat dobře fungující položky, odhadnout realizovatelnost celého výzkumného záměru (tj. tvorby výsledné podoby dotazníku), a rovněž vylepšit formulace některých položek. Součástí této fáze byly též analýzy rozdílů v odpovídání u rodičů dětí různého věku či pohlaví a odhad kriteriální validity na základě informace o udělení příslušných diagnóz odborníkem (reportovanou respondentem).

Celá položková banka byla na základě této pilotní studie rozdělena do tří „kategorií“ na položky kvality A, B, a C. Celkem 302 položek A bylo zařazeno do finální standardizace našeho dotazníku.

Standardizační studie

Samotná standardizace dotazníku byla zahájena v srpnu 2020 a trvala do konce roku, naprostá většina dat však byla získána v rozpětí srpen–říjen 2020. Naneštěstí probíhala v období pandemie

COVID-19, což se mohlo do jisté míry negativně projevit na výsledných zjištěních; domníváme se však, že efekt byl spíše malý, a v budoucnu jej budeme kontrolovat pomocí analýz provedených na nově dostupných datech.

Protože položek bylo v této fázi stále velké množství, byl zvolen design s plánovaně chybějícími daty. Celkem se standardizační studie zúčastnilo 2346 osob. Vzorek byl vytvořen příležitostně s pomocí placené reklamy na sociálních sítích, paralelně s ní jsme plánovali realizovat letákovou kampaň v ordinacích dětských lékařů, tu však zhatil nástup epidemie covid-19. Vzorek tedy není reprezentativní vůči populaci České republiky, domníváme se však, že poměrně věrně odráží populaci potenciálních uživatelů služby ePsycholog. To je mnohem zásadnější pro účely prediktivní validity, tedy předpovědi obtíží posuzovaných dětí, nereprezentativita však může limitovat interpretaci standardních T-skórů, jejichž konstrukci popisujeme dále.

Mezi respondenty, kteří dotazník vyplnili, bylo pouze 75 (3 %) mužů, průměrný věk posuzujících osob byl 33,7 let (SD = 5,7); 80 % respondentů se nacházelo v rozmezí 27–41 let. V naprosté většině šlo o vlastní rodiče dítěte; 28 respondentů (1,2 %) bylo nevlastními rodiči, 6 pěstounů či pěstounek (0,26 %), u 15 osob byl vztah k dítěti jiný (např. babička, učitel a podobně). I těchto 15 osob jsme se nakonec rozhodli ponechat v datech.

Průměrný věk posuzovaných dětí byl 5,7 let (SD = 3,4), pouze 20 % posuzovaných dětí bylo starších než 9 let. Z tohoto počtu bylo 1108 posuzovaných dívek (47 %) a 1238 chlapců (53 %). Do školy docházelo 73 % vzorku. Celkem 31 % posuzovaných dětí dříve využilo služby nějakého odborníka se vztahem k diagnostikované problematice; 478 dětí navštívilo pedagogicko-psychologickou poradnu, 186 speciálně-pedagogické centrum, 291 klinického psychologa a 168 psychiatra. Podle vyjádření rodičů byla u 111 dětí v minulosti diagnostikovaná dyslexie, u 49 dysortografie, u 72 dysgrafie, u 28 dyskalkulie, u 188 ADHD, u 92 poruchy autistického spektra (PAS), u 22 deprese, u 81 úzkosti a u 247 jiné psychické obtíže – nejčastěji šlo o vývojovou dysfázii, mentální postižení a selektivní mutismus.

Analytický postup využívající data ze standardizační studie

Na datech ze standardizační studie jsou založeny veškeré následující analýzy. Výsledný vzorek tak neobsahuje žádná data z pilotní studie (ta byla využita pouze pro zúžení položkové banky a formulační i obsahové změny v databázi položek), zároveň byla všechna data administrována tradiční formou v pevném pořadí, nikoli adaptivně. To může mít určitý vliv na výsledné parametry metody, které budou ověřeny až na datech z ostrého fungování služby ePsycholog. Adaptivní proces byl ověřen formou simulací s využitím reálných a importovaných dat, jak popíšeme níže.

Standardizační data navíc stále obsahovala položky, jejichž psychometrické parametry nebyly zcela ideální. Výsledná struktura celého dotazníku a finální seznam zařazených položek tak byly předmětem celé série analýz s řadou dílčích kroků, které zde zevrubně popíšeme.

Všechny analýzy probíhaly v prostředí R za použití řady knihoven, zejména (nikoli však výhradně) mirt, lavaan, psych, cNORM, ks, semTools, dplyr, stringr, effsize, effectsize, pROC, mice, openxlsx a parallel (Chalmers, 2012; Rosseel, 2012; Revelle, 2023; Lenhard a kol., 2018; Duong, 2022; Jorgensen a kol., 2022; Wickham, 2023; Torchiano, 2016; Ben-Shachar a kol., 2020; Robin a kol., 2011; van Buuren a Groothuis-Oudshoorn, 2011; Schauberger, a Walker, 2023; R Core Team, 2023). Odhad skórů a chyb měření individuálních respondentů byl implementován v prostředí Python 3.10 pomocí knihoven scipy (1.11.4; Virtanen a kol., 2020) a autograd[5] (1.6.2.; https://github.com/HIPS/autograd), tak, aby jej bylo možné v reálném čase používat při ostrém testování. Efektivita našeho estimátoru byla srovnávána s implementací realizovanou v R knihovně mirt (Chalmers, 2012).

Připomínáme, že žádné rozhodnutí vedoucí k vyřazení některých položek nebylo učiněno výhradně na základě statistických analýz. Ty sloužily pouze jako podklad pro další rozhodnutí, které jsme učinili na základě výzkumných výsledků, teoretických znalostí odborné literatury, a po diskuzi s odbornými garanty jednotlivých oblastí. Z těchto důvodů používáme níže zpravidla formulace jako „položky byly navrženy k vyřazení“; samotné vyřazení proběhlo v tomto případě obvykle až ve chvíli, kdy slabší fungování dané položky indikovalo více různých analýz, kdy daná položka nebyla klíčová pro zachování obsahové validity metody, a kdy vyřazení schválil garant dané oblasti.

Preliminární kroky

Deskriptivy. Prvním krokem byly pečlivé deskriptivy na úrovni jednotlivých položek. Položky s velmi vysokou prevalencí byly označeny jako potenciálně nevyhovující a navržené k vyřazení. Analyzovány byly i prevalence konkrétních odpovědí, což vedlo k různému sloučení některých odpověďových možností a dalším změnám.

Položková analýza. Druhým krokem byla tradiční položková analýza s využitím postupů klasické testové teorie (CTT). Pro každou oblast, podoblast i fasetu jsme separátně odhadli reliabilitu formou vnitřní konzistence (Cronbachova alfa a koeficient omega), korigovanou korelaci položek se škálou a ukazatele reliability po odstranění položky. Tato analýza byla realizována zvlášť pro dvě věkové kategorie i dohromady pro celý soubor. Výsledkem byla nominace některých položek k vyřazení.

Faktorová analýza. Všechny měřené oblasti byly prozkoumány pomocí metod ordinální konfirmační i explorační faktorové analýzy nad maticí polychorických korelací, a to v rámci jednotlivých faset

[5] Za optimalizaci řešení děkujeme dr. Procházkovi z Katedry pravděpodobnosti a matematické statistiky Matematicko-fyzikální fakulty Univerzity Karlovy.

i diagnostických oblastí. Položky s nízkými faktorovými náboji byly navrženy k vyřazení, zároveň byly výsledky analýzy použity jako podklad pro přesun položek mezi fasetami a oblastmi dotazníku. Výsledky také pomohly identifikovat celé fasety vhodné k vyřazení, resp. ke sloučení s dalšími (obsahově podobnými) fasetami.

Kriteriální položková analýza sloužila k ověření kriteriální validity jednotlivých položek vůči příslušným (respondety reportovaným) diagnózám. Analýza byla provedena s použitím Mannova-Whitneyho testu (případně chí-kvadrát testu testu dobré shody), ukazatelem diskriminační účinnosti každé položky bylo Cohenovo *d* a Cramerovo *V*. Položky s nízkou prediktivní účinností byly navrženy k vyřazení. Zároveň jsme identifikovali položky, které diskriminovaly i vůči jiným diagnostickým oblastem, než pro které byly původně vytvořeny. Tato analýza tak sloužila jako další podklad pro přesuny položek mezi oblastmi, stejně jako pro zařazení jedné položky do více faktorů dotazníku.

Kriteriální validita na úrovni faset a oblastí. Smysluplnost rozřazení položek do jednotlivých faset a oblastí byla potvrzena pomocí ověření kriteriální validity faset a oblastí. Pomocí t-testů (a Cohenova *d* coby velikosti účinku) byla ověřena diskriminační účinnost součtových skórů daných faset. Fasety s minimální nebo žádnou diskriminační účinností byly navrženy k odstranění a jejich položky k „rozpuštění“ v dané měřené oblasti jako celku.

Celý výše popsaný proces byl prováděn iterativně v mnoha krocích podle toho, jak byly postupně položky vyřazovány a přesouvány mezi jednotlivými dimenzemi dotazníku. Všechny změny v obsahové struktuře metody byly konzultovány s garanty jednotlivých diagnostických oblastí, kteří měli hlavní slovo ohledně výsledného obsahu „své“ oblasti, a byli zárukou obsahové validity dotazníku. Výsledkem procesu byla upravená struktura celého dotazníku se sníženým počtem položek, která byla následně testována s pomocí teorie odpovědi na položku.

Analýza s pomocí teorie odpovědi na položku (IRT)

Všechny položky byly parametrizovány pomocí dvouparametrového (2PL) IRT modelu, polytomní položky (s ordinální odpověďovou stupnicí) pomocí 2PL graded response modelu (GRM). Modely byly estimovány v R balíčku mirt pomocí EM algoritmu; vícedimenzionální modely byly odhadovány pomocí Metropolis-Hastings Robbins-Monro (MHRM) algoritmu. Shoda modelů s daty byla ověřena pomocí indexů dobré shody založené na M_2* statistice (Cai a Hansen, 2013; Maydeu-Olivares a Joe, 2006). Odhad faktorových skórů byl realizován vždy metodou maximální věrohodnosti (ML). Tato varianta nebere v potaz žádné apriorní rozdělení jako například distribuci latentního rysu. To je při adaptivním testování zpravidla méně výhodné, protože preliminární fáze může trvat déle, model později konverguje, a dochází k estimačním obtížím v uniformních odpověďových vektorech (u respondentů, kteří odpovídají systematicky stejně či podobně extrémně – tedy souhlasně či nesouhlasně – na všechny či přinejmenším většinu položek). To je důvodem, proč je pro adaptivní

testování zpravidla volena metoda EAP (expected a-posteriori) nebo MAP (maximum a-posteriori). Na druhou stranu je implementace poměrně snadná, a zejména je hlavní výhodou nezávislosti na apriorních distribučních parametrech vzorku. To je klíčové, protože nepředpokládáme, že respondenti naší metody pocházejí z jedné populace. Lze čekat, že rodiče dětí různého věku budou mít různé průměrné hodnoty odpovědí, a využívání takové obecné apriorní informace by do jisté míry mohlo zkreslovat výsledky diagnostiky.

Zatímco odhad parametrů modelů je založen na full-information přístupu a chybějící data jej příliš nelimitují, odhad shody modelu s daty byl v době vývoje dotazníku definován pouze pro kompletní dataset bez chybějících dat[6]. To byla v našem případě závažná překážka, protože některé položky byly určeny jen dětem určitého věku, zároveň byl sběr dat s ohledem na velké množství položek designován s plánovaně chybějícími daty, a nebylo tedy možné pracovat s chybějícími daty doporučenou metodou list-wise. Dřívější implementace M_2* statistiky používaly namísto chybějících dat imputované odpovědi na základě odhadnutého IRT modelu; tento postup však vede k nadhodnocení shody modelu s daty, a to zejména za situace velkého množství chybějících dat, což byl náš případ.

Vyvinuli jsme proto alternativní postup založený na imputovaných datech, avšak se škálovaným odhadem M_2* statistiky. Upozorňujeme, že tento postup není ověřený, neprošel běžným recenzním řízením, a jeho fungování není zcela jednoznačné. Výsledky je proto nutné brát jen jako orientační. Námi implementovaný postup sestával z několika kroků:

1. Imputování několika (zpravidla čtyř) datasetů s pomocí odhadnutého IRT modelu.
2. Odhad IRT modelu pro každý imputovaný dataset.
3. Odhad M_2* statistiky a indexů shody modelu s daty pro každý z odhadnutých IRT modelů.
4. Přeškálování M_2* statistiky a indexů SRMSR a RMSEA (viz níže).
5. Zprůměrování M_2* statistiky a indexů dobré shody napříč jednotlivými modely.

Škálované indexy byly získány následujícím způsobem; obě korekce jsou založeny na myšlence, že imputovaná data přispívají k celkové hodnotě M_2* statistiky plně stochasticky v souladu s modelem. V případě indexu SRMSR byla základem odhadu reziduální matice poskytovaná M_2* funkcí v balíčku mirt. SRMSR je definované jako odmocnina průměru sumy čtverců jejích hodnot. V našem případě byly vynechány hodnoty v této matici, které před imputací měly nulovou velikost vzorku (tedy páry položek, na které žádný z respondentů neodpověděl zároveň). Průměr byl následně vážený počtem respondentů, kteří na daný pár položek původně odpověděli. Rezidua párů položek s vysokým podílem imputovaných dat tak měly menší váhu při odhadu než méně imputované páry. Škálovaný

[6] V době dokončování této verze manuálu je k dispozici čerstvý preprint studie Hoovera a Thompsona (n.d.), který se však zaměřuje na modely latentních tříd (LCA), nikoli běžné IRT modely, a který postrádá softwarovou implementaci. Pro naše účely tak nebyl použitelný.

index SRMSR založený na nekompletních datech je tedy nezkresleným (ubniased) estimátorem celkového SRMSR za dvou přiměřených předpokladů: (1.) diskrepance nepozorovaných párů položek je shodná z diskrepancí pozorovaných párů a (2.) diskrepance pozorovaných párů nesouvisí s podílem chybějících dat.

V případě indexu RMSEA jsme škálovali velikost vzorku ve jmenovateli. Skutečnou velikost vzorku jsme nahradili efektivní velikostí, odhadnutou jako průměrný počet respondentů, kteří odpověděli zároveň na každý z párů položek (po vyřazení párů bez žádné společné odpovědi, protože ty přispívají k celkové M2 statistice stochasticky a nezvyšují tak diskrepanční funkci M_2–df).

Analýzu diferenciálního fungování položek (DIF) jsme realizovali metodou logistické regrese, resp. ordinální logistické regrese. Protože běžně implementované postupy pracují pouze s nominálními prediktory, vytvořili jsme vlastní, kontinuální metodu pro identifikaci DIF. Ve všech případech jsme tedy sledovali DIF podle pohlaví (chlapci vs. dívky) a podle věku dítěte (spojitý kontinuální prediktor). Pro každou položku jsme tedy sestrojili generalizovaný lineární model s logit-binomiální link funkcí; závislou proměnnou byla odpověď na danou položku, prediktory byly odhad faktorového skóre IRT modelu, pohlaví a věk dítěte. V případě, že pohlaví či věk statisticky významně a s nezanedbatelnou velikostí efektu predikovaly odpověď nad rámec odhadu faktorového skóre, byla položka shledána jako diferenciálně fungující pro děti různého pohlaví a věku. V takovém případě byla položka rozdělena na dvě položky (např. pro chlapce a dívky, případně pro starší a mladší děti, přičemž hraniční věk je variabilní podle konkrétní položky), výjimečně na tři a více (např. pro tři věkové kategorie, případně pro různě staré chlapce a dívky zvlášť). Následující kroky byly iterativně opakovány s tím, jak byly identifikovány diferenciálně fungující položky a rozdělovány do specifických položek s rozdílnými parametry. V těchto krocích jsme sledovali i to, zda lépe definované modely s méně diferenciálně fungujícími položkami budou mít vyšší kriteriální validitu (což se ve většině případů potvrdilo).

Ověření faktorové struktury uvnitř faset. Pro každou fasetu byl sestaven jednodimenzionální konfirmační IRT model. Ověřili jsme jeho shodu s daty a položky s nízkými standardizovanými faktorovými náboji jsme navrhli k vyřazení. V tomto případě byl odhad M_2* statistiky jen minimálně ovlivněn chybějícími daty, protože celá faseta byla až na výjimky vždy administrována najednou. Následně jsme realizovali DIF analýzu, sledující rozdíly v parametrech položek podle věku a pohlaví. Položky s významným diferenciálním fungováním byly navrženy na rozdělení, resp. vytvoření specifických položek pro chlapce a dívky, případně pro děti různého věku. Následně byly odhadnuty i shody empirických a modelovaných charakteristických funkcí položek, tedy tzv. shoda charakteristické funkce s daty („item fit“). Položky s nízkou shodou byly navrženy k vyřazení. Případná multidimenzionalita v rámci jednotlivých faset byla v tuto chvíli ještě jednou zkontrolována pomocí Hornovy (1965) paralelní analýzy nad maticí polychorických korelací. Konečně jsme

prozkoumali korelační matici odhadnutých faktorových skórů jednotlivých faset v rámci každé oblasti, čímž byly identifikovány fasety navržené ke sloučení.

Ověření faktorové struktury napříč fasetami uvnitř oblastí. Pro každou oblast bylo následně zkonstruováno několik modelů. V první řadě šlo o bifaktorový model s obecným faktorem a ortogonálními specifickými faktory pro každou z dimenzí – taková byla námi původně plánovaná struktura dotazníku. Tyto modely byly ve většině případů numericky nestabilní a reliabilita specifických faktorů byla zcela zanedbatelná; v některých případech docházelo k potížím s konvergencí či neúměrnou dobou estimace (v řádu vyšších desítek hodin). Následně jsme redukovali nefungující faktory a odhadli sérii redukovaných bifaktorových modelů pro každou oblast, zpravidla byl součástí i S–1 bifaktorový model[7] pro zvýšení numerické stability analýz. Nakonec jsme odhadli jednodimenzionální modely. Ty ve většině případů popisovaly data stejně dobře, případně jen nepatrně hůře ve srovnání s komplikovanějšími bifaktorovými modely. Přiklonili jsme se proto k jednodimenzionální struktuře každé z oblastí. To sice může být v rozporu s teoretickými poznatky ohledně klinického fungování dětí s určitými diagnózami, je však běžnou psychometrickou praxí. V první řadě vyšší parsiomonie (tedy statistická jednoduchost) skýtá zásadní estimační výhody a zvyšuje numerickou stabilitu všech číselných odhadů, což je vzhledem ke komplexitě našeho modelu zásadní výhodou. Druhým aspektem je fakt, že v případě posouzení jiných osob (byť velmi blízkých, jako jsou vlastní děti v naší metodě), má struktura psychologických dotazníků zpravidla jednodušší faktorovou strukturu než v případě posouzení sama sebe, pokud není posuzující „vycvičen“ vnitrosubjektové rozdíly posuzovaného dostatečně dobře vnímat. Je proto pravděpodobné, že rodiče posuzují symptomy svých dětí více jednodimenzionální ve srovnání s odborníky – psychology, psychiatry či speciálními pedagogy. I u těchto souhrnných modelů pro celé oblasti byla ověřena shoda charakteristických funkcí s daty („item fit“) a realizovali jsme DIF analýzu tak, jak je popsáno výše. Dále jsme opět ověřili kriteriální validitu celých oblastí a souvislosti s demografickými proměnnými.

Kriteriální validita oblastí. Exportované faktorové skóry z předchozího kroku byly využity pro odhad kriteriální validity. Využili jsme úsekový model, jehož prediktory byly pohlaví, věk a exportované faktorové skóry v každé oblasti zvlášť. Závislé proměnné pak tvořily diagnózy (modelované jako kategorické proměnné za využití biseriálních korelací, šlo tedy o analogii multivariační probitové regrese). Odhadli jsme kriteriální validitu pro modely s faktorovými skóry pro každou z faset zvlášť (z bifaktorových IRT modelů) i pro modely pouze s celkovým faktorovým skóre (z jednodimenzionálních IRT modelů), a ověřili inkrementální validitu specifických faktorů – tedy to,

[7] Tzv. symetrický bifaktorový model sestává z obecného faktoru, který sytí všechny položky, a ortogonálních specifických faktorů, které sytí jen vybrané položky. Každá položka je tak sycena obecným faktorem a právě jedním ze specifických faktorů. Oproti tomu S–1 bifaktorový model vynechává jeden ze specifických faktorů, část položek je tedy sycena jen obecným faktorem. Tím dosahuje vyšší numerické stability a lepší identifikaci obecného faktoru - ten je věcně ztotožněn s jedním ze specifických faktorů (Eid, 2020).

zda přispívají k predikci diagnóz (resp. ji zpřesňují) i nad rámec obecného faktoru celé oblasti. Ve většině případů se ukázalo, že specifické faktory nemají inkrementální validitu oproti celkovému skóre, což byl další argument pro využití více parsimonních modelů. S využitím těchto modelů byla realizována rovněž ROC analýza.

Souhrnný IRT model. Na základě takto vyčištěných dat jsme odhadli souhrnný IRT model pro všechny oblasti. Kvůli masivní výpočetní náročnosti nebylo možné model spustit a zejména pak vyhodnotit jeho shodu s daty na běžném počítači, a využili jsme serverový výpočet zajištěný partnery projektu, spol. Railsformers, s využitím stroje se 160 GB RAM. Tento model byl využit pro odhad parametrů položek pro všechny následující aplikace a pro prozkoumání jeho shody s daty.

Kriteriální validita. Odhadnuté faktorové skóry oblastí získané z finálního IRT modelu jsme použili pro prozkoumání prediktivní validity celého modelu (viz níže) a pro odhad parametrů pro prediktivní model používaný při vyhodnocování dotazníku (viz kapitolu Konstrukce norem a interpretace testu).

Adaptivní algoritmus

Dotazník ePsycholog využívá principů počítačového adaptivního testování (CAT) s využitím teorie odpovědi na položku. Tento postup vede zpravidla ke zrychlení doby administrace (pro dosažení shodných výsledků testování stačí administrovat méně testových položek), případně ke zvýšení reliability a validity testování.

Princip CAT je takový, že jsou v průběhu administrace respondentovi adaptivně předkládány ty položky, které o něm přinášejí nejvíce informace. Po určité úvodní fázi testování jsou odhadnuty latentní rysy (faktorové skóry) každého respondenta. Pokud má například respondent vysoký skór na škále A, není smysluplné mu dále administrovat velmi snadnou položku, u které s vysokou mírou jistoty očekáváme, že odpoví souhlasně. Protože souhlasnou odpověď dovedeme velmi přesně predikovat, samotný souhlas či nesouhlas v průměru nese velmi málo informace. Naopak, pokud bychom mu administrovali obtížnou položku, na kterou má jen 50% pravděpodobnost souhlasné odpovědi, konkrétní souhlasná či nesouhlasná odpověď je mnohem více informativní. V případě dotazníkových položek samozřejmě nejsou položky „snadné“ či „obtížné“; obtížnost si můžeme v tomto případě představit jako to, nakolik je s tvrzením reprezentovaným danou položkou pro respondenta náročné souhlasit.

Stejný princip je využit v metodě ePsycholog. Po zahájení testování jsou klientovi předloženy vybrané demografické otázky (např. věk dítěte) a následuje tzv. preliminární fáze. V této fázi není k dispozici preliminární odhad faktorového skóru, proto je potřeba vybírat následující položky jiným způsobem. V ní jsou položky vybírány do jisté míry záměrně, kromě toho však tato fáze obsahuje vysokou míru nahodilosti; vybírány jsou položky s nejvyšší informační funkcí v oblasti průměrného skóre dotazníku

s určitou mírou náhody. Výběr položek tedy probíhá tak, aby došlo k co nejrychlejší konvergenci odhadu faktorových skórů[8].

Po administraci každé položky systém zkontroluje, zda je možné odhadnout skóry na všech dimenzích dotazníku. Jakmile tato situace nastane, ukončí se preliminární fáze a nastane adaptivní fáze průchodu dotazníkem. Po zodpovězení každé předchozí položky systém odhadne míru latentního rysu a příslušnou chybu odhadu pro všechny dimenze dotazníku. Pro každou oblast zkontroluje pravidla ukončení (viz níže). Následně pro všechny dosud neadministrované položky odhadne Fisherovu informační matici. Z její diagonály jsou vynechány prvky, které odpovídají dimenzím dotazníku, které splnily pravidlo ukončení, a zbylé prvky jsou sečteny. Následně je administrována položka s nejvyšším součtem. Informační matice jsou využity pouze pro směrování adaptivního průchodu dotazníkem; pro odhad chyby měření volíme postup využívající Hessovy matice, jak je zvykem u multidimenzionálních IRT modelů[9].

Pravidla ukončení jsou tvořena sérií podmínek. Hlavním pravidlem je snížení chyby odhadu pod určitou kritickou mez, která je zvolena tak, aby reliabilita celého dotazníku dosáhla zvolených hodnot. Zároveň se kontroluje minimální a maximální přípustný počet položek pro „ukončení“ administrace dané oblasti. V poslední řadě je nastaveno globální minimum položek pro preventivní zabránění příliš rychlým, a tedy potenciálně nevalidním administracím.

Testování adaptivního algoritmu

Námi navržený adaptivní algoritmus byl doposud ověřen pouze pomocí série simulačních studií. Ověření reálných adaptivních průchodů dotazníkem, kdy jsou adaptivně administrované položky skutečně zodpovídané reálnými rodiči, bude uskutečněn až po pilotním spuštění služby ePsycholog.

Při testování adaptivního algoritmu jsme v preliminárních fázích používali zcela náhodné odpovědi. Následně jsme využili reálná data získaná v simulační studii při běžném průchodu dotazníkem (tedy neadaptivní formou). Protože jsme u žádného z respondentů neměli spolehlivě odhadnuté faktorové skóry všech dimenzí a odpovědi na položky ze všech oblastí, bylo nutné získat plný dataset bez chybějících dat a data nějakým způsobem imputovat. Běžné metody (např. multiple-imputation techniky) by byly vzhledem k velikosti datasetu výpočetně náročné a nestabilní, zvolili jsme proto

[8] Přesný postup nesdílíme, protože má jen zcela zanedbatelný vliv na samotné odhady latentních rysů (v rámci chyby měření) a není tedy podstatný pro posouzení psychometrické kvality dotazníku. Zároveň je předmětem obchodního tajemství.

[9] Přesné technické řešení numerické estimace skórů a chyb měření je výsledkem poměrně náročného vývoje. Vyzkoušeli jsme velkou řadu různých optimalizačních algoritmů s různým nastavením tak, aby výsledné odhady byly co nejvěrohodnější. Teoreticky by sice různé optimalizační postupy měly konvergovat ke stejným hodnotám, v praxi se však vyznačovaly různou mírou non-konvergence či nesmyslných odhadů, které bylo potřeba vyřešit a jejich výskyt minimalizovat. Přesný popis postupu je proto předmětem obchodního tajemství; vyžíváme však běžné nástroje v prostředí Python popsané výše.

dvoukrokový postup. Pro každého respondenta ze standardizační studie jsme odhadli faktorové skóre z oblastí, kde jsme měli k dispozici dostatečné množství položek. V prvním kroku jsme pomocí R balíčku mice (van Buuren a Groothuis-Oudshoorn, 2011) a techniky multiple imputation doplnili chybějící faktorové skóry. S pomocí těchto skórů jsme ve druhém kroku imputovali odpovědi na konkrétní položky pomocí našeho IRT modelu a charakteristických funkcí jednotlivých položek. Tím jsme získali kompletní dataset bez chybějících dat.

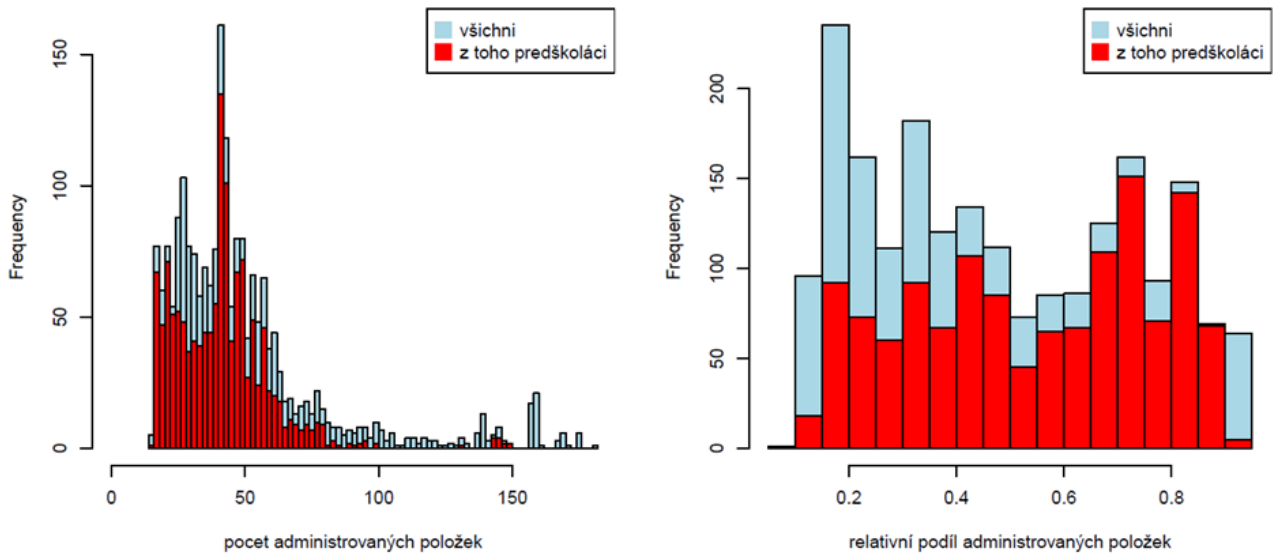
Nad tímto kompletním datasetem jsme za použití našeho IRT modelu znovu odhadli faktorové skóry všech respondentů (které se nyní v důsledku imputace mírně lišily) a ty použili jako benchmark pro naše navazující simulační studie. V nich simulování respondenti procházeli dotazníkem, ale odpovídali předem zvolenými odpověďovými možnostmi, které byly totožné s odpověďmi získanými od reálných respondentů (doplněnými o simulovaná data). Díky tomu jsme byli schopni ověřit, jak velké zkreslení nastává v důsledku adaptivního procesu oproti situaci, kdy by každý respondent obdržel všechny otázky. Současně je možné ověřit fungování námi zvolených statistických algoritmů a jejich numerické implementace.

Uvedený postup měl určité limity. V první řadě je zřejmé, že při odpovídání respondenta do určité míry záleží na pořadí položek a obsahu těch dosud zodpovězených. Lze tedy očekávat, že na určitou konkrétní položku může respondent odpovědět jinak na začátku a na konci testování. Případně se odpověď může lišit, pokud respondent dříve odpovídal na položky z jedné či jiné oblasti. Je proto nutné ověřit, zda položky nevykazují DIF v adaptivním testování ve srovnání s původní non-adaptivní standardizační studií. Za druhé může vést námi použitý postup k numerické nestabilitě odhadů faktorových skórů. Příčinou je fakt, že část odpovědí určitého respondenta je plně stochastická, resp. je zcela ve shodě s IRT modelem. Jde o položky, jejichž odpovědi původně chyběly, a byly vygenerovány ryze pravděpodobnostně za využití jejich charakteristických funkcí. Další část odpovědí je však reálná, a proto obsahuje dodatečnou nahodilost, a shoda s IRT modelem není 100%. Protože rozložení těchto dvou typů položek není náhodné, ale je systematicky rozloženo napříč jednotlivými faktory, může negativním způsobem ovlivňovat odhad faktorových skórů, a zejména jejich chyb skrze Hessovu matici. Tyto aspekty plánujeme ověřit s využitím reálných dat po spuštění pilotní verze dotazníku.

Vlastnosti adaptivního algoritmu

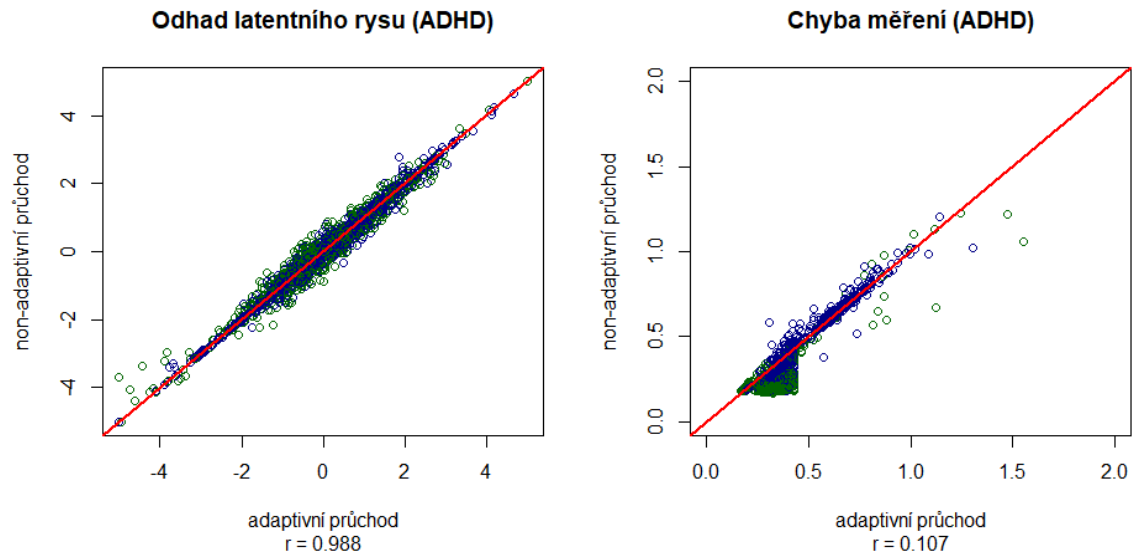
Adaptivní algoritmus přispěl k výraznému zkrácení administrace. Na obrázku 1 je patrné rozložení počtu administrovaných položek napříč více než 2000 simulovanými průchody dotazníkem (levý panel) z celkového počtu zhruba 300 položek (přesný počet se liší dle věku respondenta) a relativní úspora oproti všem dostupným položkám (pravý panel) – tedy podíl skutečně administrovaných položek vůči všem položkám dostupným pro respondenta daného věku.

Obrázek 1: Počet administrovaných položek



Adaptivní průchod měl výbornou recovery rate oproti situaci, kdy by byly administrovány všechny položky. Situaci ilustrujeme na příkladu oblasti ADHD, viz obr. 2 – pokud by rodič odpovídal stejně během adaptivní administrace jako během administrace všech dostupných položek, oba odhady by spolu korelovaly r = 0,99 – a to při v průměru 48% počtu administrovaných položek (medián 45 %). Z pravého panelu je pak patrný vliv pravidla ukončení na velikost chyby odhadu, který vytváří „zub“ pod osou prvního a třetího kvadrantu bodového grafu.

Obrázek 2: Srovnání adaptivního a non-adaptivního odhadu faktorového skóre a jeho chyby



Konstrukce norem a interpretace dotazníku

Jednotlivé skóry metody ePsycholog jsou interpretovány dvěma rozdílnými způsoby, a to buď normativně (tedy srovnáním skórů dítěte se skóry ostatních dětí ze standardizačního vzorku) za použití standardních T-skórů (s průměrem 50 a směrodatnou odchylkou 10) a percentilů, nebo kriteriálně za využití prediktivního modelu (srovnáním skórů dítěte s rodičem reportovanou diagnózou dítěte stanovenou odborníkem v praxi). Výhradně normativní interpretace je možná u diagnóz obtíží s jazykem a dyskalkulie, kde nás nedostatek dat, nízká prediktivní účinnost (zřejmě v důsledku variability reálné diagnostiky přímo v praxi) a numerická nestabilita odhadu nevedly k jistotě, že naše služba dostatečně přesně udělení diagnózy umožňuje. Jakmile budeme mít k dispozici více dat, pokusíme se i tyto oblasti o kriteriální interpretaci rozšířit.

Kriteriální interpretace testu je založena na výpočtu lineární kombinace, která využívá informace o věku a pohlaví dítěte a skóre v příslušné dimenzi dotazníku pro predikci udělení dané diagnózy s pomocí probitové regrese. Parametry modelu byly odhadnuty pomocí úsekového modelu, viz kapitolu Externí souvislosti.

Normativní interpretace je založena na věkových normách. Pro jejich konstrukci byla zvolena metoda kontinuálního normování založená na Taylorových polynomech tak, jak je implementovaná v R balíčku cNORM. Tento postup má nižší parametrickou náročnost než běžné normy, vede k vyšší numerické stabilitě, a má řadu dalších výhod. Součástí postupu je vertikální i horizontální vyhlazení skórů a tedy normalizace dat. Počet a stupně použitých polynomů byly odhadnuty pomocí cross-validace.

Psychometrické parametry metody ePsycholog

Při hodnocení psychometrických parametrů služby ePsycholog vycházíme z Messickova pojetí validity (1989, 1995). Toto pojetí je pro naše účely výhodné, protože pokrývá všechny podstatné aspekty diagnostických metod, a zároveň je založené na konstruktivistických předpokladech. Na rozdíl od realistických teorií validity (Borsboom, 2005; Borsboom a kol., 2009; Lissitz a Samuelsen, 2007; Markus a Borsboom, 2013) toto pojetí nepředpokládá, že měřený rys skutečně existuje. Takový předpoklad je totiž v případě posuzovacích škál, jako je služba ePsycholog, neudržitelný. Navíc z většiny současných výzkumných studií se zdá, že námi sledované obtíže – ADHD, PAS, specifické poruchy učení, deprese či úzkosti – nejsou monokauzální, posuzované symptomy v každé oblasti tedy nejsou způsobeny jednou společnou příčinou. Více rozebíráme tyto aspekty v kapitole Latentní procesy.

Podle Messicka je validita integrativním shrnutím způsobů, jakým je možné interpretovat testové skóry, přičemž toto shrnutí musí záviset na logické argumentaci a empirických důkazech. Tyto důkazy pocházejí z pěti hlavních oblastí:

1. **Obsah testu**
2. **Latentní procesy**
3. **Struktura testu**
4. **Externí souvislosti**
5. **Konsekvence testování**

V následujících kapitolách stručně shrneme důkazy z jednotlivých oblastí; v kapitole struktura testu navíc popíšeme výsledky analýz reliability.

Obsahová validita

Obsahová validita posuzuje, do jaké míry odpovídá obsah diagnostické metody svému účelu. Odpovídá na otázky, zda jsou testové položky dobře formulované, zda respektují aktuální teoretické znalosti, zda se ptají na relevantní symptomy.

Obsahovou validitu dotazníku jsme se snažili zajistit pomocí pečlivého výběru položek metodou dekompozice obsahového univerza a fasetových modelů každé diagnostické oblasti tak, jak je popsáno výše v kapitole Tvorba položek. Klíčovým prvkem taky byla spolupráce s garanty jednotlivých oblastí, kteří jsou odborníky na pomezí praxe a výzkumu, a v průběhu celého procesu kontrolovali dostatečnou reprezentativitu naší položkové banky. Výsledný počet položek podle jednotlivých oblastí a věku posuzovaného dítěte obsahuje tabulka 1.

Je patrné, že položky pro diagnostiku jazykových obtíží a dyskalkulie jsou určené až pro děti zhruba od šesti let věku; z toho důvodu se oblasti nevyhodnocují u mladších dětí. Vzhledem k malému počtu položek určených k měření LMP je oblast vynechána z vyhodnocení, v modelu zůstává pouze za účelem statistické kontroly všech odhadů.

Tabulka 1: Počty položek v jednotlivých oblastech

		věk (v letech)														překryvy oblastí				
oblast	položek	2	3	4	5	6	7	8	9	10	11	12	13	14	15	ADHD	depanx	PAS	čj	mat	LMP
ADHD	55	11	16	22	28	41	40	40	40	40	40	40	40	40	40	24					
depanx	142	38	41	59	61	99	98	101	101	99	97	97	97	97	97	19	96				
PAS	55	20	23	27	30	40	39	40	40	39	38	38	38	38	38	16	25	23			
čj	56	1	1	3	7	15	34	47	48	48	46	46	46	46	46	6	12	3	38		
mat	18	1	1	1	1	5	7	12	13	13	8	8	6	6	6	3	9	3	8	8	
LMP	15	0	1	1	11	12	12	12	4	4	4	4	4	4	4	4	2	2	4	2	9
celkem	279	70	79	102	126	175	192	212	205	203	195	195	193	193	193						

Poznámka. Položek – celkový počet položek; věk (v letech) – počty položek určených pro děti určitého věku; překryvy oblastí – počet položek s nenulovým faktorovým nábojem přinejmenším na obou příslušných dimenzích (na diagonále je pak počet položek s právě jedním nenulovým faktorovým nábojem na příslušné dimenzi). ADHD – poruchy pozornosti a aktivity; depanx – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; mat – poruchy počítání (dyskalkulie); LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení); celkem – celkový počet položek (z důvodu zařazení některých položek do více oblastí nemusí odpovídat součtu položek ve sloupci).

Latentní procesy

Validita latentních procesů posuzuje, do jaké míry je způsob odpovídání respondenta v souladu s teoretickými předpoklady. Jsou kognitivní procesy použité respondentem v souladu s naším očekáváním? Nedochází k nějakému systematickému zkreslení? Není respondent motivován své odpovědi upravovat v důsledku sociální žádoucnosti nebo svých vlastních záměrů?

Chování každého posuzovaného dítěte je způsobené celou řadou příčin – může jít o jeho vrozené charakteristiky, jako např. dílčí organické odlišnosti jeho mozku vedoucí k projevům atypických symptomů. Toto chování však může být způsobeno i výchovou, situačními proměnnými. Část chování reprezentuje symptomy jednotlivých oblastí, které jsou součástí našeho screeningu, nelze však říct, že jednotlivé symptomy mají jedinou společnou příčinu.

Chování dítěte sleduje rodič, který dotazník vyplňuje, a určitým způsobem jej interpretuje. Během průchodu dotazníkem mu jsou kladeny otázky na chování dítěte; rodič musí správně dotazníkové položce porozumět, vybavit si z paměti chování dítěte v příslušné situaci, a vybrat vhodnou odpověď.

Měřeným konstruktem tedy v žádném případě není „diagnóza“ či „míra obtíží dítěte“, nýbrž mentální reprezentace chování dítěte, kterou v sobě nese reportující rodič. Na míru obtíží lze z těchto dat jen nepřímo usuzovat. Z těchto důvodů se snažíme ve všech textech vyhnout diagnostice či měření obtíží; namísto toho transparentně přiznáváme, že náš screening explicitně slouží pouze pro predikci obtíží či pravděpodobnosti udělení určité diagnózy dítěti.

Během celého procesu odpovídání může docházet k řadě zkreslení. Rodič si např. nemusí konkrétní chování vybavovat – z těchto důvodů je možné (a v případě nejistoty vhodné!) libovolnou dotazníkovou položku nezodpovědět a přeskočit ji. Rodič také může záměrně klamat; ať už bagatelizovat obtíže dítěte, nebo je naopak nadhodnocovat. Rodič může otázce (či některé odpovědi) nesprávně porozumět. Z těchto důvodů je důležité, aby byly výstupy našeho dotazníku chápány jako informace pro rodiče, případně v některých situacích také pro odborníka. Pokud však odborník pozná, že je rodič motivován zkreslit výsledky našeho screeningu (v situacích, kdy je pro něj získání diagnózy zásadní, např. během občansko-právních sporů, a podobně), může být validita naší metody značně narušena. **V podobných situacích je zcela nežádoucí, aby třetí strana využila výsledky naší screeningové metody.**

Náš standardizační vzorek tvořili prakticky výhradně rodiče či zákonní zástupci dítěte (nejčastěji pak matky). Je sice možné, aby dotazník vyplnila jiná blízká osoba (například vyučující, vychovatel, prarodič a podobně), v takovém případě je však nutné výsledky interpretovat s nejvyšší opatrností, protože se mohou lišit typické motivy pro vyplnění i okruh situací, kdy dítě s respondentem přichází do kontaktu (včetně četnosti kontaktu a tudíž i příležitostí pozorovat posuzované chování).

Struktura testu

Vnitřní faktorová struktura testu byla ověřována již v průběhu vývoje tak, jak je popsáno v kapitole Vývoj metody. Zde popíšeme faktorovou strukturu výsledného modelu a různé odhady reliability.

Shoda výsledného IRT modelu s daty byla vynikající, $M_2^*(df = 32681) = 41952{,}8$, $p < 0{,}001$, $RMSEA = 0{,}012$ s $_{90\%}CI = [0{,}011–0{,}012]$, $SRMSR = 0{,}056$, $TLI = 0{,}986$. Tyto ukatatele mohou být zkreslující vzhledem k imputaci dat, přestože jsou škálované pomocí postupu popsaného výše. Výbornou shodu modelu s daty však měla i většina dílčích modelů uvnitř jednotlivých oblastí a faset, které jsou chybějícími daty ovlivněné minimálně. Součástí finální verze dotazníku rovněž nezůstala žádná položka, která by měla neakceptovatelnou shodu charakteristické funkce s daty. Domníváme se proto, že shoda našeho IRT modelu s daty je skvělá a více než excelentní pro účely odhadu faktorových skórů a jejich interpretaci.

Jednotlivé oblasti spolu přiměřeně korelovaly, viz tabulku 2, která obsahuje korelace faktorů (parametry modelu očištěné o chybu měření). Velikost těchto korelací je smysluplná a odpovídá teoretickým předpokladům.

Tabulka 2: Korelace latentních rysů (parametry IRT modelu)

Posuzovaná oblast	adhd	depanx	PAS	čj	LMP	mat
ADHD	1	–	–	–	–	–
depanx	0,560	1	–	–	–	–
PAS	0,331	0,453	1	–	–	–
čj	0,331	0,340	0,497	1	–	–
LMP	0,211	0,220	0,630	0,672	1	–
mat	0,163	0,276	0,383	0,689	0,772	1

ADHD – poruchy pozornosti a aktivity; depanx – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení); mat – poruchy počítání (dyskalkulie).

Reliabilitu testu je obtížné vzhledem k jeho adaptivní povaze posoudit. Pro její odhad není možné využít celkový IRT model, protože je v něm na jednu stranu značné množství chybějící dat, na stranu druhou respondenti vyplnili k jednotlivým oblastem prakticky všechny položky, což však nenastává při adaptivním průchodu dotazníkem.

Někteří respondenti navíc dosáhli maximálního nebo minimálního možného skóre (zodpověděli všechny položky v určité oblasti zcela souhlasně či nesouhlasně), takové odpověďové vektory není

možné vyhodnotit. V několika málo případech estimace faktorových skórů a zejména chyb odhadu selhala (a to jak v naší implementaci, tak i při odhadu skrze R balíček mirt). Parametry vzorku po vyčištění nicméně reportuje tabulka 3.

Tabulka 3: Reliabilita dotazníku a počet administrovaných položek

Posuzovaná oblast	SD	RMSE	n	M	Mde	r
ADHD	1,17	0,438	1977	12,3	11	0,861
depanx	1,14	0,556	1750	23,9	20	0,763
PAS	1,29	0,434	1797	15,3	12	0,886
čj	1,57	0,378	602	16,4	11	0,942
mat	0,94	0,515	384	5,9	5	0,697
LMP	1,14	0,865	99	2,8	1	0,426

Poznámka. SD – směrodatná odchylka odhadnutých skórů. RMSE – root mean square error (lze chápat jako „průměrnou" chybu měření). n – počet respondentů s plauzibilním odhadem faktorového skóre (celkové $N = 2058$). M, Mde – průměr a medián počtu administrovaných položek s nenulovým nábojem na dané oblasti. *r* – odhad empirické reliability adaptivního průchodu. ADHD – poruchy pozornosti a aktivity; depanx – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; mat – poruchy počítání (dyskalkulie); LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení).

Reliabilita jednotlivých dimenzí (vyjma LMP, která není použita pro diagnostické závěry) se pohybuje v rozmezí 0,697–0,942 s průměrem 0,830. Tyto hodnoty jsou podle nás více než dostatečné pro screeningové účely.

Externí souvislosti

Externí souvislosti odpovídají dřívějšímu pojetí empirické validity, případně konstruktové validity podle Cronbacha a Meehla (1955), která umísťuje měřený atribut do tzv. nomologické sítě. Validita ve smyslu externích souvislostí popisuje, zda testové skóry přiměřeně korelují s jinými testovými metodami, případně zda dobře predikují externí kritérium. Tato oblast důkazů ověřuje tzv. konvergentní a divergentní validitu (tedy korelace s testy měřícími stejné a rozdílné atributy), diferenciální validitu (v našem případě schopnost rozlišit jednotlivé diagnózy), případně kriteriální a prediktivní validity (predikovat existující kritérium, případně kritérium, které nastane v budoucnosti).

V případě služby ePsycholog bohužel nemáme srovnání výsledků našeho screeningu a následného posouzení odborníkem; tvorba takových dat by byla neúměrně drahá, a tyto informace budeme

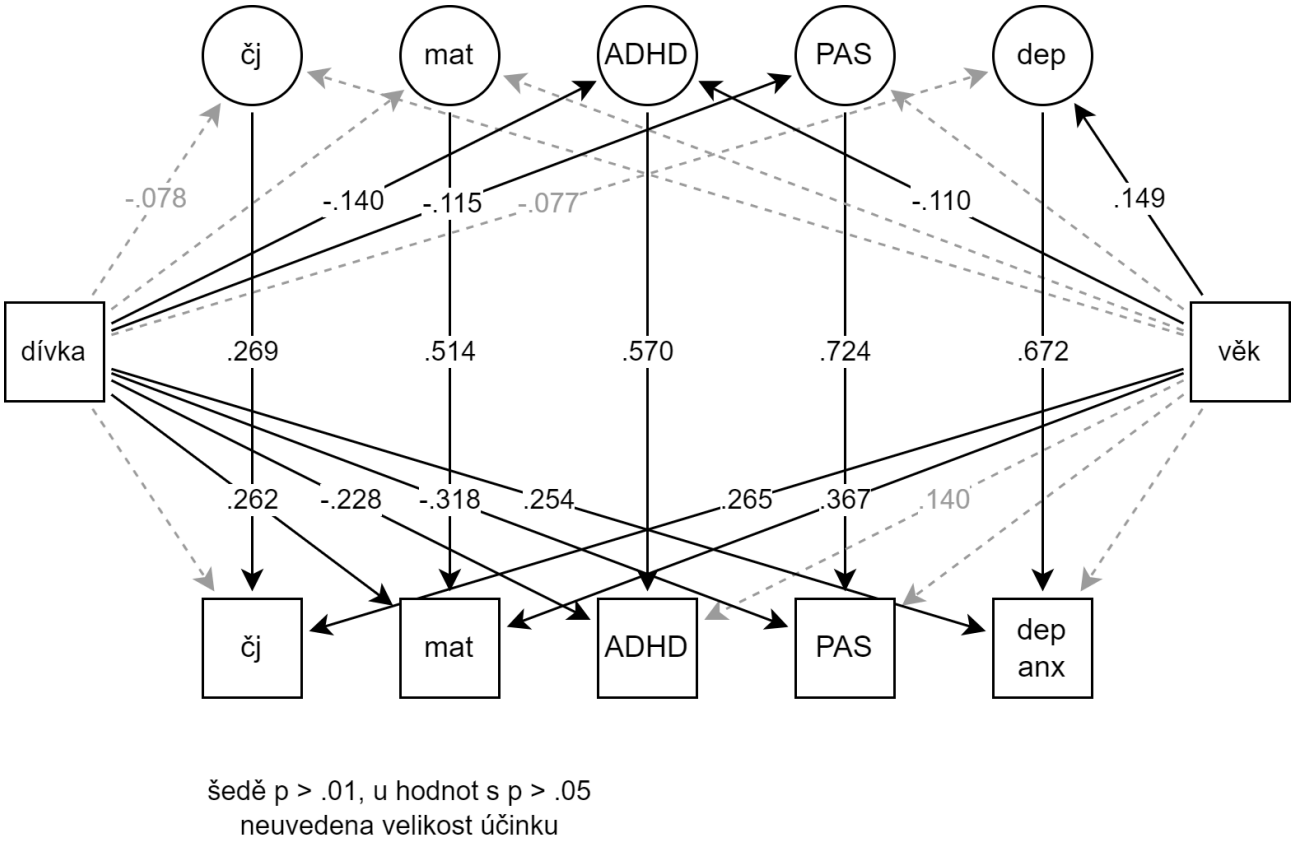
sledovat po pilotním spuštění naší služby. Za tímto účelem jsme připravili dotazníky pro odborníky, které budeme prostřednictvím rodičů průběžně sbírat a vyhodnocovat, a na jejich základě upravovat parametry naší screeningové metody a její vyhodnocování. Aktuálně však máme k dispozici informaci poskytnutou rodičem, kde a kým byla udělena jaká diagnóza. Tyto informace jsou pro nás klíčové, protože jsou základem prediktivního modelu použitého pro náš screeningový závěr.

Během vývoje služby jsme vyzkoušeli větší množství úsekových a strukturních modelů, zde reportujeme jen finální model použitý pro predikci diagnóz. Reálně jde však o dva modely – jeden pro předškoláky, druhý pro děti, které již navštěvují školu. Původní záměr byl využít veškeré skóry dotazníku pro predikci všech diagnóz. Tento model však nefungoval významně lépe ve srovnání s více parsimonním modelem, kde je každá z diagnóz predikovaná pouze „příslušným“ faktorovým skóre. To je důkazem vysoké diskriminační validity našeho dotazníku. Součástí modelů byl dále ještě věk a pohlaví, které jsou pro vyhodnocení dotazníku využívány.

Všechny udělené diagnózy dyslexie, dysgrafie či dysortografie byly sloučeny tím způsobem, že proměnná byla kódovaná 1, pokud dítě má alespoň jednu z udělených diagnóz, jinak byla kódovaná 0. Stejný postup byl zvolen pro diagnózy úzkosti a deprese. Model byl odhadnut metodou robustních diagonálně vážených čtverců (WLSMV), chybějící data byla ošetřena metodou pair-wise. Pokud dítě v minulosti nenavštívilo žádného odborníka, informace o diagnózách byla překódovaná na nulu.

Model pro školní děti popsal data skvěle, dokonce se neodlišoval od dat, $\chi^2(df = 20) = 22{,}9$, $p = 0{,}295$, $CFI = 0{,}997$, $TLI = 0{,}992$, $RMSEA = 0{,}014$ s $_{90\%}CI = [0–0{,}036]$, $SRMR = 0{,}037$. To znamená, že povolení dodatečných regresních cest (například predikce diagnózy ADHD faktorovým skórem PAS) by nevedla k jeho dalšímu zlepšení. Reziduální kovariance jednotlivých diagnóz byly vesměs nízké a nesignifikantní, výjimkou je poměrně očekávaná reziduální závislost specifických poruch učení v češtině a matematice, $r = 0{,}720$, dále pak PAS a úzkostně-depresivních obtíží, $r = 0{,}666$, PAS a ADHD, $r = 0{,}654$, a konečně PAS a obtíží českého jazyka, $r = -0{,}456$. Je potřeba si uvědomit, že počet dětí s diagnózou byl poměrně malý, a tedy vysokou reziduální kovarianci mohly způsobit jednotky dětí, které měly obě diagnózy udělené společně. Výsledný model (s vynecháním reziduálních kovariancí) je uvedena na obrázku 3.

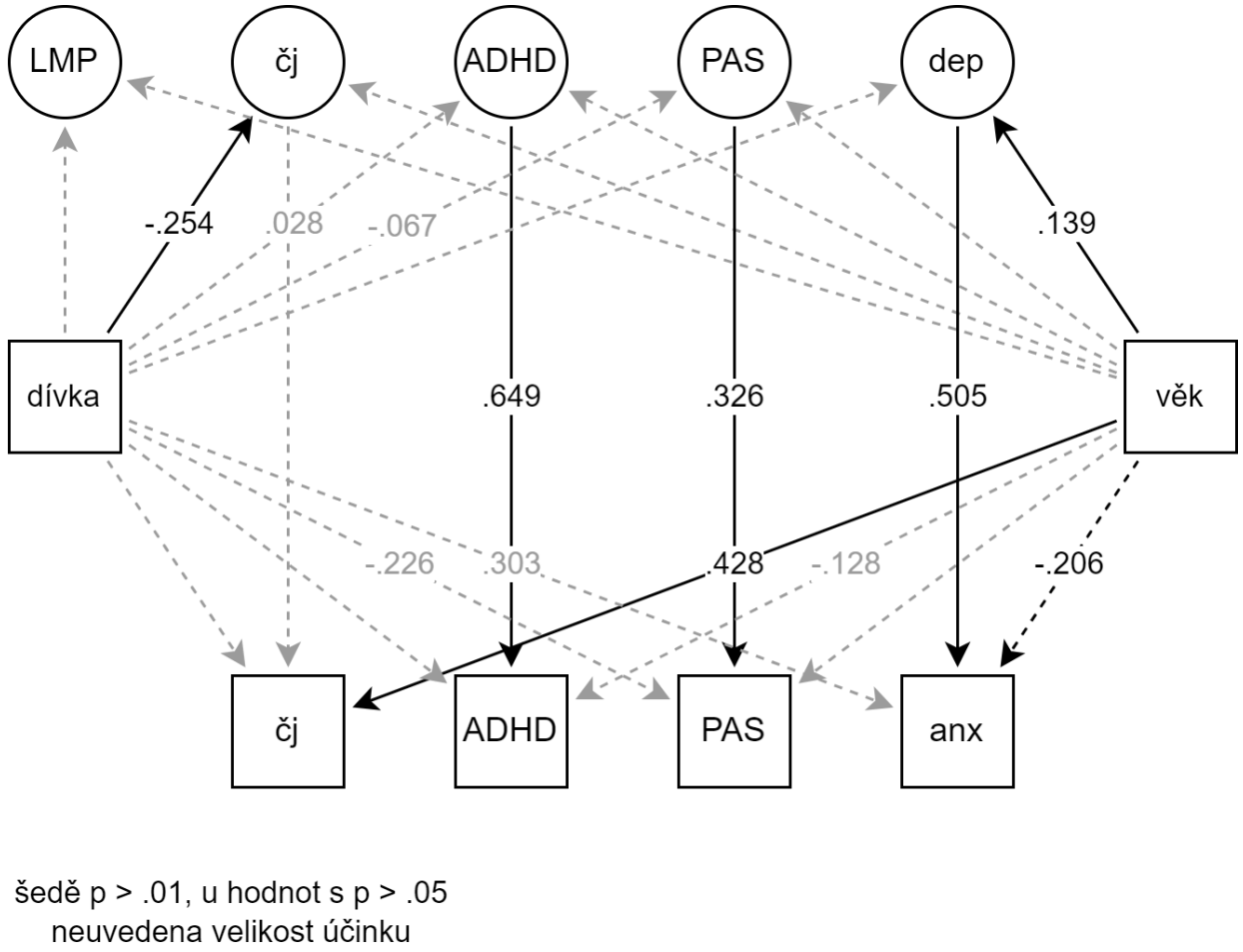
Obrázek 3: Úsekový model kriteriální validity pro školní děti



Poznámka. ADHD – poruchy pozornosti a aktivity; depanx – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; mat – poruchy počítání (dyskalkulie); LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení); dívka – pohlaví dítěte (referenční kategorie byli chlapci).

Rovněž i model pro předškolní děti fungoval skvěle a statisticky významně se nelišil od dat, $\chi^2(16) = 10{,}0$, $p = 0{,}864$, $CFI = 1{,}000$, $TLI = 1{,}044$, $RMSEA = 0{,}000$ s $_{90\%}CI = [0–0{,}014]$, $SRMR = 0{,}058$. Tento model navíc obsahoval i faktorové skóre LMP, pouze jako statistickou kontrolu bez vlivu na jeho interpretaci. Protože žádné předškolní dítě nemělo diagnostikovanou depresi, v modelu figuruje samostatně diagnóza depresivity. Ze stejného důvodu byla z modelu odstraněna diagnóza dyskalkulie a rovněž i faktorový skór dyskalkulie, protože děti neměly zpravidla dostatek položek pro jeho odhad. V tomto modelu byly všechny reziduální kovariance zanedbatelné ($|r| < 0{,}42$) a nesignifikantní. Diagram modelu je zobrazen na obrázku 4.

Obrázek 4: Úsekový model kriteriální validity pro předškolní děti



Poznámka. ADHD – poruchy pozornosti a aktivity; dep – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení); dívka – pohlaví dítěte (referenční kategorie byli chlapci).

Parametry uvedených strukturních modelů byly použity pro tvorbu kompozitního skóre (lineární kombinace) a byly využity pro predikci jednotlivých diagnostických oblastí pomocí ROC analýzy. Výsledky obsahuje tabulka 4. Je patrné, že plocha pod křivkou (AUC) byla vynikající a pro screeningové účely zcela dostatečná. V každé z oblastí (vyjma poruch čtení a psaní u předškoláků, kde jsme neměli dostatek dat) jsme rovněž identifikovali dva kritické skóry signalizující vysoké a nízké riziko výskytu dané diagnózy. Pro obě tyto hranice reportujeme senzitivitu a specificitu. Tabulka konečně obsahuje i informace o způsobu interpretace daného skóre – tedy normativní (za pomoci percentilu a T-skóre), případně kriteriální (za pomoci prediktivního modelu).

Tabulka 4: ROC Analýza, senzitivita a specificita

Školáci						
		mírné podezření		výrazné podezření		
	AUC	spec	senz	spec	senz	interpretace
ADHD	0,804	0,751	0,759	0,938	0,509	predikce
PAS	0,888	0,851	0,813	0,933	0,604	predikce
čeština	0,703	0,709	0,628	0,866	0,398	normy
dyskalkulie	0,880	0,820	0,857	0,973	0,714	normy
depanx	0,836	0,709	0,628	0,890	0,345	predikce

Předškoláci						
		mírné podezření		výrazné podezření		
	AUC	spec	senz	spec	senz	interpretace
ADHD	0,845	0,642	0,914	0,925	0,486	predikce
PAS	0,775	0,787	0,722	0,929	0,417	predikce
čeština	0,820	–	–	0,745	0,800	–
úzkost	0,691	0,885	0,455	0,949	0,273	predikce

Poznámka. Kritický skór nízkého rizika byl zvolen s ohledem na maximalizaci senzitivity, tedy vyloučení faktu, že dítě nebude identifikováno jako potenciálně rizikové v situaci, kdy ve skutečnosti rizikové je. Naopak cut-off vysokého rizika byl zvolen s maximalizací specificity; tedy vyloučení falešně negativního závěru a negativního screeningového závěru v situaci, kdy dítě ve skutečnosti obtížemi trpí. AUC – plocha pod křivkou; spec – specificita; senz – senzitivita; Interpretace – způsob interpretace (predikce – prediktivní, normy – normativní, „–“ – není součástí výstupu, ale v modelu je informace brána v potaz). ADHD – poruchy pozornosti a aktivity; depanx – úzkostné a depresivní poruchy; PAS – poruchy autistického spektra; čj – poruchy čtení a psaní; mat – poruchy počítání (dyskalkulie); LMP – opožděný psychomotorický vývoj (není součástí vyhodnocení).

Konsekvence testování

Konsekvence testování popisují důsledky, které s sebou diagnostika nese. Může jít o dopady určitých diagnostických rozhodnutí, udělení či naopak neudělení diagnózy. Relevantní jsou též politické a společenské dopady samotného procesu testování či etika celého procesu.

Pro úvahu o konsekvencích testování bohužel nedisponujeme daty. Lze nicméně očekávat, že neposkytnutí adekvátní péče dětem se specifickými potřebami a psychickými obtížemi může mít

dopady na jejich well-being i vzdělávací výsledky. Pokud tedy aplikace ePsycholog přispěje k časnější diagnostice obtíží, mohla by navýšit i pravděpodobnost získání adekvátní péče.

Naopak značným rizikem jsou důsledky chyb při diagnostice. Falešně negativní závěr (tedy negativní výsledek screeningu u dítěte, které ve skutečnosti určitými obtížemi trpí) by mohl rodiče podpořit v rozhodnutí, že je návštěva odborníka zbytečná, a oddálit okamžik, kdy bude dítěti poskytnuta adekvátní péče. Riziko vzniku této situace jsme se snažili zredukovat tím, že jsme nastavili poměrně vysokou citlivost (senzitivitu) našeho dotazníku, a dostatečné procento dětí je proto identifikováno pomocí závěru „mírné podezření".

Současně lze však očekávat negativní dopady falešně pozitivních závěrů, kdy dítě ve skutečnosti obtíže nemá, ale naše aplikace návštěvu odborníka doporučí. V tomto případě lze uvažovat o rizicích spojených s fenoménem sebenaplňujícího se proroctví, stigmatizací či zvýšenými obavami na straně rodiče. Tuto situaci jsme se snažili naopak minimalizovat tím, že závěr „výrazné podezření" byl konstruován s ohledem na specificitu, a tedy minimalizaci falešně negativních případů. Veškeré výstupy naší aplikace a texty s ní spojené současně co nejvíce zdůrazňují fakt, že jde o prostý screening. Negativní dopady může redukovat i nabídka okamžité telefonické konzultace, kterou poskytujeme.

Domníváme se, že naše aplikace může být užitečná i v širším společenském kontextu. Protože podíl dětí s psychickými obtížemi v populaci neustále roste, může být naše služba dalším podnětem, který zaměří pozornost veřejnosti na děti se specifickými potřebami.

Podrobnější vyhodnocení konsekvencí testování bude realizováno s využitím dat z ostrého provozu naší služby a bude doplněno v dalších verzích tohoto manuálu.

Reference

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

Ben-Shachar, M., Lüdecke, D., & Makowski, D. (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software, 5*(56), 2815. https://doi.org/10.21105/joss.02815

Borsboom, D., Cramer, A. O. J., Kievit, R. A., Scholten, A. Z., & Franic, S. (2009). The End of Construct Validity. In R. W. Lissitz (Ed.), *The Concept of validity: Revisions, new directions, and applications* (pp. 135–170). Information Age Publishing.

Borsboom, D. (2005). *Measuring the Mind.* Cambridge University Press.

van Buuren, S., Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), 1-67. https://doi.org/10.18637/jss.v045.i03

Cai, L., & Hansen, M. (2013). Limited-information goodness-of-fit testing of hierarchical item factor models. *British Journal of Mathematical and Statistical Psychology, 66*(2), 245–276. https://doi.org/10.1111/j.2044-8317.2012.02050.x

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281–302. https://doi.org/10.1037/h0040957

Duong, T. (2022). *ks: Kernel Smoothing.* R package. https://CRAN.R-project.org/package=ks

Eid, M. (2020). Multi-Faceted Constructs in Abnormal Psychology: Implications of the Bifactor S - 1 Model for Individual Clinical Assessment. *Journal of Abnormal Child Psychology, 48*, 895-900. https://doi.org/10.1007/s10802-020-00624-9

Guttman, L. (1959). Introduction to Facet Design and Analysis. In *Proceedings of the Fifteenth International Congress of Psychology* (pp. 130-132). Amsterdam: North Holland. https://doi.org/10.1016/0001-6918(59)90023-X

HIPS (2023). Autograd: Efficiently computes derivatives of numpy code. Github. https://github.com/HIPS/autograd

Hoover, J.C., & Thompson, W. J. (n.d.). Modifying the M2 Statistic to Handle Missing Data. https://dynamiclearningmaps.org/modifying-m2-statistic-handle-missing-data

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. Psychometrika, 30, 179–185. http://www.ncbi.nlm.nih.gov/pubmed/14306381

Chalmers, R.P. (2012). mirt: A Multidimensional Item Response Theory Package for the R Environment. *Journal of Statistical Software, 48*(6), 1-29. https://doi.org/10.18637/jss.v048.i06

International Test Commission and Association of Test Publishers (2022). *Guidelines for technology-based assessment.* https://www.intestcom.org/page/28

International Test Commission (2012). *International Guidelines on Quality Control in Scoring, Test Analysis, and Reporting of Test Scores.* https://www.intestcom.org/page/17

Jorgensen, T. D., Pornprasertmanit, S., Schoemann, A. M., & Rosseel, Y. (2022). *semTools: Useful tools for structural equation modeling.* R package. https://CRAN.R-project.org/package=semTools

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A Continuous Solution to the Norming Problem. *Assessment, 25*(1), 112–125. https://doi.org/10.1177/1073191116656437

Lissitz, R. W., & Samuelsen, K. (2007). A Suggested Change in Terminology and Emphasis Regarding Validity and Education. *Educational Researcher, 36*(8), 437–448. https://doi.org/10.3102/0013189X07311286

Markus, K. A., & Borsboom, D. (2013). *Frontiers of Test Validity Theory.* Routledge.

Maydeu-Olivares, A., & Joe, H. (2006). Limited information goodness-of-fit testing in multidimensional contingency tables. *Psychometrika, 71*(4), 713–732. https://doi.org/10.1007/s11336-005-1295-9

Messick, S. (1989). Meaning and Values in Test Validation: The Science and Ethics of Assessment. *Educational Researcher, 18*(2), 5–11. https://doi.org/10.3102/0013189X018002005

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist, 50*(9), 741–749. https://doi.org/10.1037/0003-066X.50.9.741

R Core Team (2023). *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software, 48*(2), 1-36. https://doi.org/10.18637/jss.v048.i02

Revelle, W. (2023). *psych: Procedures for Psychological, Psychometric, and Personality Research.* Northwestern University, Evanston, Illinois. https://CRAN.R-project.org/package=psych

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C., & Müller, M. (2011). pROC: an open-source package for R and S+ to analyze and compare ROC curves. *Bioinformatics, 12*, 77. https://doi.org/10.1186/1471-2105-12-77

Shye, S. (1978). *Theory construction and data analysis in the behavioral sciences.* Jossey-Bass.

Schauberger, P., Walker, A. (2023). *openxlsx: Read, Write and Edit xlsx Files.* R package. https://CRAN.R-project.org/package=openxlsx

Torchiano, M. (2016). Effsize - a package for efficient effect size computation. *Zenodo.* https://doi.org/10.5281/zenodo.1480624

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., ... & van Mulbregt, P. (2020). SciPy 1.0 Contributors. SciPy, 1, 261-272.

Wickham, H., François, R., Henry, L., Müller, K., & Vaughan, D. (2023). *dplyr: A Grammar of Data Manipulation.* R package. https://CRAN.R-project.org/package=dplyr

Wickham H (2022). *stringr: Simple, Consistent Wrappers for Common String Operations.* R package https://CRAN.R-project.org/package=stringr