STUDIA PAEDAGOGICA

Ročník 27, číslo 1, 2022
https://doi.org/10.5817/SP2022-1-4

HODNOCENÍ BAKALÁŘSKÝCH PRACÍ JAKO PŘIJÍMACÍ KRITÉRIUM DO NAVAZUJÍCÍHO MAGISTERSKÉHO STUDIA: PSYCHOMETRICKÁ KAZUISTIKA

BACHELOR'S THESIS RATING AS AN ADMISSION CRITERION TO THE MASTER'S PROGRAM OF PSYCHOLOGY: A PSYCHOMETRICAL CASE STUDY

Hynek Cígler[a], Stanislav Ježek[a], Jan Širůček[a], Lenka Lacinová[a]

[a] Fakulta sociálních studií, Masarykova univerzita

ABSTRAKT

V roce 2020 epidemie covid-19 znemožnila konání přijímací zkoušky do navazujícího magisterského studia psychologie na Masarykově univerzitě formou běžného písemného znalostního testu. Zvolili jsme proto náhradní formu v podobě hodnocení bakalářských prací. Tento článek je „psychometrickou kazuistikou", která dokumentuje celý postup od tvorby hodnoticích kritérií s ohledem na obsahovou validitu přes volbu designu až po výsledky. Každou práci hodnotili dva náhodně vylosování hodnotitelé, jejichž „přísnost" byla vyvážena využitím logistického lineárního testového modelu (LLTM) v rámci paradigmatu teorie odpovědi na položku (IRT). Výsledkem byl jednodimenzionální skór, férově vyvážený napříč dvěma přijímacími termíny ($n_1 = 82$, $n_2 = 48$) a 18 hodnotiteli. Jeho reliabilita, $r_{xx'} = 0{,}869$, byla srovnatelná s písemnými testy v jiných letech. Přísnost hodnotitelů a obtížnost kritérií se nelišily napříč oběma termíny, výsledné hodnocení se zdá být validním ukazatelem a srovnatelně férové s písemným testem. Navržená procedura tak může posloužit jiným pracovištím a k jiným účelům, než je jen přijímací test; z těchto důvodů sdílíme analytický skript a veškeré materiály nezbytné pro replikaci celé procedury.

KLÍČOVÁ SLOVA

lineární logistický testový model; LLTM; high-stake testy; hodnoticí škála; přijímací test; teorie zobecnitelnosti

ABSTRACT

The COVID-19 pandemic influenced admissions testing for the master's degree program of psychology at Masaryk University in 2020. The administration of the standard paper-and-

pencil knowledge test was not possible; therefore, we chose the bachelor's thesis ratings instead. This paper is a psychometrical case study that covers the development of criteria related to content validity, design selection, and results. Two randomly selected raters evaluated each thesis, and we equated their severity using a linear logistic test model (LLTM) under the item response theory (IRT) paradigm. This procedure resulted in unidimensional and unbiased scores equated across 18 judges and 2 terms ($n_1 = 82$, $n_2 = 48$). The reliability was comparable to the standard tests, $r_{xx'} = 0.869$, and judge severity and criteria difficulty did not differ across them. The resulting ratings seem to be valid and no less fair than the written exam. The proposed method can serve other departments and other goals, not only as an entrance test. We share an analytical script and all the necessary materials to enhance the use of this method.

KEYWORDS

Linear Logistic Test Model; LLTM; high-stakes testing; rating scale; entrance test; generalizability theory

KORESPONDUJÍCÍ AUTOR
Hynek Cígler, Katedra psychologie, Fakulta sociálních studií,
Masarykova univerzita, Joštova 10, 602 00 Brno, Česká republika
e-mail: cigler@fss.muni.cz

Úvod

V České republice došlo 11. března 2020 k uzavření vysokých škol v důsledku vysoce infekčního respiračního onemocnění covid-19 způsobeného koronavirem SARS-CoV-2. Na Fakultě sociálních studií Masarykovy univerzity (FSS MU) platil i po zbytek semestru zákaz prezenční výuky a jiných setkání se studenty tváří v tvář. Toto omezení se týkalo rovněž prezenční účasti uchazečů o studium na přijímacích zkouškách (PZ) do navazujícího magisterského studia psychologie (NMgr.). V relativně krátkém čase tak bylo nutné zajistit alternativní a dostatečně objektivní podobu PZ, která by naplňovala běžné psychometrické standardy. Katedra psychologie FSS MU zvolila jako přijímací kritérium hodnocení bakalářských prací. Předložený článek je „psychometrickou kazuistikou" zvoleného postupu. Ten je popsán od prvních příprav až po výsledné řešení a dává k dispozici veškeré potřebné informace: zvolená kritéria včetně jejich podrobné argumentace, anonymizovaná data i analytické skripty. Výsledné hodnocení totiž nebylo založeno na prostém součtu bodů, ale vyvažovalo přísnost jednotlivých hodnotitelů prostřednictvím raschovského logistického lineárního testového modelu (LLTM) v rámci teorie odpovědi na položku. Věříme, že výsledný formát PZ je nejen vysoce funkční, ale že může sloužit i jako návod pro jiná vysokoškolská pracoviště – a to nejen v situaci ohrožení zdraví osob pandemií koronaviru.

1. Kontext

1.1 Běžná podoba PZ a charakteristika studia

Do NMgr. programu psychologie na FSS MU se mohou přihlásit uchazeči, kteří z psychologie získali bakalářský diplom; důvodem tohoto omezení je fakt, že magisterský diplom z psychologie opravňuje v České republice absolventa k výkonu vázané živnosti psychologické poradenství a diagnostika a je podmínkou pro další vzdělávání a následnou pozici klinického psychologa jako výkonu nelékařské zdravotnické profese.

Samotné přijímací řízení do NMgr. psychologie se v běžné situaci sestávalo z 60položkového znalostního testu s časovým limitem 60 minut. Položky jsou ve formátu multiple-choice s jednou správnou odpovědí, za chybné odpovědi se neodečítaly žádné body. Celkem tak bylo možné získat 0–60 bodů.

Test se konal dvakrát ročně, odděleně pro studenty s nástupem do jarního a podzimního semestru, a jeho vyhodnocení bylo normativní – k přijetí bylo navrženo určité množství uchazečů s největším počtem bodů dle kapacity oboru. Reliabilita testu, odhadnutá pomocí Bentlerova koeficentu r_{glb} (Bentler, 2009; Bentler & Woodward, 1980; Revelle & Zinbarg, 2009; Sijtsma, 2009), se pohybovala v rozmezí 0,90–1,00 s průměrem 0,95; Cronbachova alfa, která je ovšem podhodnocena v důsledku předpokladů tau-ekvivalence a jednodimenzionality, se nacházela v rozmezí 0,78–0,98 s průměrem 0,86. Faktorová struktura byla na základě Hornovy paralelní analýzy zpravidla jednodimenzionální nebo dvoudimenzionální, v takovém případě byl sekundární faktor tvořený většinou metodologicko-statistickými položkami. Příloha 2 v online suplementu Cígler et al. (2020) poskytuje podrobnější informace o psychometrických parametrech testů.

1.2 Alternativní podoba PZ

Katedra psychologie po uzavření vysokých škol zvažovala různé alternativní podoby PZ, mj. osobní pohovory pomocí videokonference, přijetí všech uchazečů (a jejich následnou selekci během prvního semestru prostřednictvím vybraných kurzů s výrazně zvýšenou náročností), hodnocení motivačního dopisu apod. Podoba přijímacího řízení na univerzity, zejména na zdravotnické obory, je přitom častým tématem výzkumných studií. V České republice je navíc k dispozici monografie Charváta a Viktorové (2014). Během přijímacího řízení mohou být hodnoceny kognitivní i nekognitivní charakteristiky uchazečů ověřované prostřednictvím tradičního testu s uzavřenými či otevřenými otázkami, esejemi či jinými psanými texty, nestrukturovanými či polostrukturovanými pohovory (zpravidla před komisí, případně v podobě většího množství individuálních „minipohovorů"), motivačními dopisy či jinou formou individuálních prohlášení. Důležitým zdrojem informací může být i analýza předchozích výsledků – studijního průměru, prací, stáží apod. (Salvatori,

2001; Zamanzadeh et al., 2020). Ověřování nekognitivních aspektů a využívání osobních pohovorů je nicméně do jisté míry kontroverzní; ukazatele validity se liší napříč studiemi a výsledek může být z různých důvodů problematický, neférový a méně validní (Salvatori, 2001; Zamanzadeh et al., 2020).

Během naší rešerše jsme nicméně nenarazili na žádnou studii, ve které by byla jako přijímací kritérium využita předchozí bakalářská práce. Příčinou může být fakt, že většina studií se zaměřuje na přijímací řízení do bakalářských studijních programů, kde předchozí absolventská práce většího rozsahu není k dispozici. Na rozdíl od České republiky rovněž v řadě západních zemí vůbec není tradicí vypracování rozsáhlé bakalářské (a obecně diplomové) práce.

Obecně se pak soudí, že kombinace více různých zdrojů informací zvyšuje validitu přijímacího řízení (AERA et al., 2014; Zamanzadeh et al., 2020). Nejlepším prediktorem budoucího studijního úspěchu bývá předchozí studijní výsledek, operacionalizovaný zpravidla jako Grade Point Average (GPA). Ten však může být neférový vůči absolventům náročných škol; jeho validita je navíc zpravidla ověřována vůči budoucímu GPA na univerzitě, a je proto otázkou, zda by měl podobnou validitu i vůči jinak definovaným kritériím (Charvát & Viktorová, 2014; Salvatori, 2001).

V našem případě nepřipadala v úvahu žádná forma hromadně administrovaného písemného testu. Nechtěli jsme se též spoléhat na jakoukoli formu individuálních pohovorů, ověřování nekognitivních aspektů, motivace či osobnostních charakteristik. Tyto varianty byly vyhodnoceny jako potenciálně neférové, nevalidní a s nízkou reliabilitou při přiměřených nárocích na pracovníky katedry; jejich zevrubnější diskuze by však byla rozsáhlou odbočkou od hlavního tématu článku.

Jako nejvýhodnější bylo nakonec zvoleno hodnocení bakalářské práce. Domnívali jsme se, že je u něj možné zajistit největší objektivitu, férovost a reliabilitu ve smyslu shody posuzovatelů. Zároveň jsme toho názoru, že kvalitní bakalářská práce dobře odráží schopnost orientovat se v současné odborné literatuře, formulovat relevantní výzkumné otázky a zasadit je do teoretického rámce, navrhnout a realizovat výzkumný záměr a v neposlední řadě odborně a kriticky uvažovat nad výzkumnými zjištěními. Tyto kompetence považujeme za klíčové předpoklady ke studiu v NMgr. programu Psychologie na FSS MU. Psané eseje navíc mají uspokojivou prediktivní validitu vůči budoucímu studijnímu úspěchu srovnatelnou s písemnými testy, a to okolo $r = 0{,}5$, což je srovnatelné nebo nepatrně nižší ve srovnání se standardizovanými znalostními testy (Salvatori, 2001). Lze uvažovat, že bakalářská práce může mít prediktivní validitu ještě vyšší oproti relativně krátké eseji.

Hlavní nevýhodou hodnocení bakalářských prací je však „nemožnost opravy" – při opakovaném pokusu je totiž hodnocen stále stejný text. Pro širší perspektivu je nicméně vhodné uvést, že opakované absolvování přijímacího řízení je problém i z hlediska znalostních testů, u kterých opakováním roste pravděpodobnost přijetí díky pozitivní chybě měření.

Zásadním úkolem tedy bylo navrhnout takovou podobu hodnocení bakalářských prací, která by zajistila dostatečnou férovost, reliabilitu a byla by dostatečně obsahově validní. Zároveň muselo ohodnocení vyhovět poměrně náročným požadavkům univerzity a fakulty na strukturu a podobu PZ a být dostatečně efektivní i časově úsporné.

Při hodnocení bakalářské práce a textových výstupů uchazečů vůbec lze využít řadu různých analytických postupů s různou mírou subjektivity, a to od objektivního kódování jednoznačných kritérií přes počítačově asistované hodnocení (zpracování přirozeného jazyka, NLP, a machine learning) až po subjektivnější hodnocení komplexnějších aspektů textu (Breland et al., 1999). Obecně platí, že při ověřování kvality hodnocení není vhodné spoléhat se jen na celkovou shodu posuzovatelů, ale je nezbytné zaměřit se na dílčí kritéria a stanovit i další indikátory kvality (Wind & Peterson, 2018).

Reliabilita při hodnocení textů (zejména esejí) bývá velmi různorodá. Vnitřní konzistence hodnoticích kritérií se zpravidla pohybuje kolem 0,7, ale může klesnout až k 0,5 i v případě tak oblíbených testů, jako je např. Graduate Management Admissions Test (GMAT). Reliabilita ve smyslu shody posuzovatelů nicméně bývá ještě nižší a rovněž různorodá; zpravidla se pohybuje okolo 0,5, může ale klesnout až k 0,2 (Breland et al., 1999; Salvatori, 2001). Obecně je pak reliabilita ovlivněna mírou zaškolení hodnotitelů a standardizací hodnoticího procesu. Důležitým moderátorem je též počet hodnocených oblastí či kritérií, počet hodnocených prací a počet hodnotitelů. Zvyšování počtu prací má smysl zejména tehdy, pokud tím vzroste i počet hodnocených oblastí (Breland et al., 1999). Zároveň nelze říci, že navyšování počtu hodnotitelů / témat / hodnocených prací automaticky posiluje reliabilitu, záleží vždy na konkrétní situaci – z těchto důvodů bývá oblíbeným designem hodnocení dvou prací dvěma hodnotiteli s reliabilitou zpravidla v rozmezí 0,7–0,8 (což je využíváno i v high-stakes testech typu GMAT či GRE). Reliabilita při hodnocení jedné práce dvěma hodnotiteli se pak většinou pohybuje v rozmezí 0,5–0,6 (Breland et al., 1999).

Velká variabilita panuje i na poli statistického zpracování výsledků, přičemž se často liší vlastní skórovací systém a způsob ověření reliability. Systematická review Winda a Petersona (2018), byť v oblasti ověřování jazykových kompetencí, jmenuje mezi nejčastěji používanými postupy odhad reliability jako shody posuzovatelů tradičními postupy (Cohenovo kappa, vnitrotřídní korelace apod.; 31 % studií), raschovský model (19 %) či teorii zobecnitelnosti (9 %). Dále výzkumníci občas využívají jiné IRT přístupy, faktorovou analýzu a strukturní modelování či hierarchické (smíšené) lineární modely. (Wind & Peterson, 2018)

V našem případě se nabízí několik možných přístupů. Na součtu či průměru hodnocených kritérií je postavena teorie zobecnitelnosti (Generalizability Theory, GT) (Brennan, 2001; Cronbach et al., 1963). Ta předpokládá,

že hodnoticí škála je intervalová a že každé kritérium má shodně silný (tau-ekvivalentní), lineární vztah s hodnoceným atributem. To však může být problém zejména u různorodých kritérií, hodnocených na krátkých, ordinálních odpověďových škálách. Nelineární vztah položky a latentního rysu předpokládá teorie odpovědi na položku (IRT), která poskytuje velké množství vhodných modelů. Ty se liší v předpokladech, množství parametrů, náročnosti jejich odhadu a s tím souvisejících požadavcích na velikost vzorku. Shodnou diskriminační účinnost všech položek/hodnotitelů předpokládá například multifasetový Raschův model (Bond & Fox, 2009), lineární logistický testový model (LLTM) (Fischer, 1973) či s ním spřízněné explanační IRT modely (De Boeck et al., 2012) – ty všechny jsou vlastně logistickou variantou teorie zobecnitelnosti. Na předpoklad tau-ekvivalence naopak rezignuje řada dalších modelů, mezi nimi i Hierarchical Rater Model (HRM) (Casabianca et al., 2016; Patz et al., 2002); ty však obecně vyžadují větší množství respondentů k dostatečné identifikaci parametrů.

1.3 Požadavky na přijímací zkoušku do NMGR psychologie na FSS

Požadavky na PZ, které jsme brali v úvahu při tvorbě hodnoticích kritérií bakalářské práce, vycházely ze tří oblastí. První a nejdůležitější z nich byly požadované kompetence uchazečů ověřované právě prostřednictvím PZ, a tedy obsahová validita zkoušky.

Druhou oblastí byly formální požadavky univerzity a fakulty v kombinaci s kapacitními možnosti oboru, stejně jako časové možnosti a počet potenciálních hodnotitelů z řad pracovníků katedry. Posledním, třetím požadavkem bylo zajištění dostatečné reliability ve smyslu vnitřní konzistence i shody posuzovatelů (férovosti hodnocení).

2. Obsahová validita

2.1 Požadované kompetence uchazečů

Oblasti, které jsou z hlediska úspěchu v magisterském studiu zásadní a pro něž lze najít indikátory v bakalářské práci, jsou následující:

Psychologické znalosti získané předchozím studiem psychologie

Studující si volí své téma, a proto lze bakalářskou práci považovat za vzor studentova *teoretického maxima* – tedy toho, jak dobře dokázal studující zvládnout teorii v oblasti, která jej zajímá, mohl se na ni zaměřit a měl na to dostatek času. Projev znalostí zde spatřujeme ve schopnosti jasně deklarovat psychologické konstrukty a teorie popisující jev, kterým se studující v práci zabývá, a představit je ve vzájemných vztazích v jasně strukturovaném teoretickém rámci. Projevem schopnosti je také zkoumatelná výzkumná otázka, popř.

i ověřitelná hypotéza, která z tohoto rámce jasně vyplývá a jejíž zodpovězení teorii potenciálně obohatí. (Z těchto požadavků vychází později kritérium teoretický rámec, viz tabulku 1.) Za poslední indikátor znalostí a schopnosti s nimi pracovat považujeme formulaci zjištění na základě výsledků analýz tak, aby byla v souladu s navrženým teoretickým rámcem a vhodně do něj přispívala (kritérium *diskuze A*, viz tabulku 1). Spíše než šíře znalostí nás tak zajímá, jak dobře dokáže uchazeč porozumět vybrané oblasti teorie a jak dokáže s poznatky pracovat. To je z hlediska úspěchu ve studiu klíčové.

Funkční schopnost pracovat se současnou odbornou literaturou

Psychologie se rychle rozvíjí, a tak je nutné umět kombinovat poskytnuté výukové materiály s aktuálními poznatky ve studované oblasti. Protože je přirozené, že výzkum vychází se současných poznatků, je seznam v práci použitých zdrojů bezprostředním validním indikátorem schopnosti práce s odbornou literaturou. (*Literatura*.)

Metodologické a analytické znalosti a dovednosti

I s přihlédnutím k faktu, že design, metody i analýza dat bývají v pracích hojně konzultovány, a odrážejí tak i schopnosti a znalosti vedoucích a konzultantů, stále je lze využít pro hodnocení dovedností autora. Pro adekvátní realizaci a popis navržených postupů je totiž nezbytné porozumění všem analýzám a metodologickým krokům. Sekce metoda, výsledky a diskuze tak obsahují řadu indikátorů schopností a dovedností v této oblasti. (Kritéria *design A*, *design B1–B5*.)

Schopnost kriticky hodnotit poznatky

Tato schopnost se projevuje především v diskuzi, v níž má docházet ke kritickému zhodnocení vlastních zjištění z hlediska všech potenciálně limitujících faktorů. Je ale také indikována alespoň občasným kritickým přístupem ke zjištěním, která jsou využívána jako součást teoretického rámce i samotnou potřebou doprovázet klíčová tvrzení nějakými empirickými doklady. (*diskuze B*)

Formální náležitosti

Vědecká metoda má své náležitosti, standardy, formát předávání informací. Kvalitní práce by tyto náležitosti měla splňovat; jednoznačně nekvalitní grafická úprava, nevhodný jazyk či pravopisné chyby snižují důvěryhodnost prezentovaných myšlenek a v odborném textu nemají své místo. Kvalitní bakalářská práce by měla splňovat v současnosti obecně akceptované obsahové požadavky výzkumných zpráv v psychologii. (*formalni*)

Z uvedených kritérií je patrné, že se všechna zaměřují výhradně na schopnost studovat a na predikci úspěchu v NMgr. studiu, nikoliv na schopnost vykonávat psychologické řemeslo. Otázka, jaké schopnosti či dovednosti jsou

zásadní pro úspěch v psychologické profesi, je samozřejmě vysoce relevantní. Nepovažujeme ji však za vhodnou, nebo snad dokonce nezbytnou součást přijímací zkoušky. Absolventi NMgr. psychologie mohou působit v mnoha rozdílných profesích: od klinické psychologie či zdravotnictví přes psychoterapii a poradenství až po výzkum. Řada z nich se navíc uplatní zcela mimo obor, kde mohou své studijní poznatky dobře zužitkovat. Domníváme se, že naším úkolem není selektovat vybraný „typ" uchazečů, ale nabízet kvalitní a všeobecné psychologické vzdělání. Způsob jeho využití je pak na každém uchazeči zvlášť.

Žádné z kritérií se rovněž nezabývá motivací ke studiu, předchozí praxí, zkušenostmi a podobně, přestože součástí požadovaných dokumentů byl vedle bakalářské práce na pokyn fakulty i motivační dopis a doklady o praxi či dalším vzdělání. S ohledem na jejich vysoké obsahové i formální odlišnosti však nebylo možné vyvinout standardizovaný způsob jeho hodnocení, který by některé uchazeče nezvýhodňoval a jiné naopak nepoškozoval a který by netrpěl subjektivním posouzením hodnotitelů. Rozhodli jsme se proto motivační dopisy použít pouze jako „vstupní filtr"; do dalšího hodnocení by proto nepostoupili a nula bodů by obdrželi ti uchazeči, jejichž motivační dopis by svědčil o celkové neznalosti obsahu psychologické praxe a předmětu psychologie jako vědy. U žádného z uchazečů jsme nicméně k tomuto kroku nepřistoupili.

2.2 Formální a ekonomické požadavky na PZ

Formální požadavky fakulty byly poměrně svazující. PZ se musela konat ve dvou termínech, ke kterým uchazeči mohli dodat své bakalářské práce, přitom však muselo být rozhodnutí o přijetí či nepřijetí uchazečům sděleno v co nejkratším čase. To vedlo k paradoxní situaci, kdy jsme o přijetí uchazečů z prvního termínu museli rozhodnout dříve, než bylo známo, kolik uchazečů předloží svoji práci k hodnocení ve druhém termínu. Kapacitní limit je navíc sdílen mezi PZ do jarního a podzimního semestru, což veškeré plánování dále komplikuje (část kapacitního limitu musela být ponechána pro lednový termín zkoušky, o jejíž podobě rovněž panovaly pochybnosti).

Dále bylo nutné udělit body všem uchazečům, kteří splnili formální podmínky, tj. získali bakalářský diplom v oboru psychologie a předložili veškeré požadované dokumenty. Udělené body musely být celočíselné v rozpětí 0–60, bodová hranice nutná pro přijetí je arbitrárně stanovena na 60 %, tedy 36 bodů. Studenti s bodovým ziskem 36 a více bodů tak měli být navrženi k přijetí, studenti s menším ziskem bodů nepřijati.

Kombinace výše uvedených požadavků vedla k paradoxu, kdy vyhodnocení přijímacího řízení bylo normativní (k přijetí měl být navržen určitý počet uchazečů) i kriteriální zároveň (tito uchazeči museli získat určitý minimální počet bodů v PZ). Do podoby PZ navíc vnášela značnou nejistotu

existence dvou termínů – stanovit bodovou hranici a náročnost PZ bylo nutné ihned po prvním termínu tak, aby po druhém termínu nebylo ke studiu příliš mnoho či naopak příliš málo studentů. Z důvodu zajištění férovosti však musely být podmínky pro přijetí totožné napříč oběma termíny a přísnost hodnocení se nesměla lišit.

2.3 Reliabilita

Poslední třetí okruh požadavků se zaměřoval na „technickou" podobu hodnocení. Katedra předem počítala se zhruba 15 hodnotiteli (nakonec jich měla k dispozici 18), a bylo tedy nutné zajistit jejich obdobnou přísnost; hodnoticí kritéria musela být jednoznačná, aby minimalizovala potenciální neshody a subjektivitu. Kritérií musel být dostatečný počet pro zajištění přiměřené vnitřní konzistence hodnocení, naopak jejich příliš velké množství by zvyšovalo časovou náročnost hodnocení a mohlo by paradoxně vést k tomu, že by hodnotitelé pracovali s nižší úrovní pozornosti, a reliabilita by tak klesala. Dále bylo zřejmé, že hodnotitelé budou muset být zaškoleni a že navržený design hodnocení musí brát v potaz jejich časové možnosti a jiné závazky týkající se výuky a výzkumu.

Všechny výše uvedené požadavky definovaly detaily našeho výzkumného záměru, tedy vývoj dostatečně kvalitního hodnoticího schématu bakalářských prací. Ten společně s výsledky popíšeme na následujících stranách.

3. Metoda

3.1 Vzorek

Vzorek tvoří uchazeči o studium, kteří ve stanoveném termínu zaslali veškeré požadované dokumenty. Celkem jde o $N_1 = 82$ studentů a studentek v prvním termínu a $N_2 = 48$ v termínu druhém. Věk, pohlaví a jiné sociodemografické údaje nebyly záměrně sledovány.

V obou termínech hodnotilo bakalářské práce stejných 18 hodnotitelů z řad akademických a výzkumných pracovníků. Vybráni byli tak, aby reprezentovali všechna výzkumná a pedagogická zaměření (aby tedy nešlo např. jen o vyučující statistických a metodologických kurzů).

3.2 Kritéria hodnocení bakalářských prací

Kritéria jsou dílem čtyřčlenné přijímací komise. Dva ze členů nezávisle na sobě navrhli dvě různé podoby hodnocení, třetí člen komise pak oba návrhy agregoval do výsledné podoby. Ta navíc prošla připomínkovým řízením a po vzájemných konzultacích ještě doznala mírných úprav. Celé posuzování bakalářských prací tak probíhalo zcela nezávisle na hodnocení u jejich obhajob. Při tvorbě kritérií byla dodržována následující pravidla:

1. Kritéria byla volena tak, aby skýtala minimální prostor pro subjektivní interpretaci.
2. Kritéria byla hodnocena na krátké škále 0–1, nebo 0–2 body, aby byl rovněž snížen prostor pro subjektivní posouzení – aby každý bodový stupeň mohl být přesně operacionalizován.
3. V návaznosti na předchozí deziderata se kritéria zaměřovala především na přítomnost/nepřítomnost obecně žádoucích prvků prací, nikoli na míru jejich přítomnosti či naplnění. Explicitně jsme se vyhýbali hodnocení toho, zda byla výzkumná očekávání daty podpořena, či ne, jestli jsou výsledky v souladu se současnou teorií, či ne apod.
4. Kritéria byla volena tak, aby byla nezávislá na zvoleném designu a epistemologickém přesvědčení autora. Kritéria musela být vhodná pro kvantitativní i kvalitativní práce, či dokonce teoretickou stať.
5. Kritéria musela být nezávislá na formálních náležitostech pracoviště, kde práce vznikla (např. délka či požadovaná šablona práce).
6. Některá kritéria byla použita jako tzv. screeningová kritéria; při jejich nenaplnění práce nebyla podle dalších kritérií dále hodnocena.
7. Všechna kritéria navíc vycházela z požadavků na obsahovou validitu, které jsme popsali v úvodu.

Výsledkem je sada 11 bodovaných kritérií doplněná o 4 screeningová binární kritéria. Ta sloužila pro identifikaci jak prací, které by nemohly získat mnoho bodů (např. neúplnost), což znamená úsporu času hodnotitelů, tak prací, které jsou ve zjevném rozporu s požadavky na akademickou integritu. Při nesplnění všech čtyř kritérií už práce nebyla dále hodnocena. Jednotlivá kritéria byla hodnocena na škále 0–1, případně 0–2 body; celkem tak bylo možné získat až 16 bodů hrubého skóru. Podrobný popis kritérií včetně jejich bodové hodnoty je součástí tabulky 2. Tyto body nicméně nebyly konečným hodnocením, výsledné pořadí bylo založeno na hodnoceních korigovaných prostřednictvím statistického modelu (viz níže). Hodnoticí kritéria byla předem ve zjednodušené a hrubé podobě oznámena uchazečům e-mailem[1]. Navržená kritéria se přitom liší od běžných kritérií používaných během obhajob a státních zkoušek. Na našem pracovišti jsou kritéria spíše implicitní, posudky nemívají pevnou strukturu. Mnohem větší roli má hodnocení kreativity a invence autora či další komplexnější aspekty, které naopak nebyly záměrně hodnoceny v rámci přijímací zkoušky. Kritéria hodnocení na jiných pracovištích pak mohou být ještě výrazně odlišnější.

[1] https://psych.fss.muni.cz/cosedeje/aktuality/informace-pro-uchazece-prijimaci-zkouska-do-podzimniho-semestru-2020

Tabulka 1
Hodnotící kritéria

ID položky	proměnná (počet bodů)	podrobné instrukce
Screeningová kritéria Pokud v této sekci nejsou splněny první tři podmínky, nebodovat dál. Pokud jsou splněny, zkontrolovat literaturu („literatura"). Pokud není ve screeningu uděleno čtyřikrát ANO, pak práce není dále bodována.		
X_vejce	VejceVejci	Kontrola plagiátů prostřednictvím nástroje Masarykovy univerzity VejceVejci. ANO = v pořádku. NE = netriviální shoda (> 5% shoda s jinou prací)[2].
X_tema	Psychologické téma	NE volíme jen tehdy, když jde o zjevně a nepochybně nepsychologické téma.
X_cele	Celá studie	NE dáváme, když jde jen o projekt či nedokončené torzo. Pokud je na vině koronavirus a autor/ka omezení diskutuje, lze i přes to udělit ANO.
X_ literatura	Literatura	ANO = Alespoň jeden bod v kritériu „literatura". NE = nula bodů v kritériu „literatura".
Hodnotící kritéria Nehodnotit, pokud ve screeningu nebylo zvoleno čtyřikrát „ano".		
literatura	Literatura (0–2)	**Hodnocení 2 body:** Referencí je dostatečný počet [> 30] V referencích převažují časopisecké empirické studie [> 1/3] V referencích převažují zahraniční zdroje [> 1/2] Reference mají jednotný formát [APA či jiný] Namátková kontrola 3 odkazů v textu datovaných po r. 2000 nezjistila nesoulad mezi odkazovaným tvrzením a charakterem citovaného zdroje. Minimum sekundárních citací [< 10] **Hodnocení 1 bodem:** Jako 2 body, avšak: V seznamu jsou ovšem navíc učebnice, slovníky, encyklopedie [> 4], nebo… … maximálně jeden z parametrů uvedených výše není splněn. **Hodnocení 0 body:** Nesplňuje podmínky pro udělení 1 nebo 2 bodů.

[2] Veškerá podezření na plagiát byla vždy přezkoumána přijímací komisí, protože procentuální shoda s jinými texty nemusí být validním ukazatelem.

ID položky	proměnná (počet bodů)	podrobné instrukce
teorie	Teoretický rámec (0–2)	**Hodnocení 2 body:** Teoretický rámec má vzhledem k cílům práce jasnou, pochopitelnou strukturu. Je zřejmé, které pojmy jsou klíčové, a tyto jsou důsledně definovány. Výzkumná otázka je jasně stanovena v návaznosti na teoretický rámec. Autorské prvky – citované teze jsou prezentovány kriticky, hodnoceny, uváděny do souvislostí **Hodnocení 1 bodem:** Jako 2, ale bez autorských prvků; nebo… … kvalita některých požadavků na 2 je snížená, avšak akceptovatelná. **Hodnocení 0 body:** Alespoň jeden z prvních 3 požadavků na 2bodové hodnocení je zcela nenaplněný.
A	Design A (0/1)	Zvolený design v principu umožňuje získat odpověď na výzkumnou otázku. Pohybujeme se na úrovni observační, korelační, experimentální studie, IPA, GT analýza dat atp.
B1	Design B1 (0/1)	**Kvalitativní práce** Je formulována smysluplná výzkumná otázka a výsledky analýzy (témata, kategorie, teorie, ...) na ni skutečně odpovídají. **Kvantitativní práce** Jsou jasně stanoveny hypotézy, popř. jasně deklarovaný explorační charakter práce.
B2	Design B2 (0/1)	**Kvalitativní práce** Metoda je podrobně a konkrétně popsána, včetně odkazů na primární metodologickou literaturu. **Kvantitativní práce** Je zřejmé, z jaké populace a jakým postupem byl získán vzorek. Velikost vzorku je založena na úvaze o síle testu.
B3	Design B3 (0/1)	**Kvalitativní práce** Ve výsledcích je zřetelný prvek vlastní analýzy/syntézy (tj. není to jen pouhá deskripce dat bez autorova analytického přínosu). **Kvantitativní práce** Metody měření jsou představeny s věcně adekvátní zmínkou o validitě a reliabilitě.
B4	Design B4 (0/1)	**Kvalitativní práce** Zakotvenost výsledků v datech je doložena vhodně zvolenými citacemi/úryvky. **Kvantitativní práce** Design je jasně popsán.

ID položky	proměnná (počet bodů)	podrobné instrukce
B5	Design B5 (0/1)	**Kvalitativní práce** Vzorek je tvořen na základě věcně zdůvodněných kritérií (ne autoritou). **Kvantitativní práce** Analytické modely jsou vhodně zvolené, bez vyložených přešlapů v oblasti (ne)testování hypotéz.[3] Jsou přítomny deskriptivy a velikosti účinku.
D1	Diskuse A (0–2)	**Hodnocení 2 body** Zjištění bylo sděleno netriviálním způsobem s explicitní reflexí přínosu (nevadí, když se nepovede nic přinést). Zjištění byla začleněna do teoretického rámce (s citacemi) s autorským pohledem. **Hodnocení 1 bodem** Jedno chybí, nebo nemá úroveň. **Hodnocení 0 body** Ani jedno nemá úroveň (formalismus, alibismus...)
D2	Diskuse B (0–2)	**Hodnocení 2 body** Přítomnost reflexe interní validity (kredibility). Přítomnost reflexe externí validity (zobecnitelnosti). **Hodnocení 1 bodem** Jedno chybí, nebo nemá úroveň (např. „nelze zobecňovat“). **Hodnocení 0 body** Oboje chybí nebo nemá úroveň.
formalni	Formality (0–2)	Pravopisná, typografická a jazyková úprava. **Hodnocení 2 body** Pravopisná i grafická úprava je perfektní, jen s minimem nedostatků. **Hodnocení 1 bodem** Práce obsahuje jen málo pravopisných chyb a občasné typografické nedostatky, které nekomplikují čtení. **Hodnocení 0 body** Práce obsahuje značné množství pravopisných nebo typografických nedostatků, které komplikují čtení.

3.3 Procedura

Pro zajištění reliability – shody posuzovatelů – byla každá práce hodnocena dvěma hodnotiteli. V prvním termínu každý z nich posuzoval 9 nebo 10 prací, ve druhém termínu 5 nebo 6. Práce byly hodnotitelům přiřazovány náhodně pomocí algoritmu naprogramovaného v prostředí R (viz Cígler et al., 2020). Ten kromě náhodného přiřazení prací hodnotitelům zajišťoval i vyvážený počet prací na jednoho hodnotitele. Následně přijímací komise ověřila,

[3] Jinými slovy to znamená, že použité statistické postupy odpovídají hypotézám a umožňují jejich ověření. Zároveň jsou statistické informace doplněny o vhodný slovní komentář, ze kterého není patrné, že by autor práce jím použitým statistickým analýzám nerozuměl.

zda není vylosovaným hodnotitelem práce její vedoucí ani oponent (v tomto případě jej změnila). Domnívali jsme se totiž, že předchozí znalost práce, zkušenost z průběhu obhajoby a osobní vazba na uchazeče by mohly hodnocení potenciálně zkreslovat. Dále pak komise zkontrolovala jazyk práce. Několik prací bylo totiž předloženo v jiném než českém, slovenském či anglickém jazyce: jmenovitě němčině a ruštině. I u těchto prací byl vybrán hodnotitel tak, aby na dostatečné úrovni ovládal příslušný jazyk.

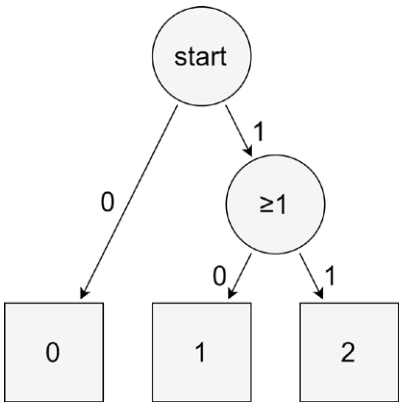
Hodnotitelé absolvovali krátké zaškolení prostřednictvím videokonference před prvním kolem hodnocení; videozáznam (54 minut) z tohoto školení měli nadále k dispozici. Kromě toho dostali hodnoticí tabulku v programu MS Excel s rozepsanými kritérii, do které zapisovali svá hodnocení.

Hodnocení prací hodnotiteli následně sloužilo jako podklad pro čtyřčlennou přijímací komisi, která ověřovala konzistenci hodnocení a řešila sporné případy. Kromě předsedy komise byli všichni její členové zároveň hodnotiteli. V případě, že se hodnotitelé neshodli na splnění či nesplnění screeningových kritérií, přiklonila se komise k jednomu či druhému stanovisku podle svého hodnocení příslušných kritérií. Pokud komise rozhodla ve prospěch uchazeče, zpravidla chybělo hodnocení bodovaných kritérií jedním z hodnotitelů. V takovém případě práci ohodnotil ten z členů přijímací komise, který nebyl vzhledem k práci v pozici střetu zájmů. Na podkladě takto korigovaných hodnocení pak komise rozhodla o přijetí či nepřijetí uchazečů.

3.4 Volba psychometrického modelu a způsob hodnocení

Pro validní posouzení prací byla klíčová volba psychometrického modelu, který by umožnil vyvážení „přísnosti" jednotlivých hodnotitelů a pracoval by s binárními, resp. ordinálními položkami. Zároveň nesměl být omezen chybějícími daty, tedy faktem, že každého uchazeče hodnotili pouze dva hodnotitelé, a malým počtem uchazečů. Mezi různými alternativními postupy byl zvolen tzv. multifasetový Raschův model (Bond & Fox, 2009) v rámci teorie odpovědi na položku (Item Response Theory, IRT). Na rozdíl od víceparametrových modelů neumožňuje rozdílnou diskriminační účinnost různých položek, má tak menší počet parametrů a snižuje náročnost na velikost vzorku. Model byl nicméně odhadován jako LLTM (lineární logistický testový model) v R balíčku lme4 (Bates et al., 2015; De Boeck et al., 2012; Fischer, 1973). Výhodou této volby byla maximální analytická flexibilita díky použití běžného balíčku pro generalizovaný lineární smíšený model (GLMM) bez nutnosti využít specializovaného softwaru. Ordinální položky na škále 0–2 jsme pro něj parametrizovali podle Tutzova sekvenčního Raschova modelu, SRM (Tutz, 1990), protože běžnější ordinální IRT modely (např. RSM, GPCM či GRM) není možné odhadnout jako LLTM model (viz De Boeck et al., 2012, pro podrobější diskuzi); výsledky by však měly být v praxi podobné.

To znamená, že každá ordinální položka byla rozdělena na dvě dílčí binární podpoložky. První položka nesla informaci, zda hodnotitel udělil hodnocení 0 nebo 1 a více bodů. Pokud udělil 0 bodů, druhá podpoložka byla chybějící. V opačném případě obsahovala informaci, zda uchazeč získal 1 nebo 2 body. V grafické podobě námi zvolený SRM prezentuje obrázek 1.



Obrázek 1
Poznámka: Ovály/uzly označují rekódované položky, čtverce pak celkový počet bodů. Popisky hran grafu označují počet bodů získaných v dané rekódované položce. V případě pozorované odpovědi 0 na první uzel (start) je odpověď ve druhém uzlu (≥ 1) *datová hodnota „chybějící"* (NA).

Výsledný dataset tak obsahoval 15 binárních položek skórovaných 0;1. Tato data byla převedena do dlouhého formátu, kde na každém řádku byla informace, zda respondent od hodnotitele získal 0 nebo 1 bod a ID hodnotitele, uchazeče a položky. Raschův IRT model nad těmito daty byl odhadnut jako GLMM model s binomickou logit-link funkcí.
Model byl charakterizován odpověďovou funkcí s formální definicí

$$\ln \frac{P_{ipr}}{1-P_{ipr}} = \theta_p + \beta_i + \rho_r \qquad (1)$$

kde levá strana představuje modelem predikovaný logaritmus šance správné odpovědi; P_{ipr} je tedy pravděpodobnost, že uchazeč *p* získá v kritériu *i* od hodnotitele *r* 1 bod. Klíčový parametr θ_p reprezentuje kvalitu práce jako úroveň latentního rysu uchazeče *p*. Snadnost položky *i* je označena β_i (opačná hodnota tradičního parametru obtížnosti v IRT) a konečně ϱ_r je mírnost či shovívavost hodnotitele *r* (opak náročnosti či přísnosti; tato slova používáme jako synonyma). Parametry uchazečů θ_p a hodnotitelů ϱ_r byly modelovány jako náhodné (random effects), zatímco snadnost položek ϱ_r byla pevným parametrem (fixed effect). Pro postup posouzení shody modelu s daty prostřednictvím ukazatelů infit a outfit viz online suplement, přílohu 3.

Alternativním modelem, který jsme zvažovali, byl prostý průměr bodů za oba hodnotitele. Tento model však rovněž není zcela triviální a nese s sebou určité předpoklady; minimálně předpoklad intervalových položek, stejné přísnosti a váhy (rozlišovací účinnosti) všech hodnotitelů a shodné rozlišovací účinnosti položek. Pro odhad reliability takovéhoto modelu by navíc bylo stejně nezbytné využít pokročilejších analýz v rámci teorie zobecnitelnosti. Běžné odhady v rámci klasické testové teorie (např. korelace obou hodnocení napříč hodnotiteli a pracemi) by byly totiž zkreslené nenaplněním řady předpokladů, zejména nezávislosti jednotlivých pozorování.

3.5 Statistická analýza

Veškeré analýzy byly provedeny v prostředí R (R Core Team, 2020). Odhad LLTM modelu byl realizován prostřednictvím knihovny lme4 (Bates et al., 2015), pro dílčí analýzy byla použita řada dalších knihoven v čele s balíčkem psych a DescTools (Gagolewski, 2020; Gu et al., 2014; Phillips, 2017; Revelle, 2020; Sarkar, 2008; Signorell et al., 2020; Wickham, 2007). Veškeré analytické skripty, rozšířené výsledky a doplňující materiály jsou dostupné na Open Science Framework (Cígler et al., 2020).

V následujícím textu jsou výsledky z důvodu přehlednosti prezentovány pro oba termíny současně – vyjma pasáží, kde je to explicitně zmíněno. Při čtení je proto nutné mít na paměti, že vyhodnocení fungování PZ proběhlo hned po prvním termínu a po druhém termínu byl celý postup replikovaný. Celý analytický postup je po technické stránce popsán právě tak, jak probíhalo skutečné vyhodnocení PZ[4].

4. Výsledky

4.1 Screeningová kritéria

V prvním termínu úvodní kritéria jednoznačně splnilo 60 uchazečů (73 %), 13 nikoli (16 %) a u zbylých 9 uchazečů (11 %) se hodnotitelé neshodli. To značí relativně nízkou míru shody posuzovatelů, Cohenova kappa $\varkappa = 0{,}673$

[4] Jedinou odchylkou je fakt, že z důvodu analytické chyby na prvním termínu byl model parametrizován s nenulovým průsečíkem; obtížnosti položek proto byly vyjádřeny rozdílem obtížnosti oproti první položce s odhadem fixovaným na nulu. Z důvodu replikovatelnosti celé procedury byla stejná parametrizace využita i na druhém termínu; hodnota průsečíku byla vzata v úvahu při transformaci na bodové skóre, tvorbě grafů a dalších postupech. V následujícím textu však pro snadnější interpretaci a přehlednost využíváme mírně odlišnou parametrizaci s nulovým průsečíkem a odhadovanou obtížností všech položek. Tyto dvě parametrizace se nicméně neliší v odhadu úrovně latentního rysu (náhodné faktory mají vždy nulový průměr), pouze se posunuly odhady parametrů položek o již zmíněnou konstantu. Rozdíl tak nemá žádný praktický význam na výsledky hodnocení.

s 95% *CI* = [0,476; 0,871]. Sporné případy rozhodla přijímací komise třikrát ve prospěch uchazeče (jednou plagiát, jednou literatura a jednou nedokončenost), ve zbylých šesti případech v jeho neprospěch (čtyřikrát nedostatečná literatura, dvakrát nedokončená práce). Do druhé fáze hodnocení tak postoupilo $n_1 = 63$ uchazečů, jejichž hodnocení bylo použito pro odhad IRT modelu.

Ve druhém termínu kritéria splnilo 32 uchazečů (67 %), 7 nikoli (15 %) a 9 (19 %) případů bylo sporných. Shoda posuzovatelů byla ještě nižší než v prvním termínu, $\varkappa = 0{,}485$ s 95% *CI* = [0,196; 0,775], přičemž přijímací komise rozhodla sporné případy dvakrát ve prospěch uchazeče (psychologické téma) a sedmkrát v jeho neprospěch (jedna práce nebyla dokončená, u zbylých nevyhovoval seznam literatury). Do druhé fáze hodnocení tak postoupilo $n_2 = 34$ uchazečů. Příčiny nízké shody posuzovatelů u screeningových kritérií jsou podrobně popsané v diskuzi.

V obou termínech dohromady 54krát udělil hodnotitel pouze tři body ve screenigových kritériích, devětkrát dva body, dvakrát jeden bod a v jediném případě žádný bod. Při celkem 194 hodnoceních byly uděleny všechny body ve screeningových kritériích, což odpovídá $n = 97$ uchazečům, jejichž práce byly hodnoceny celé.

U 17 vyřazených prací jsme měli k dispozici body udělené alespoň jedním hodnotitelem, což nám umožnilo ověřit kriteriální validitu screeningových kritérií. Zprůměrovali jsme proto počet bodů hrubého skóru uděleného oběma hodnotiteli a porovnali vyřazené a nevyřazené uchazeče Mannovým-Whitneyho testem, $p(114) < 0{,}001$. Rozdíl průměrů byl velký, Cohenovo $d = 1{,}16$ s 95% *CI* = [0,61, 1,70]. Protože ty nejhorší práce pravděpodobně vyřadili (a tedy nebodovali) oba hodnotitelé a nemáme k nim data, skutečný rozdíl by měl být ještě vyšší. To značí vysokou prediktivní validitu screeningových kritérií.

4.2 Deskriptivní statistiky

Deskriptivní statistiky položek těch respondentů, kteří splnili screeningová kritéria, obsahuje tabulka 2; je nicméně nutné vzít v úvahu, že korigovaná korelace položky s celkovým skórem a Cronbachova alfa po vyřazení položky jsou stejně jako odhady reliability ve smyslu klasické testové teorie (CTT) silně zkresleny hierarchickou strukturou dat. Cronbachova alfa a McDonaldova omega hodnoticí škály jsou $\alpha = 0{,}834$ a $\omega = 0{,}859$ pro první termín, resp. $\alpha = 0{,}843$ a $\omega = 0{,}860$ pro termín druhý. Přesnější odhad reliability součtového skóru, který bere v potaz hierarchickou strukturu dat (a neshodu hodnotitelů), poskytuje teorie zobecnitelnosti: koeficient zobecnitelnosti byl $\Phi = 0{,}821$ pro první a $\Phi = 0{,}822$ pro druhý termín. Postup odhadu je uveden v online suplementu (příloze 7).

Průměr prostého součtu položek byl $M = 11{,}64$ ($SD = 3{,}60$) v prvním a $M = 11{,}50$ ($SD = 3{,}50$) ve druhém termínu; v obou případech bylo rozložení mírně negativně zešikmené.

Tabulka 2 obsahuje i shodu posuzovatelů v bodovaných kritériích, vyjádřenou v případě binárních kritérií pomocí Cohenova kappa, v případě ordinálních kritérií pomocí váženého kappa. Ve všech případech je shoda sice nízká, ale přiměřená. Je však zřejmé, že hodnocení práce dvěma hodnotiteli s možností zohlednění jejich přísnosti zde má potenciál reliabilitu hodnocení výrazně zlepšit. Z důvodu malého počtu uchazečů reportujeme výsledek pouze pro oba vzorky dohromady (i tak je patrný příliš široký interval spolehlivosti u většiny položek).

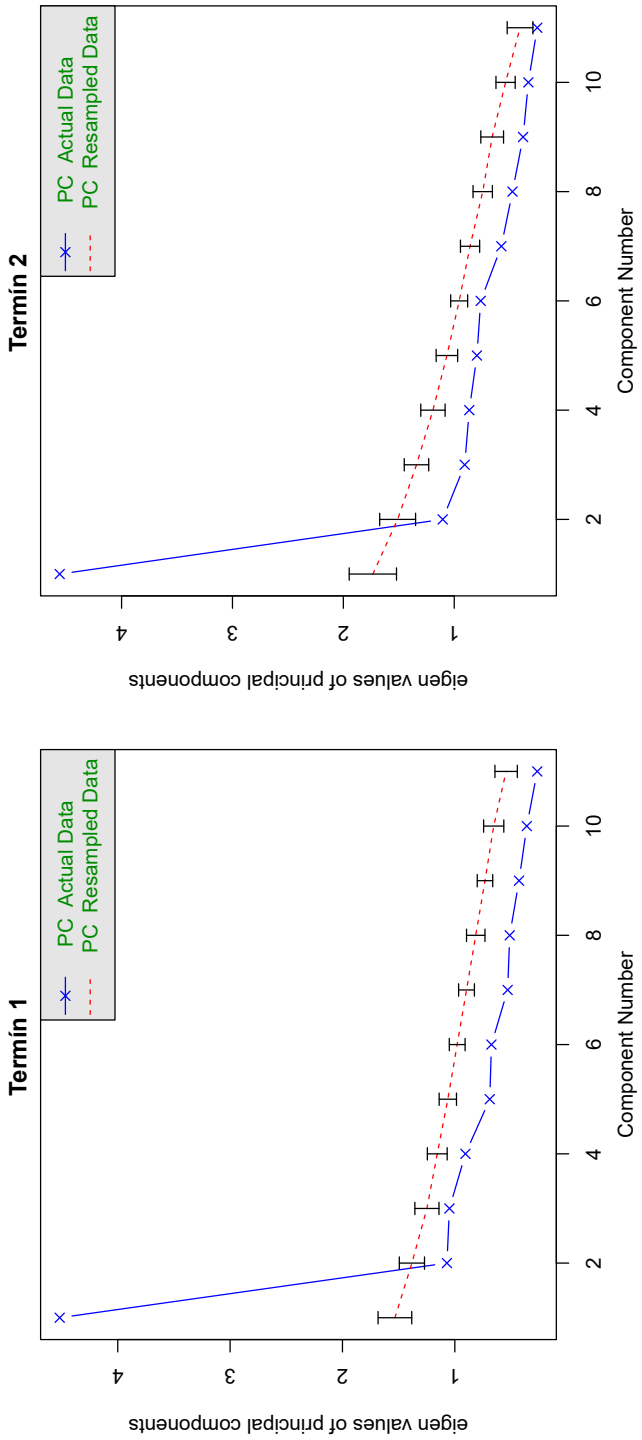
Tabulka 2
Deskriptivní statistiky položek

	termín 1 (n_1 = 63)					termín 2 (n_2 = 34)					shoda posuzovatelů (n = 97)	
Kritérium	M	SD	P	r_{drop}	α_{del}	M	SD	P	r_{drop}	α_{del}	$\varkappa$	[*95% CI*]
literatura	1,67	0,47	0,83	0,55	0,82	1,72	0,45	0,86	0,46	0,83	0,84	[0,75; 0,90]
teorie	1,26	0,72	0,63	0,63	0,81	1,46	0,58	0,73	0,70	0,81	0,39	[0,20; 0,58]
A	0,87	0,33	0,87	0,43	0,83	0,88	0,32	0,88	0,51	0,83	0,34	[0,07; 0,61]
B1	0,83	0,37	0,83	0,44	0,83	0,75	0,44	0,75	0,40	0,84	0,41	[0,19; 0,64]
B2	0,50	0,50	0,50	0,40	0,83	0,49	0,50	0,49	0,34	0,84	0,29	[0,10; 0,48]
B3	0,75	0,44	0,75	0,52	0,82	0,72	0,45	0,72	0,58	0,83	0,46	[0,27; 0,65]
B4	0,88	0,33	0,88	0,53	0,82	0,88	0,32	0,88	0,45	0,84	0,36	[0,09; 0,64]
B5	0,68	0,47	0,68	0,53	0,82	0,63	0,49	0,63	0,56	0,83	0,33	[0,13; 0,53]
D1	1,21	0,78	0,60	0,69	0,80	1,16	0,70	0,58	0,66	0,82	0,54	[0,39; 0,68]
D2	1,22	0,71	0,61	0,72	0,80	1,09	0,71	0,54	0,67	0,82	0,53	[0,36; 0,70]
formální	1,77	0,51	0,88	0,28	0,84	1,72	0,48	0,86	0,49	0,83	0,50	[0,16; 0,85]

Pozn.: M = průměr; SD = směrodatná odchylka; P = popularita (v případě ordinálních položek byl průměr vydělený dvěma); r_{drop} = korelace položky se součtem zbylých položek; α_{del} = Cronbachova alfa po odstranění položky; $\varkappa$ = Cohenovo kappa (v případě vícebodových položek vážená) pro oba termíny dohromady.

4.3 Faktorová struktura hodnocení

Základním předpokladem následujících analýz a zejména použitého LLTM modelu je alespoň přibližně jednodimenzionální struktura hodnocení, resp. lokální nezávislost jednotlivých kritérií. Faktorová struktura byla prozkoumána prostřednictvím explorační faktorové analýzy. Hornova paralelní analýza (Horn, 1965) nad maticí Pearsonových korelací indikovala v obou termínech jednoznačně jednodimenzionální strukturu, viz sutinové diagramy na obrázku 2 (pro způsob vypořádání se s hierarchickou strukturou dat viz online suplement, přílohu 4).



Obrázek 2
Sutinový graf s Hornovou paralelní analýzou pro oba termíny.

4.4 Parametry LLTM modelu a shoda s daty

Vzhledem k potřebě rozhodnout o přijetí a nepřijetí uchazečů byla procedura popsaná v následujícím textu realizována ihned po prvním termínu. Po druhém termínu byla následně replikována na datech uchazečů z druhého termínu a výsledky (parametry modelu apod.) byly porovnány. Následně pak byla procedura replikována na sloučených datech z obou termínů za účelem dosažení co největší stability odhadu parametrů. V následujícím textu nejprve popíšeme parametry celého modelu ($n = 97$). Následně představíme způsob stanovení kritického skóru a způsob převodu na bodovou škálu v prvním vzorku ($n_1 = 63$). Nakonec popíšeme srovnání obou vzorků ($n_1 = 63$, $n_2 = 34$). Toto pořadí neodpovídá časové posloupnosti realizovaných analýz; věříme však, že bude pro čtenáře výrazně srozumitelnější.

Parametry položek odhadnutého modelu nad všemi daty obsahuje tabulka 4; grafy charakteristických funkcí (resp. skórovací funkce) včetně rozložení odhadů schopností respondentů a shovívavosti hodnotitelů jsou na obrázku 3. Další vizualizace náhodných parametrů modelu je součástí online suplementu, viz přílohu 1.

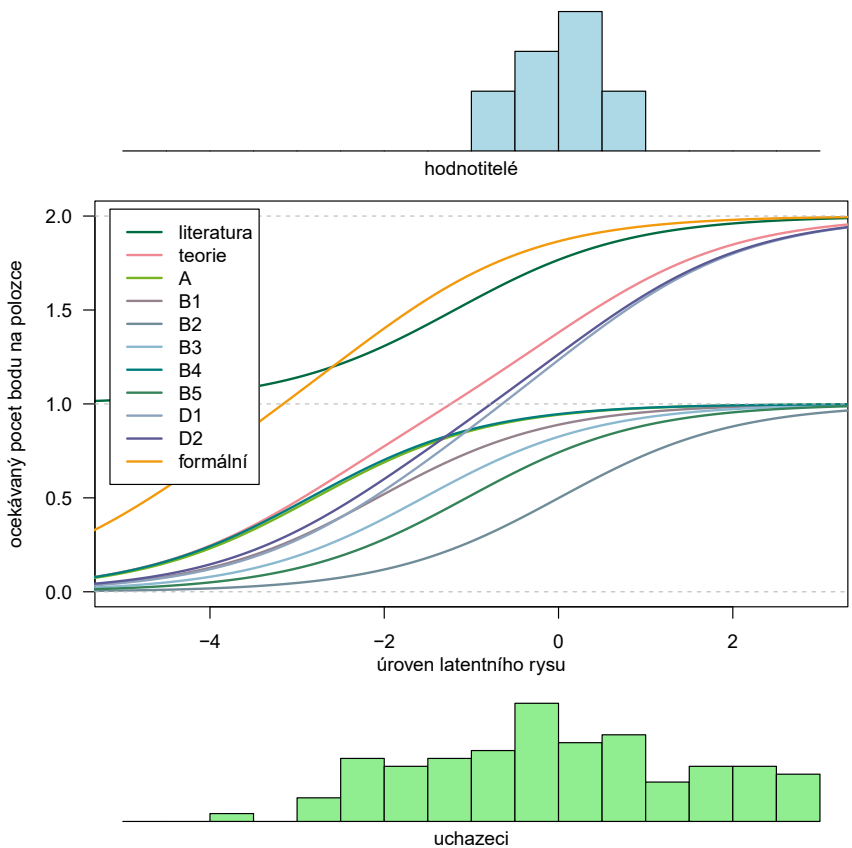
Tabulka 4
Odhad parametrů LLTM modelu

	snadnost			outfit[b]		infit[b]	
položka kategorie	est	SE	[95% CI]	χ^2/df	z	χ^2/df	z
literatura$_2$	1,19	0,29	[0,63; 1,76]	0,69	−3,35	0,90	−0,30
teorie				0,75	−3,74	0,91	−0,39
teorie$_1$	2,86	0,34	[2,2; 3,52]	0,87	−1,35	1,04	0,22
teorie$_2$	−0,16	0,29	[−0,72; 0,4]	0,64	−4,03	0,83	−0,59
A	2,80	0,33	[2,15; 3,45]	0,49	−6,23	0,91	−0,14
B1	2,08	0,31	[1,48; 2,68]	0,76	−2,59	0,98	0,04
B2	−0,01	0,28	[−0,55; 0,54]	1,29	2,63	1,20	0,83
B3	1,55	0,29	[0,98; 2,13]	0,66	−3,74	0,89	−0,31
B4	2,86	0,34	[2,2; 3,52]	0,37	−8,29	0,80	−0,50
B5	1,05	0,29	[0,49; 1,62]	0,73	−2,86	0,90	−0,29
D1				0,61	−6,27	0,83	−0,86
D1$_1$	1,99	0,30	[1,4; 2,59]	0,78	−2,26	0,98	0,04
D1$_2$	−0,40	0,29	[−0,97; 0,17]	0,44	−7,04	0,69	−1,19
D2				0,52	−8,28	1,03	0,19
D2$_1$	2,22	0,31	[1,61; 2,82]	0,45	−6,96	0,76	−0,74
D2$_2$	−0,79	0,29	[−1,35; −0,22]	0,58	−4,82	0,82	−0,61

	snadnost			outfit[b]		infit[b]	
položka kategorie	est	SE	[95% CI]	χ^2/df	z	χ^2/df	z
formální				1,45	5,53	1,03	0,19
formální$_1$	4,56	0,50	[3,59; 5,53]	1,56	4,76	0,90	0,05
formální$_2$	2,04	0,31	[1,44; 2,64]	1,34	3,04	1,06	0,29
náhodné efekty (SD)							
uchazeči[a]	1,73		[1,42; 2,08]				
hodnotitelé[a]	0,53		[0,25; 0,76]				

N = 97

Pozn.: est = odhad parametru; SE = standardní chyba odhadu; CI = interval spolehlivosti. Spodní index u některých položek označuje subpoložky vytvořené pro účely odhadu prostřednictvím Tutzova sekvenčního modelu. [a] Interval spolehlivosti byl odhadnut neparametrickým bootstrapem s 1000 permutovanými vzorky. [b] Pro postup výpočtu viz online suplement, přílohu 3.



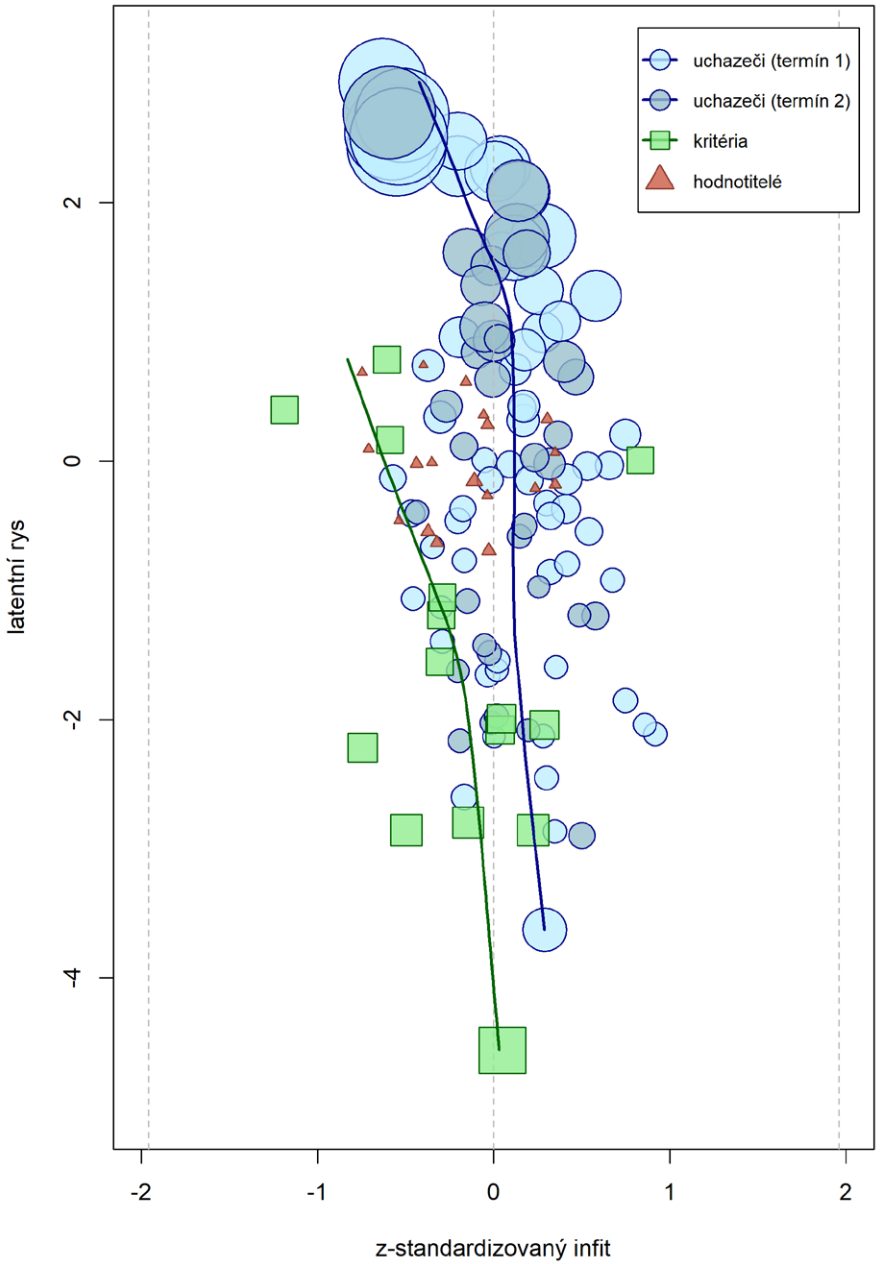
Obrázek 3

Charakteristické funkce položek, rozložení shovívavosti hodnotitelů a schopností uchazečů

Model popsal data přespříliš dobře, $\chi^2(2789) = 2311{,}9$; $p = 1{,}000$. Je nutné mít na paměti, že test maximální věrohodnosti využívající zdrojová data, tedy tzv. full-information, může shodu s daty výrazně nadhodnocovat (Cai & Hansen, 2013; Maydeu-Olivares et al., 2011). To souvisí rovněž i se shodou odpovědí na jednotlivé položky s modelem (outfit): kromě extrémně snadného (formální náležitosti) a obtížného kritéria (B2) všechny položky vyhovovaly Raschovu modelu příliš dobře (overfit). Nicméně v případě, že se zaměříme na infit vážený podle informační funkce, výsledek je velmi dobrý; žádná z položek se statisticky významně neliší od Raschova modelu (z-standardizované hodnoty jsou v rozmezí ±1,96) a χ^2/df statistika infitu leží v rozmezí 0,69–1,20, což lze vzhledem k velikosti vzorku hodnotit jako dobrý výsledek (Bond & Fox, 2009).

Závěrem lze říct, že pozorované odpovědi na položky byly celkově příliš deterministické; v případě, že se zaměříme jen na neextrémní odpovědi, tak položky Raschovu modelu vyhovovaly velmi dobře. Srovnání schopnosti respondentů, obtížnosti položek a přísnosti hodnotitelů včetně míry shody s modelem obsahuje obrázek 4.

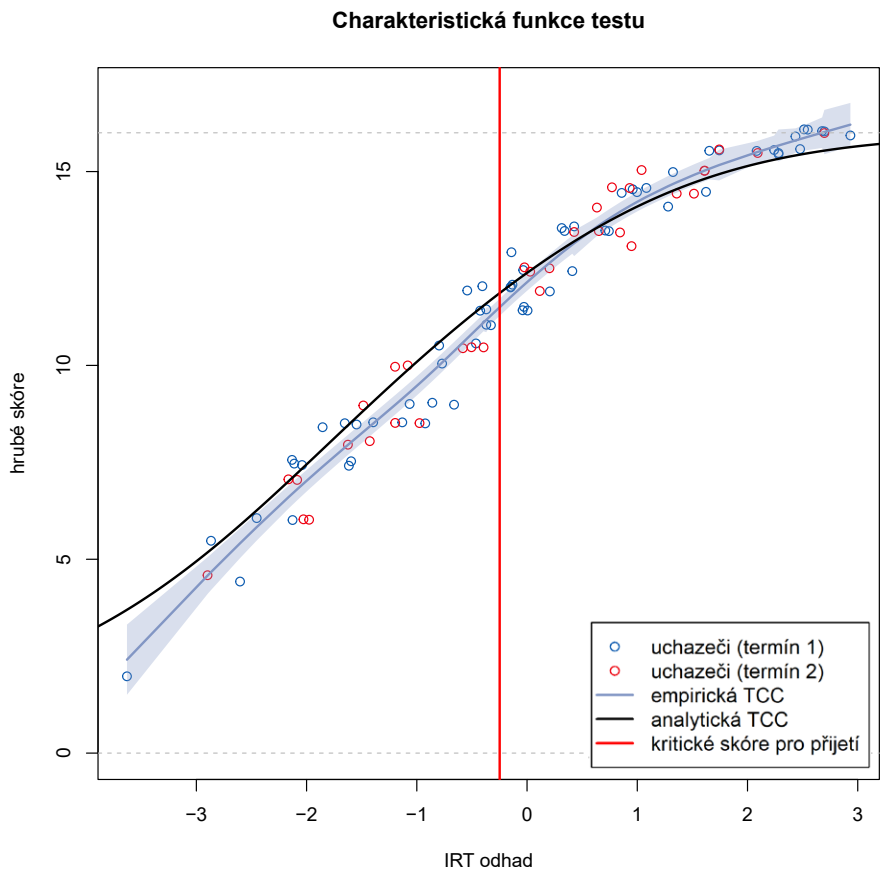
Podobné výsledky jsme pozorovali u uchazečů a hodnotitelů. Infit žádného z uchazečů ani hodnotitelů nebyl statisticky významný, z hlediska outfitu byl nápadný underfit u dvou hodnotitelů a v budoucnu by pravděpodobně bylo možné zvýšit reliabilitu hodnocení bakalářských prací jejich vyřazením z hodnoticího panelu.



Obrázek 4

Pozn.: Obtížnost hodnotitelů a položek je pro přehlednost vyjádřena v tradiční IRT metrice, tedy vyšší hodnota znamená obtížnější položku / přísnějšího hodnotitele. Velikost bodů znázorňuje chybu odhadu. Ta je na stejné škále napříč všemi body, nicméně není v měřítku vzhledem k úrovni latentního rysu. Body je proložena polynomická regresní přímka příslušné barvy; nikoli však pro hodnotitele, protože jejich rozmístění je spíše nahodilé.

Výsledné modelem implikované IRT skóry silně korelovaly s původním hrubým skórem (průměrné bodové hodnocení oběma hodnotiteli bez zvážení rozdílu v jejich shovívavosti). Pearsonova korelace byla $r = 0{,}969$ s 95% *CI* [0,954; 0,979], Spearmanova $\varrho = 0{,}986$ s 95% *CI* [0,979; 0,999]. Charakteristická funkce testu (TCC) je na obrázku 5.



Obrázek 5

Pozn.: Empirická charakteristická funkce testu (TCC) je polynomickou regresí pozorovaného hrubého skóru na IRT odhad latentního rysu. Analytická TCC je spočítána na základě parametrů modelu pro průměrně přísného hodnotitele. Právě rozdílná přísnost hodnotitelů je příčinou odlišnosti obou křivek. Aby se body nepřekrývaly, byl v případě hrubého skóru použit slabý jitter (body byly náhodně rozmlženy kolem jejich skutečné pozice na ose y).

4.5 Reliabilita a validita IRT modelu

Na základě modelu jsme odhadli reliabilitu parametrů schopnosti uchazeče, položky a hodnotitele. Také jsme korelovali odhady shovívavosti hodnotitelů a snadnosti položek odhadnuté separátně v obou termínech; jde o důležitý ukazatel test-retest reliability odhadu parametrů, a tedy i stability celého modelu. Výsledky jsou v tabulce 5. Reliabilita odhadu schopností uchazečů byla velmi dobrá, reliabilita odhadu obtížnosti položek pak výborná, což svědčí o dobré stabilitě modelu. Horší výsledek pozorujeme v případě reliability shovívavosti hodnotitelů. Korelace odhadů napříč termíny je nicméně uspokojivá: $r = 0{,}688$. Reliabilita hodnotitelů je však podhodnocena malou rozdílností v jejich přísnosti, viz tabulku 4, případně obrázek 4. Při pohledu na *RMSE* je patrné, že parametry hodnotitelů jsou reálně odhadnuté výrazně přesněji než v případě položek či uchazečů. Další vybrané analýzy pro ověření validity modelu jsou součástí online suplementu, přílohy 5.

Tabulka 5
Reliabilita IRT modelu

	termín 1		termín 2		dohromady		korelace termínů (test-retest)	
	$r_{xx'}$	*RMSE*	$r_{xx'}$	*RMSE*	$r_{xx'}$	*RMSE*	Pearson	Spearman
uchazeči	0,870	0,625	0,848	0,606	0,869	0,598		
hodnotitelé	0,576	0,318	0,486	0,426	0,703	0,286	0,688	0,711
položky	0,935	0,367	0,913	0,554	0,955	0,320	0,933	0,882

$r_{xx'}$ = odhad reliability; *RMSE* = standardizovaná průměrná chyba odhadu parametrů (Root Mean-Square Error)

4.6 Transformace IRT odhadu na výsledné hodnocení

Výstupem výše představeného IRT modelu byl odhad kvality bakalářské práce na logitové škále. Ten by bylo přímo možné využít pro seřazení uchazečů, nicméně jsme byli svázáni výše uvedenými formálními požadavky na přijímací řízení – IRT odhady latentního rysu proto bylo nutné převést na bodovou hodnotu.

Protože musely být body uděleny všem uchazečům (tj. i těm, kteří neprošli screeningovými kritérii), za každé screeningové kritérium byl udělen právě jeden bod. Po prvním termínu PZ jsme dospěli k rozhodnutí navrhnout k přijetí přibližně 35 uchazečů. Museli jsme vyhovět dvěma požadavkům, aby kritériem pro přijetí bylo 36 bodů a aby se body pohybovaly na škále zhruba 0–60. Protože čtyři body byly uděleny za screeningová kritéria, IRT odhady latentního rysu měly být přeškálovány do rozmezí přibližně 4 až 56

tak, aby 35. uchazeč dosáhl zhruba 32 bodů. Tohoto stavu jsme dosáhli s použitím následujícího algoritmu:

V prvním kroku jsme zjistili odhad latentního rysu 35. respondenta, který byl přibližně $\theta_{krit} = -0{,}248$. Ve druhém kroku byl určen počet N_{item} virtuálních binárních položek s obtížností $b_1 = 1{,}192$ (obtížnost první položky)[5], jejichž hypotetické absolvování by u daného uchazeče vedlo očekávanému hrubému skóru $c = 32$ bodů. Rovnici je možné zapsat jako

$$c = N_{item} \frac{\exp(\theta_{krit} + b_1)}{1 + \exp(\theta_{krit} + b_1)} \tag{2}$$

a její vyřešení pro počet položek vedlo k odhadu $N_{item} = 44{,}453$. Na základě těchto parametrů byl tedy odhadnut očekávaný pravý skór $E(\tau_p)$ každého respondenta p v testu jako

$$E(\tau_p) = \sum_{i=1}^{4} x_{ip} + N_{item} \frac{\exp(\theta_p + b_1)}{1+\exp(\theta_p + b_1)},\ N_{item} = 0, pokud\ \left(\sum_{i=1}^{4} x_{ip}\right) < 4, \tag{3}$$

kde χ_{ip} je počet získaných bodů na screeningovou položku i uchazečem p, θ_p je odhad jeho latentního rysu a N_{item} je konstanta definovaná výše. Pro uchazeče, kteří nezískali čtyři body ve screenigových kritériích, byla logistická část rovnice fixována na nulu. Výsledný počet bodů byl nakonec zaokrouhlen na celá čísla.

Výsledkem uvedené procedury byly udělené body v rozpětí 1–49 bodů s průměrem $M = 27{,}48$ ($SD = 17{,}23$) a mediánem $Mdn = 33$ (na prvním termínu). Navrženo k přijetí bylo přesně podle zadání 35 uchazečů (43 %), kteří získali 36 a více bodů.

4.7 Srovnání termínů a jejich vyvážení

Kritickým úkolem bylo zajistit férovost napříč oběma termíny. Za tímto účelem byl zcela nezávisle odhadnut totožný model pro druhý zkušební termín a byla ověřena shoda shovívavosti hodnotitelů a snadnosti položek. Korelace odhadů napříč termíny je reportována výše v tabulce 5, shoda byla rovněž ověřována vizuální inspekcí bodových grafů. Nepozorovali jsme u žádného nežádoucí artefakty či drift v odhadu parametrů pro žádného hodnotitele ani položku. Mimoto byl rozdíl v odhadech srovnán chí-kvadrát testem. Hodnotitelé se napříč termíny ve své shovívavosti statisticky význam-

[5] Důvodem použití koeficientu b_1 byla již dříve zmíněná analytická chyba při odhadu modelu s interceptem (který je ve zde prezentovaném modelu bez interceptu shodný právě obtížností první položky). Racionálnější by bylo zvolit koeficient b_1 buď nulový, nebo raději rovný průměrně obtížné položce.

ně nelišili: $\chi^2(18) = 6{,}10$; $p = 0{,}996$; a stejně tak ani snadnost položek: $\chi^2(15) = 14{,}67$; $p = 0{,}475$. Test-retestová korelace odhadů snadnosti položek byla navíc velmi vysoká: $r(14) = 0{,}933$.

Vzhledem k minimálním odlišnostem byly oba datasety sloučeny do jediného, $N = 130$, z nichž $n = 97$ uchazečů splnilo screeningová kritéria a bylo zahrnuto do finálního IRT modelu. Celá výše popsaná procedura byla beze změny replikovaná a LLTM model znovu odhadnut. To vše s tím rozdílem, že kritická hodnota latentního rysu pro přijetí, θ_{krit} nebyla odhadována znovu, ale převzali jsme ji z prvního termínu. Vyhodnocení druhého termínu na rozdíl od prvního proto bylo ryze kriteriální, čímž byla zajištěna zcela totožná náročnost PZ pro oba termíny.

Spearmanova korelace originálního počtu bodů u uchazečů z prvního termínu s novým odhadem byla téměř perfektní, $\varrho = 0{,}998$; změna v počtu bodů nebyla větší než tři body; posun o tři body v libovolném směru nastal pouze ve 2 případech (2 %), v 11 případech o dva body (13 %), ve 32 případech o jeden bod (39 %) a u 37 uchazečů (45 %) nedošlo k žádnému bodovému posunu. Rozdíl nebyl systematický, párový t-test $t(81) = -0{,}205$; $p = 0{,}838$ a nelišilo se ani rozložení bodů, Kolmogorův-Smirnoffův test $D = 0{,}116$; $p = 0{,}513$. Shodou okolností by nedošlo k žádné změně v rozhodnutí o přijetí u žádného uchazeče z prvního termínu (nikdo po replikaci procedury nepřekonal nebo naopak neklesl pod kritérium 36 bodů). Kdyby však k takové změně došlo, ponechali bychom rozhodnutí z prvního termínu.

Postup vedl k navržení 54 uchazečů (42 %) na přijetí do studia, z toho 19 nově navržených z druhého termínu. Pravděpodobnost přijetí byla shodná v obou termínech, Fisherův exaktní test $p(130) = 0{,}854$; a uchazeči se napříč termíny nelišili ani ve výsledném počtu bodů, Welchův t-test $t(96{,}3) = 0{,}675$; $p = 0{,}501$; ani v jeho rozložení, Kolmogorův-Smirnoffův test $D = 0{,}106$; $p(130) = 0{,}884$.

5. Diskuze

Výsledky naší studie ukazují, že hodnocení kvality studentských prací, v tomto případě bakalářských diplomových prací v rámci přijímacího řízení, může být realizováno s vysokou mírou reliability a férovosti. Reálná náročnost veškerých analýz není tak vysoká, jak se může zdát z předchozího textu; při rutinní práci by psychometrik patrně provedl méně kontrolních testů či analýzy méně dokumentoval. Naším cílem nicméně byla co nejvyšší transparentnost a rovněž i podpora dobré praxe v České republice skrze tento článek.

Nejpřínosnějším zjištěním naší studie z hlediska zajištění férovosti hodnocení je fakt, že jednotliví hodnotitelé se příliš nelišili svou přísností a rovněž mezi nimi panovala vysoká shoda v tom, co je dobrá a co špatná práce. Přesto jsme ve výsledném hodnocení zohlednili jejich přísnost a výsledná

reliabilita ve smyslu vnitřní konzistence byla nad očekávání dobrá. Hodnota $r_{xx'} = 0{,}87$ je plně srovnatelná, či dokonce lepší než reliabilita písemných testů v předchozích semestrech, případně vnitřní konzistence didaktických testů státní maturity (CERMAT, 2020). V souladu s doporučením Winda a Petersona (2018) jsme nicméně shodu hodnotitelů prozkoumali s využitím více různorodých kritérií.

Výše uvedená reliabilita se vztahuje na IRT odhady faktorových skórů, které však těsně korelují s prostým součtem bodů v jednotlivých kritériích. Reliabilita prostého součtu prizmatem teorie zobecnitelnosti je rovněž vysoká, a to jak po vyvážení na přísnost hodnotitelů ($\varrho^2 = 0{,}85–0{,}86$), tak bez něj ($\Phi = 0{,}82$). Je proto důležité si uvědomit, že náš poměrně komplikovaný IRT model nepřinášel podstatně více informací oproti výrazně jednoduššímu postupu vyhodnocení. Lze sice uvažovat, že v případě variabilnějších hodnotitelů či více zešikmených dat by byl IRT odhad výrazně vhodnější, ale pro podporu této domněnky nejsou k dispozici data. Nutno však podotknout, že odhad IRT reliability může být mírně podhodnocený v důsledku zkreslené práce s rozptylovými komponentami (Cai & Hansen, 2013; Maydeu-Olivares et al., 2011). O tom svědčí rovněž fakt, že po korekci na nereliabilitu je test-retest korelace odhadů přísnosti hodnotitelů vyšší než 1, konkrétně 1,30, což je nesmysl; a nejpravděpodobnější příčinou je prosté podhodnocení reliability. Protože přesnost odhadů parametrů uchazečů souvisí s přesností odhadu benevolentnosti hodnotitelů, lze se domnívat, že ta bude podhodnocena rovněž. Celkově je vnitřní konzistence i shoda posuzovatelů shodná či vyšší, než reportují jiné studie u hodnocení textového materiálu (Breland et al., 1999; Salvatori, 2001). To není překvapivé vzhledem k délce, rozsahu a náročnosti přípravy bakalářské práce ve srovnání s typickými esejemi. Svou roli může hrát i podoba volených kritérií.

Nižší se zdá být shoda hodnotitelů na screeningových kritériích, tedy na vyřazení prací z dalšího hodnocení. Při bližším pohledu na příčiny této neshody je však patrné, že hodnotitelé si určitých nedostatků byli vědomi a reflektovali je, či se přímo obraceli na přijímací komisi s dalšími doporučeními k revizi jejich vlastního hodnocení. Zároveň v případě, že se posouzení obou hodnotitelů lišilo, zpravidla i ten z nich, který práci ve screeningových kritériích hodnotil dostatečně, vyjadřoval ve slovním komentáři určitou nespokojenost a zároveň intenci studenta nepoškodit a chybovat spíše v jeho prospěch. Lze se navíc oprávněně domnívat, že práce, u nichž byly pochybnosti ve screeningových kritériích, by neuspěly v „plném" hodnocení; není např. dost dobře možné napsat kvalitní text bez použití adekvátní literatury. Tomu odpovídá i velký rozdíl v průměrném počtu bodů hrubého skóru přijatých a nepřijatých prací. Případná neshoda v screeningových kritériích by tak neměla mít vliv na přijetí uchazeče; bylo přijato 42 % uchazečů, screeningem jich však neprošlo pouze 25 %.

Velmi dobrou zprávou byl fakt, že odhad parametrů modelu i výsledné pořadí uchazečů se prakticky nelišilo při použití modelů z prvního, druhého nebo obou termínů dohromady. Po přepočítání modelu na datech všech uchazečů by dokonce nedošlo k potenciální změně žádného z rozhodnutí o přijetí, změna v počtu bodů byla zcela zanedbatelná. Je tedy patrné, že samotné udělení bodů je kriteriální a stabilní v čase, a navržené hodnoticí schéma je tak možné použít podobnou populací hodnotitelů i v budoucnu. Předem stanovené kritérium lze tak použít pro kriteriální rozhodování o přijetí či nepřijetí namísto prostého pořadí uchazeče.

Určitá omezení skýtal formát přijímacího řízení, který byl formálně vyžadován fakultou a univerzitou; zejména kombinace normativního (kapacita oboru) a kriteriálního (36bodová hranice) rozhodování. Převod IRT odhadu na bodovou hodnotu není příliš transparentní a je zbytečný; v budoucnu by bylo lepší se bez něj obejít. Přesto však IRT odhady byly férově převedeny na arbitrární bodovou škálu, a ta může být pro uchazeče a stake-holdery srozumitelnější než číslo s řadou desetinných míst. Zásadnějším limitem je to, že v rozporu s běžnými doporučeními (AERA et al., 2014; Zamanzadeh et al., 2020) je přijetí založeno pouze na jediné zkoušce. Bohužel to však odpovídá běžné situaci, tedy jedinému znalostnímu testu, a omezeným zdrojům našeho pracoviště.

Za jeden z hlavních přínosů práce považujeme samotné navržení hodnoticích kritérií. I když diskuse o hodnotitelných kvalitách prací probíhá díky jejich obhajobám prakticky neustále, jen zřídka se naskytne situace, která by ospravedlnila investici do rigorózní studie relevance a psychometrického potenciálu jednotlivých hodnoticích kritérií. Výsledky, které uvádíme, zejména ukazatele reliability, svědčí o tom, že lze najít shodu v kritériích napříč různě orientovanými psychology a různými typy prací. Zároveň lze naleznout kritéria, které umožní jejich jednodimenzionální a zároveň komplexní hodnocení.

Je až s podivem, že naše směs formálních a věcných kritérií se chovala jednodimenzionálně. Domníváme se, že jde o něco, co je poplatné současné situaci v oboru psychologie, kdy i formální aspekty prací, jako je například počet cizojazyčných citovaných zdrojů, indikují kvalitu. Aktuálně korelují s obsahovou kvalitou, ale za nějakou dobu už nemusí, protože je například začnou dodržovat všichni (nebo nikdo). Je tedy důležité upozornit na důležité zpětnovazební procesy a vývoj v oboru, které mohou přínosnost jednotlivých kritérií a validitu měřítka z nich složeného dramaticky proměnit, podobně jako prosakování položek psychodiagnostických testů do didaktických materiálů na všech úrovních škol. I když tedy zde prezentované hodnocení nebude možné beze změn opakovaně používat po mnoho let (a snad ubudou situace, které by to vyžadovaly), věříme, že i do budoucna bude možné stejným postupem vytvořit v dané chvíli stejně dobře fungující kritéria.

V tomto ohledu může být nevýhodou, že se navržená kritéria liší od běžných kritérií během obhajoby bakalářské práce a rovněž tato kritéria nemusí být vnímána stejně na různých pracovištích. Docházelo tak k situacím, kdy uchazeči s výbornou známkou z obhajoby získali nižší výsledek a nebyli přijati. Tuto diskrepanci jsme se pokoušeli uchazečům vysvětlit právě skrze transparenci celého hodnoticího procesu. Při vyřizování případných odvolání jsme také uchazečům poskytovali podrobnou informaci o tom, za která kritéria bylo uděleno kolik bodů. Pokud by se vnímání námi navržených kritérií výrazně lišilo napříč různými katedrami psychologie, věříme, že náš text tak může podnítit diskuzi o tom, jakými aspekty se kvalitní bakalářské práce vyznačují. Je možné, že některá kritéria z námi navržených by na jiných pracovištích nemusela být oceňovaná – a určitý konsenzus by byl všestranně přínosný.

Dalším limitem navržené přijímací zkoušky je nesporný fakt, že kvalita (a tedy i hodnocení) bakalářských prací je závislá nejen na studentovi, ale zčásti i na jeho vedoucím. Je evidentní, že někteří vedoucí bakalářských prací jsou studentům více nápomocni, a mohou tak „přispět" k vyšší šanci na přijetí.

Na základě předložených dat přesto soudíme, že námi navržený způsob hodnocení bakalářských prací je férový, validní a reliabilní, přestože prediktivní validitu bude možné ověřit až v budoucnu. Zároveň věříme, že může být inspirací pro jiná pracoviště, a to nejen v době koronavirové epidemie; navržený způsob hodnocení je možné využít kdykoli a s libovolnými daty – např. i seminárními pracemi, slohovými pracemi či při hodnocení jakéhokoli jiného kvalitativního produktu. Za tímto účelem jsme poskytli podrobnou dokumentaci celého postupu včetně dat a analytických skriptů, které nabízíme k veřejnému použití. Pro účely aplikace našeho postupu na jiných pracovištích sdílíme i informace o časové náročnosti pro hodnotitele a další doplňková zjištění, která jsou součástí online suplementu (příloha 6). Předchozí text nicméně nelze vnímat jako validizaci námi navržených kritérií pro používání v běžné praxi. Domníváme se, že jejich dlouhodobé využívání v nezměněné podobě by mohlo vést k budoucím úpravám bakalářských prací tak, aby splňovaly arbitrárně vybraná kritéria a k zafixování jejich podoby v čase. V případě potřeby doporučujeme vyvinout si svá vlastní kritéria a používat je po omezený čas, případně je neustále mírně upravovat.

V každém případě se však zvolený formát přijímacího řízení (který hodnotí spíše specifické zaměření autora a jeho dílčí dovednosti) odlišuje od běžného písemného testu, který ověřuje spíše obecné znalosti a jejich šíři. Je otázkou, nakolik se bude lišit populace studentů jednou či druhou podobou přijímací zkoušky; tyto rozdíly by bylo velmi vhodné v budoucnu prozkoumat.

Literatura

AERA, APA, & NCME. (2014). *Standards for educational and psychological testing.* American Educational Research Association.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Bentler, P. M., & Woodward, J. A. (1980). Inequalities among lower bounds to reliability: With applications to test construction and factor analysis. *Psychometrika*, *45*(2), 249–267. https://doi.org/10.1007/BF02294079

Bentler, P. M. (2009). Alpha, dimension-free, and model-based internal consistency reliability. *Psychometrika*, *74*(1), 137–143. https://doi.org/10.1007/s11336-008-9100-1

Bond, T. G., & Fox, C. M. (2009). *Applying the Rasch model : Fundamental measurement in the human sciences.* Lawrence Erlbaum Associates, Inc.

Breland, H. M., Bridgeman, B., & Fowles, M. E. (1999). Writing assessment in admission to higher education: Review and framework. *ETS Research Report Series*, *1999*(1), 1–42. https://doi.org/10.1002/j.2333-8504.1999.tb01801.x

Brennan, R. L. (2001). *Generalizability theory.* Springer-Verlag.

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

Casabianca, J. M., Junker, B. W., & Patz, R. J. (2016). Hierarchical rater models. *Handbook of Item Response Theory*, *1*, 449–465. https://doi.org/10.1201/9781315374512

CERMAT. (2020). *Neagregovaná položková data didaktických testů maturitní zkoušky*. Portál s výsledky evaluačních projektů. https://vysledky.cermat.cz/statistika/Default.aspx

Cígler, H. (2018). *Matematické schopnosti: Teoretický přehled a jejich měření* (EDIS). Masarykova univerzita.

Cígler, H., Ježek, S., Širůček, J., & Lacinová, L. (2020). *Hodnocení bakalářských prací jako přijímací kritérium do navazujícího magisterského studia: Psychometrická kazuistika*. Masarykova univerzita. https://doi.org/10.17605/OSF.IO/QX5U7

Cronbach, L. J., Rajaratnam, N., & Gleser, G. C. (1963). Theory of generalizability: A liberalization of reliability theory. *British Journal of Statistical Psychology*, *16*(2), 137–163. https://doi.org/10.1111/j.2044-8317.1963.tb00206.x

De Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., & Partchev, I. (2012). The estimation of item response models with the lmer function from the lme4 package in R. *Journal of Statistical Software*, *39*(12), 1–28.

Fischer, G. H. (1973). The linear logistic test model as an instrument in educational research. *Acta Psychologica*, *37*(6), 359–374. https://doi.org/10.1016/0001-6918(73)90003-6

Gagolewski, M. (2020). *R package stringi: Character string processing facilities.*

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika*, *30*, 179–185. https://doi.org/10.1007/BF02289447

Charvát, M., & Viktorová, L. (2014). *Tvorba, administrace a analýza testů studijních předpokladů*. Univerzita Palackého v Olomouci. https://www.researchgate.net/publication/291295001

Maydeu-Olivares, A., Cai, L., & Hernández, A. (2011). Comparing the fit of item response theory and factor analysis models. *Structural Equation Modeling: A Multidisciplinary Journal*, *18*(3), 333–356. https://doi.org/10.1080/10705511.2011.581993

Nakagawa, S., Johnson, P. C. D., & Schielzeth, H. (2017). The coefficient of determination R 2 and intra-class correlation coefficient from generalized linear mixed-effects models revisited and expanded. *Journal of The Royal Society Interface*, *14*(134), 20170213. https://doi.org/10.1098/rsif.2017.0213

Patz, R. J., Junker, B. W., Johnson, M. S., & Mariano, L. T. (2002). The hierarchical rater model for rated test items and its application to large-scale educational assessment data. *Journal of Educational and Behavioral Statistics*, *27*(4), 341–384. https://doi.org/10.3102/10769986027004341

Phillips, N. (2017). *Yarrr: A companion to the e-Book "YaRrr!: The pirate's guide to R"* (R package version 0.1.5). https://cran.r-project.org/package=yarrr

R Core Team. (2020). *R: A language and environment for statistical computing* (4.02). R Foundation for Statistical Computing.

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research* (2.0.9). Northwestern University.

Revelle, W., & Zinbarg, R. E. (2009). Coefficients Alpha, Beta, Omega, and the glb: Comments on Sijtsma. *Psychometrika*, *74*(1), 145–154. https://doi.org/10.1007/s11336-008-9102-z

Salvatori, P. (2001). Reliability and validity of admissions tools used to select students for the health professions. *Advances in Health Sciences Education*, *6*(2), 159–175. https://doi.org/10.1023/A:1011489618208

Sarkar, D. (2008). *Lattice: Multivariate data visualization with* R. Springer. http://lmdvr.r-forge.r-project.org

Signorell, A., Aho, K., Alfons, A., Anderegg, N., Aragon, T., Arachchige, C., Arppe, A., Baddeley, A., Barton, K., Bolker, B., Borchers, H. W., Caeiro, F., Champely, S., Chessel, D., Chhay, L., Cooper, N., Cummins, C., Dewey, M., Doran, H. C., … Zeileis, A. (2020). *DescTools: Tools for descriptive statistics* (R package version 0.99.32). https://cran.r-project.org/package=DescTools

Sijtsma, K. (2009). On the use, the misuse, and the very limited usefulness of Cronbach's Alpha. *Psychometrika*, *74*(1), 107–120. https://doi.org/10.1007/s11336-008-9101-0

Tutz, G. (1990). Sequential item response models with an ordered response. *British Journal of Mathematical and Statistical Psychology*, *43*(1), 39–55. https://doi.org/10.1111/j.2044-8317.1990.tb00925.x

Wickham, H. (2007). Reshaping data with the {reshape} package. *Journal of Statistical Software2*, *21*(12), 1–20. http://www.jstatsoft.org/v21/i12/

Wind, S. A., & Peterson, M. E. (2018). A systematic review of methods for evaluating rating quality in language assessment. *Language Testing*, *35*(2), 161–192. https://doi.org/10.1177/0265532216686999

Zamanzadeh, V., Ghahramanian, A., Valizadeh, L., Bagheriyeh, F., & Lynagh, M. (2020). A scoping review of admission criteria and selection methods in nursing education. *BMC Nursing*, *19*(121), 1–17. https://doi.org/10.21203/rs.3.rs-31484/v1