Univerzita Karlova v Praze

Filozofická fakulta

Katedra Psychologie

Bakalářská práce

Nikola Frollová

Metodologické inovace v psychologickém výzkumu

Methodological innovation in psychological research

Praha, 2016

Vedoucí práce: Ing. Mgr. Marek Vranka

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

V Praze dne 20. 4. 2014 ……

Nikola Frollová

Ráda bych zde poděkovala mému vedoucímu Ing. Mgr. Vrankovi za trpělivost a péči a také kamarádovi a spolužákovi Šimonovi Kucharskému za podporu.

Abstrakt:

Rok 2012 otevřel intenzivní diskusi o metodologických problémech psychologického výzkumu, jež vedou k nadměrné publikaci málo spolehlivých a replikovatelných výsledků. Navrhovaná zlepšení spočívají zejména v realizaci replikačních studií, otevřeného sdílení dat, omezování „stupňů volnosti výzkumníka“ a zlepšení porozumění a aplikace statistických metod pro analýzu dat. Bakalářská práce má za cíl zmapovat zmíněné problémy, navrhovaná řešení jako i jejich možnou kritiku a problémy s implementací do praxe. Návrh výzkumu spočívá v měření prevalence sporných vědeckých postupů mezi psychology v České republice.

Klíčová slova:

Replikace, předregistrace, pochybné výzkumné praktiky, publikační zkreslení

Abstract:

The year 2012 opened an intense debate about the methodological problems of psychological research that lead to excessive publication of unreliable and nonreplicable results. The proposed improvements consist of conducting more replication studies, sharing data openly, limiting the "degrees of researcher freedom ", and improving the understanding and application of statistical methods for data analysis. This bachelor thesis aims to map the aforementioned problems, proposed solutions, as well as possible criticisms and problems with their implementation in practice. The research proposal consists of measuring the prevalence of questionable scientific practices among psychologists in Czech Republic.

Keywords:

Replication, pre-registration, questionable research practices, publication bias

Obsah

Literárně přehledová část

Úvod

Pochybování o psychologii ve smyslu vědní disciplíny není nic nového. Důvodem je zřejmě fakt, že něčemu jako je lidská se psychika se velice těžko nastavují měřitelné parametry. Jak se dá například změřit stres? Můžeme s jistou říci, že ho operacionalizujeme správně a tedy volíme vhodné nástroje k jeho měření? Srovnáme-li to například s exaktnějšími vědami, u nichž nám k udělání závěru postačí třeba jen mikroskop- například k prokázání existenci mikroorganismu. To psychologii dělá zranitelnější a výzkum náchylnější k chybám a následně pak další skepsi o možnostech zkoumání psychiky jako takové.

Rádi bychom věřili, že výsledky publikovaných psychologických výzkumů lze s důvěrou využívat při navrhování nových studií nebo je přímo aplikovat v praxi. Nejnovější události v akademické sféře však opět otřásly základy této důvěry. Vypadá to, že se teď nacházíme v situaci, kdy je zřejmě nutné používané výzkumné praktiky přehodnotit.

Tato bakalářská práce má dva cíle. Následující stránky by měly poskytnout jednak základní přehled problematických aspektů v psychologickém výzkumu a také přehled metodologických opatření, které by mohly psychologický výzkum zkvalitnit. Cílem práce není podat vyčerpávající a komplexní výklad všech teoretických konceptů, jež s touto problematikou souvisí. Práce se zaměřuje na pokrytí moderních teorií a současných vědeckých názorů v oblasti metodologie psychologického výzkumu.

Nejprve popíšu pochybné výzkumné praktiky výzkumníků, k jejichž používání vede dále rozebírané publikační zkreslení, a jaké dopady má jejich používání na celkové vědecké poznání. Dále pak rozeberu replikace jako možný prostředek pro spolehlivé zjištění skutečných efektů. Nakonec zmíním předregistraci jako jedno z možných řešení vedoucí k odstranění popsaných problémů.

V návaznosti na problematiku nastíněnou v teoretické části práce, obsahuje druhá část této práce návrh kvantitativního výzkumu pochybných výzkumných praktik v českém akademickém a výzkumném prostředí.

1. Krize ve vědě

O aktuálních vědeckých poznatcích se dočteme v recenzovaných odborných časopisech, které by měly být jakousi zárukou jejich kvality. Nedávné kauzy však důvěryhodnost výsledků publikovaných v časopisech významně narušily. Je tedy potřeba podívat se kriticky na to, podle jakého klíče jsou články do časopisů vybírány, a jak se to odráží na jejich důvěryhodnosti.

Chris Chambers, experimentální psycholog působící na Cardiff University ve Velké Británii se nechal slyšet, že: *„Ve vysoce impaktovaných časopisech považují často psychologii za jakousi show plnou zábavných triků.“* (Yong, 2012).

Mezi události, díky nimž se vznesla otázka, s jakou důvěrou můžeme přistupovat k prestižním časopisům, patří například odhalení nizozemského psychologa Diederika Stapela. Jeho hvězdná kariéra s desítkami prestižních publikací byla ve skutečnosti založena na zfalšovaných datech (Enserink, 2011). Další událostí byla publikace studie Daryla Bema údajně přinášející důkazy o schopnosti předvídat budoucnost ve významném odborném časopise *Journal of Personality and Social Psychology* (Bem, 2011). Studie byla záhy podrobena kritice (Wagenmakers, Wetzels, Borsboom, & van der Maas, 2011), jež ukázala, že zjištěné výsledky jsou s vysokou pravděpodobností pouhým produktem pochybných výzkumných praktik (ang. *questionable research practicies*), jako je například používání vícero proměnných a analýz, z nichž se pak do publikace vyberou pouze ty, u nichž se podařilo zjistit statisticky významný výsledek (Simmons, Nelson, & Simonsohn, 2011). Jak ale ukázal následný průzkum mezi psychology (John, Loewenstein, & Prelec, 2012), tyto praktiky vůbec nejsou ojedinělé, pouze se o nich a jejich negativním vlivu na důvěryhodnost závěrů obvykle dostatečně nemluví.

Časová koexistence zmíněných kauz způsobila iniciaci mnohých změn, jednak v recenzním a publikačním procesu, ale především také při navrhování samotných studií. Zdá se, že klíčem ke zlepšení výzkumné praxe v psychologii je osvojení si principů otevřené vědy – to znamená zcela transparentní zveřejňování kompletních informací o realizovaných výzkumech, včetně použitých materiálů a nasbíraných dat, a to i v případě, že výsledky nebyly publikovány. Jelikož lze čekat, že k tak radikálnímu posunu od způsobů, jak se informace o realizovaných studiích zveřejňují dnes, dojde až v delším časovém horizontu, v mezidobí lze zvýšit důvěryhodnost publikovaných studií pomocí před-registrací. Provést inventuru v dosud publikovaných výsledcích a vytřídit zrno od plev zase umožní větší ochota

provádět a publikovat replikační studie. O předregistracích i replikacích je podrobněji pojednáno v dalších částech této práce.

V následující kapitole bude ale nejdřív představena jedna z hlavních příčin současných problému – výzkumné praktiky zvyšující riziko získaní nespolehlivých a pouze iluzorních výsledků.

2. Pochybné výzkumné praktiky

Ne všechna pochybení ve výzkumu musí být přímo vědomými podvody, jako tomu bylo v případě Stapela. Mnohem větší část z nich spadá do kategorie tzv. pochybných výzkumných praktik (PVP). Mezi PVP například patří selektivní publikování pouze některých experimentálních manipulací, zastavení sběru dat na základě získání statisticky významného výsledku, zaokrouhlování p-hodnot, post-hoc interpretace a tvorby hypotéz (prezentování explorační analýzy jako analýzy konfirmační), manipulace s odlehlými hodnotami (odstranění odlehlých hodnot nebo naopak jejich ponechání, pokud vedou k statisticky významnému výsledku) (Fiedler & Schwarz, 2015).

P-hacking je hovorovým výrazem pro sérii postupů, které zásadně ovlivňují důvěryhodnost výsledků statistické analýzy. Jedná se o proces statistické analýzy, kdy se tak dlouho zkouší různé možnosti, dokud se nedocílí požadované hodnoty statistické významnosti – většinou se jedná o překročení hranice $p = 0{,}05$. Zjednodušeně řešeno je p-hacking sérií PVP. Ve studii má výzkumník možnost sám dělat mnoho rozhodnutí o klíčových aspektech, například o tom, kolik probandů sesbírá, které proměnné bude měřit, a jak bude analyzovat výsledky a také kdy – jestli v průběhu experimentu nebo po konci sběru dat. Těmto širokým možnostem se někdy také říká stupně volnosti výzkumníka, v narážce na „stupně volnosti“ známé ze statistiky. Zmíněné volby mohou být provedeny nevinně a s čistým úmyslem, ale dávají vědcům svobodu manipulovat parametry analýzy do té doby, než je přivedou k pozitivním výsledkům.

Například se nabízí otázka, jak si ve studii poradit s odlehlými hodnotami? Simmons et al. (2011) pročetli zhruba 30 článku z různých psychologických žurnálů a zjistili velké rozdíly v postupech při vyřazování odlehlých hodnot. Většina výzkumníků vyloučila například některé odpovědi ze studie, pokud byly zvoleny příliš rychle. Ovšem otázkou zůstává, co to znamená “příliš rychle“. Pro některé autory to byly reakční časy o dvě směrodatné odchylky odlišné od průměru, pro jiné to byl reakční čas rychlejší než 200ms. Rozhodování o těchto specifikacích nemusí být vždy nutně špatně, nicméně může být

potencionálním zdrojem pro dále zmíněný koncept sebestředného ospravedlnění se (ang. *self-serving justifications*) (Simmons, Nelson, & Simonsohn, 2011).

Běžné postupy, které vedou k p-hackingu zahrnují: Provedení analýzy v polovině experimentů a následovné rozhodnutí, zda pokračovat ve sběru dat; sběr mnoha různých proměnných a následně rozhodnutí, které z nich zahrnout do analýzy, rozhodování o tom, zda zahrnout odlehlé proměnné do analýzy na základě statistické významnosti a zahrnutí nebo vyloučení mediátorů či moderátorů na základě post-hoc analýzy (Head, Holman, Lanfear, Kahn, & Jennions, 2015). P- hacking nemusí být vždy úmyslný. Gelman a Loken (2013) ve svém článku *The garden of forking paths* používají metaforu větvících se cest, kde křižovatky symbolizují rozhodnutí o příslušném analytickém kroku. I když výzkumník neprovede u každé křižovatky sérii analýz a až poté si podle výsledku zvolí, kterou cestou půjde, a má pocit, že všechna rozhodnutí udělal nejčestněji, jak mohl, je tato cesta velmi problematická. Ačkoliv se cesta zdá být přímá, tedy bez počítání různých analýz a následného výběru těch pro studii nejlichotivějších, velké množství rozhodnutí, které má výzkumník zdánlivě korektně vyargumentované, vychází z toho, jak konkrétní data vypadají (Gelman & Loken, 2013).

Gelman a Loken (2013) popisují různé způsoby, jak vědci mohou analyzovat vztahy mezi proměnnými. Buď mohou použít jednoduchý t-test či ANOVU například s kontrolou pohlaví, protože si výzkumník předem určí, že zrovna tuto proměnou z nějakého důvodu potřebuje kontrolovat – to jsou praktiky, které by většina akademické obce shledala za legitimní. Pak přicházejí způsoby, které jsou považovány za PVP (Francis, 2013; Simmons et al., 2001). Patří sem například postup provedení testu ANOVA s kontrolou pohlaví, jestliže až na základě analýzy výzkumník zjistí, že je tam nějaká asociace s pohlavím. Problémem je, že s tímto vztahem explicitně výzkumník nepočítá dopředu a stejně tak by tam mohl zpětně zařadit i kontrolu právě věku, kdyby v něm zjistil spojitost. Nejpochybnějším způsobem pak je, když výzkumník zkouší různé verze testu ANOVA s různými kontrolami a interakcemi, a následně sleduje, které vyjdou nejlépe, a podle toho postup analýzy popíše. Tomuto druhu p-hackingu se někdy říká cherry picking či fishing neboli lovení v datech. V praxi to vypadá tak, že výzkumníkovi nevyjde efekt, který zkoumá, ale protože ví, že publikovat negativní nález by bylo obtížné či přímo nemožné (viz publikační zkreslení popsané v další kapitole), vyřeší to tím, že nejdříve udělá velkou korelační matici, najde proměnné, co spolu souvisejí, zkoumá mezi nimi vztah a pak článek napíše stylem, kde tento vztah prezentuje a tváří se, že za tímto účelem od začátku sbíral data (Gelman & Loken, 2013).

Simmons a kolegové (2011), ve své studii ukazují, že i když psychologové v článku předpokládají malou pravděpodobnost falešné pozitivity nálezu (neboli nesprávné zamítnutí nulové hypotézy), flexibilita při sběru dat, analýze výsledků a jejich reportování ve skutečnosti tuto pravděpodobnost významně zvyšuje. Ukázalo se, že v mnoha případech existuje větší pravděpodobnost najít důkaz, že vliv existuje, než správně najít důkazy o tom, že tomu tak není (Simmons, Nelson, & Simonsohn, 2011).

Simmons s kolegy pro demonstraci tohoto problému uskutečnili dva experimenty, které byly navrženy tak, aby prokázaly falešný efekt: konkrétně, že některé písničky dokážou změnit věk posluchače. V první studii zkoumali na vysokoškolských studentech, zda poslech dětské písně vyvolává věkový kontrast, a lidé se pak tudíž budou cítit starší. Vybrali dětskou píseň a kontrolní instrumentální skladbu. Po poslechu části skladby, účastníci výzkumu na škále odpověděli, jak staří se momentálně cítí. Další kontrolní proměnnou byl věk otce participantů, aby se kontrolovaly rozdíly věků mezi účastníky. Analýza kovariance ukázala, že účastníci se po dětské písničce skutečně cítili subjektivně starší.

Ve druhé studii zkoumali obrácený efekt: zdali se po poslechu skladby vyvolávající stáří, budou účastníci cítit mladší. Použili stejný design jako v první studii, navíc s kontrolou věku otce na jejich skutečný věk. Ukázalo se, že posloucháte-li píseň song *When I'm Sixty-four* od Beatles, sníží se subjektivně prožívané stáří posluchače o rok a půl.

Tyto dvě studie byly provedeny s reálnými účastníky, experimentátory a s legitimní statistickou analýzou. Nicméně nesmyslná hypotéza se potvrdila, protože věk participantů s kontrolou věku otce nebyly zdaleka jediné proměnné, které byly ve studii reálně měřeny. Do studie byly zahrnuty desítky jiných proměnných od věku matky, přes náladu až chuť k jídlu. Autoři zkrátka provedli spoustu různých analýz, se spoustou závislých proměnných a různými kontrolami skupin, ale v demonstrativním článku reportovali pouze ty proměnné, které jim vyšly signifikantně.

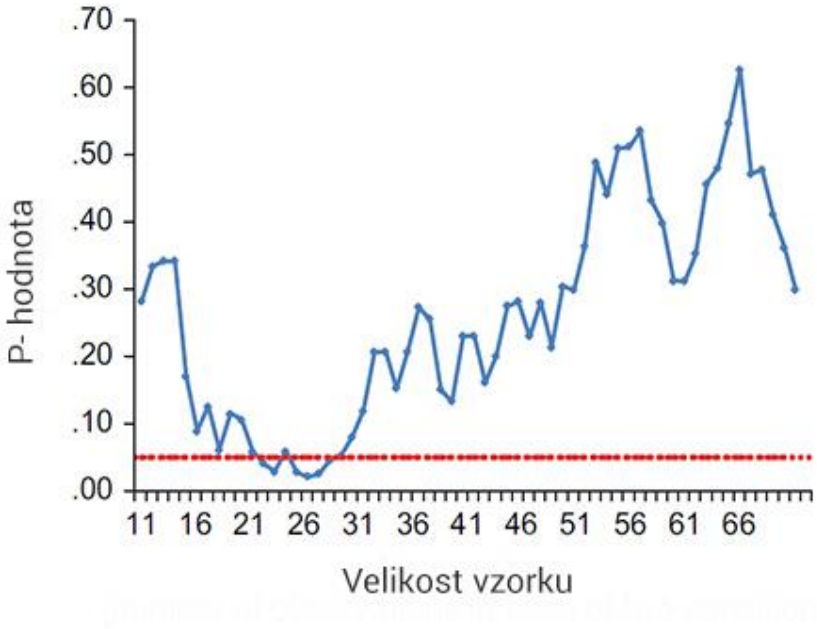
Simmonse tedy zajímalo, jak stupně volnosti výzkumníků mohou ovlivnit statistickou významnost. Použil počítačové simulace experimentálních dat pro odhad toho tohoto ovlivnění a došel k těmto čtyřem "stupňům volnosti", které nejvíce ovlivňují výzkum: (a) vybírání si mezi závislými proměnnými, (b) flexibilní volba velikosti vzorku (c) přidání kontrolující proměnné a (d) reportování vybraných experimentálních podmínek, které vyšly signifikantně (Simmons, Nelson, & Simonsohn, 2011).

Pro demonstraci rozsahu a závažnosti tohoto problému Simmons s kolegy provedli počítačovou simulaci vymyšlených dat, kde generovali náhodné vzorky a podrobovali je různým statistickým exploračním analýzám, a pozorovali, jak často je alespoň jedna z výsledných p-hodnot v každém vzorku nižší než hladina významnosti, respektive za jakých okolností se výzkumník může dostat za hranici signifikantního výsledku. Simulované data samozřejmě ve skutečnosti pocházela z distribucí, které se vzájemně vůbec nelišily, a tedy jakýkoliv zjištěný statisticky významný rozdíl byl nutně pouze falešným pozitivem. Simmons s kolegy simuloval například důsledky použití více závislých proměnných. Situace, kdy výzkumník měří dvě proměnné a reportuje jen tu, která vyjde signifikantně. Tak se zdvojnásobí šance na falešně pozitivní výsledek. (viz tabulka I). Následně do statistických postupů výzkumník vybere ten test, který vychází nejlépe.

	Signifikance		
Stupně volnosti výzkumníka	**p < .1**	**p < .05**	**p < .01**
Situace A: dvě závislé proměnné (r = .50)	17.80%	9.50%	2.20%
Situace B: přidání 10 pozorování	14.50%	7.70%	1.60%
Situace C: kontrola pohlaví a interakce pohlaví s intervencí	21.60%	11.70%	2.70%
Situace D: Vyloučení jedné ze tří podmínek	23.20%	12.60%	2.80%
Kombinace situací A a B	26.00%	14.40%	3.30%
Kombinace situací A, B, a C	50.90%	30.90%	8.40%
Kombinace situací A, B, C, a D	81.50%	60.70%	21.50%

Tabulka I. *Pravděpodobnost získání falešně pozitivních výsledků. Převzato a přeloženo z Simmons, Nelson & Simonsohn (2011).*

Kromě toho může výzkumník manipulovat s počtem účastníků ve studii na základě statistické signifikanci při průběžné analýze dat. Simmonsova simulace ukazuje, jak se s touto manipulací p-hodnota dramaticky mění (viz tabulka I). Po každém přidání účastníka, výzkumník provede t-test. Dostane-li se na požadovanou p-hodnotu, sběr dat ukončí. To je samozřejmě problematické, protože jsou jednotlivé hodnoty z výběru náhodné, souhrnná statistika (většinou průměr) fluktuje. Pokud děláme statistický úsudek po každém dalším výběru, zvyšujeme riziko chyby prvního druhu (zvláště pak s malými vzorky, kdy jedna odlehlá hodnota může velmi ovlivnit celkovou statistiku) tyto fluktuace mohou v některých fázích výběru vykazovat signifikantní efekt, přičemž jsou ale pouze výsledkem těchto fluktuací.



Graf I. *Ilustrativní simulace p-hodnot získaných výzkumníkem, který průběžně přidává pozorování. Po každém přidání účastníka provede t-test. Přerušovaná čára je kritérium p ≤ .05. Převzato, přeloženo a upraveno z Simmons, Nelson & Simonsohn (2011)*

Přibližně 70% vědců se k této praktice přiznalo (John, Loewenstein, & Prelec, 2011). Řešením by mělo být uvést velikost vzorku dopředu.

Další velice oblíbenou praktikou je přidání nějaké kontrolované proměnné, velice často vybíranou je pohlaví. Tedy analýza mužů a žen samostatně nebo dohromady, podle toho, co vychází lépe. Další zajímavou situací je vyloučení jedné ze tří podmínek, pokud nám ve studii nevychází signifikantně, tedy takový design výzkumu, ve kterém se nesrovnávají dvě experimentální skupiny, nýbrž tři a do článku se pak uvádějí jen vybrané skupiny, u nichž byl nalezen signifikantní efekt. V tabulce je možné vidět, že jestliže zkombinujeme všechny zmíněné triky, můžeme dosáhnout falešné pozitivity až s 61% pravděpodobností. Tedy existuje neuvěřitelná vysoká pravděpodobnost toho, že v datech, kde žádný vztah není, máme možnost naleznout signifikantní efekt. Stačí jen vyloučení dvou závislých proměnných a pravděpodobnost výskytu falešného výsledku se téměř zdvojnásobí. Jak to poté asi může vypadat, když jich někdo vyloučí 19? Protože přesně k tolika vyloučeným proměnným na základě nesignifikantních výsledků došla výzkumnice Vohs, která se k tomu veřejně přiznala, což je jen dalším důkazem toho, že se takové chování v akademické obci zřejmě nepovažuje za problém (Rohrer, Pashler, & Harris, 2015).

Výše zmíněný Daryl Bem napsal v roce 1987 metodologickou příručku pro studenty o tom, jak napsat vědecký článek. V sekci analýza dat Bem studentům přímo doporučuje v případě nálezu nesignifikaních výsledků tzv. „zkoumat problematiku ze všech úhlů“, což znamená například zanalyzovat pohlaví odděleně či přidat dodatečné kontroly. Dále pak, pokud má student pocit, že v datech vidí nějaký zajímavý efekt, má vystavět novou hypotézu a pro tu v získaných datech najít oporu. A pokud nic vysledovat nejde, Bem doporučuje hledat co nejvíce, cokoliv zajímavého (Bem, 1987). Není pak divu, že vědcům, jako je Vohs přijde redukce experimentálních podmínek na základě signifikance normální, když tomu zřejmě byli učeni na univerzitách. Mohlo by to také být vysvětlením toho, proč jsou pro některé výzkumníky replikační krize a opatření navrhovaná pro její překonání tak nepochopitelná.

Dalším možným důležitým aspektem PVP je samotné podvádění. Soudit nečestnost není tak jednoduché, nemůžeme jednoduše rozdělit lidi na ty dobré a špatné. Jde především o jejich motivaci k takové činnosti (Mazar, Amir, & Ariely, 2008). Nepoctivci jsou často lidé, s normálními hodnotami, dobrými morálními standardy s vysokým míněním o sobě, co se morálky týče. Přesto se stane, že se ocitnou v situaci, kdy jsou v pokušení podlehnout sobeckým motivům na úkor překročení hranice toho, co obvykle považují za morálně přijatelné. Většinou není složité si to nějakým způsobem obhájit. Tomuto novému morálnímu problému se říká sebestředné ospravedlnění (ang. *self-serving justifications*) (Shalvi, Gino, Barkan, & Ayal, 2015). Tento mechanismus umožňuje lidem překročit své morální hranice, aniž by ze sebe měli špatný pocit nebo by u sebe registrovali nemorální chování. Například ve výzkumu, kde je běžnou praxí, že statisticky významné výsledky jsou finančně odměňovány a publikovány v prestižních časopisech, se zdánlivě může nemorální chování obhájit. Na straně druhé statisticky nesignifikantní výsledky výzkumu jsou téměř úplně ignorovány, a to navzdory skutečnosti, že někdy z nich zjistíme více. V důsledku toho dojde ke střetu zájmů, kdy po investici významných finančních prostředků a stovek hodin práce ve výzkumu se stane, že výzkumník stojí před nesignifikantním efektem s mizivými vyhlídkami na úspěšnou publikaci (John, Loewenstein, & Prelec, 2012).

Ukazuje se tedy, že v psychologii jsou PVP vcelku běžné. John, Loewnstein a Prelec navrhli průzkum, jenž byl poslán e-mailem 5 964 vědcům z oblasti psychologie. Z toho jim odpovědělo 2155. Polovina vědců přiznala, že publikují pouze ty experimenty, které potvrzují jejich hypotézy. Skoro polovina vědců někdy odstranila taková data, která se jim nehodila, nebo publikovali nečekané zjištění, jako kdyby ho předpokládali od začátku (35%). Dvě

procenta vědců přiznali, že data někdy zfalšovali. V průměru se respondenti domnívali, že tyto praktiky byly obhajitelné (John, Loewenstein, & Prelec, 2012)

Zdá se, že nečestnost je do určité míry lidská přirozenost, jde tedy o to to přijmout a hledat způsoby jak změnit prostředí ve výzkumu, aby to bylo takové prostředí, ve kterém máme větší šanci zůstat na správné cestě.

Důležité je tedy získaná data ve výzkumu sdílet, aby mohly být výsledky výzkumu podrobeny kontrole. Wichters s kolegy (2006) ukázal ve své přehledové studii, jak vědci často nespolupracují v objasňování svých poznatků. Ze 141 autorů článků publikovaných v rámci APA, 103 z nich nereagovalo na výzvu o sdílení dat déle než 6 měsíců, přestože u každé publikace v APA časopisech se autoři zavazují ke zpřístupnění dat. V roce 2012 bylo tedy navrženo, že by výzkumníci měli zveřejňovat spolu s prací veškeré údaje související s výzkumem, zejména také data (Wichters et al, 2006).

Fakt, že je analýza dat ve výzkumu problematická, dokládá Brian Nosek et al. (2015) ve své studii. Stupně volnosti výzkumníkovi umožňují zvolit statistiku nevědomě podle špatných kritérií- Data můžou být analyzována velice různě. Noskův tým pozval přes šedesát analytiků, aby na základě dat zanalyzovali, zda fotbaloví rozhodčí dávají více červených karet hráčům tmavé pleti nebo hráčům se světlou pletí. Tuto výzvu přijalo 29 z nich. I přes to, že data byla vždy stejná, došli výzkumníci k různým závěrům. Dvacet analytiků došlo k tomu, že fotbaloví rozhodčí dali více červených karet hráčům s tmavou pletí, devět jich nenalezlo významný vztah mezi barvou pleti a červenými kartami. Volbou různých analytických postupů můžeme dospět k různým výsledkům. Studie ukazuje, že data lze opravdu analyzovat různě. Všechny postupy statistiků byly správné. Zdá se, že je tedy velice snadné vědomě či nevědomě dělat během analýzy taková rozhodnutí, která povedou k signifikantnímu výsledku. Proto by měli výzkumníci publikovat jiné možné analýzy nebo alespoň specifikovat postup analýzy ještě před jejím začátkem (Nosek et al, 2015).

Brian Nosek et al. (2015) si myslí, že výsledky jejich výzkumu by mohly pomoci vysvětlit úpadek v psychologii a dalších vědách. Existuje velká pravděpodobnost, že původního signifikantního efektu bylo docíleno právě některou PVP (Brian Nosek et al, 2015). Chování výzkumníků zřejmě není úmysl, ale spíše výsledek dvou faktorů, za prvé nejednoznačnost v tom, jaký postup či rozhodnutí ve výzkumu učinit, a za druhé touha výzkumníka najít statisticky významný výsledek (Simmons, Nelson, & Simonsohn, 2011).

Simmons, Nelson a Simonsohn (2011), doporučují použití bayesianské statistiky namísto statistiky frekvencionistické, ačkoliv i ta má svá omezení (Simmons, Nelson a Simonsohn, 2011). Frekvencionistická statistika stojí na hypoteticko-deduktivním a falzifikačním základu. Vědci vymyslí hypotézy, dedukují důsledky pro pozorování, které testují. Vědecké hypotézy mohou být odmítnuty, ale nikdy nejsou zcela prokázány. V bayesianském přístupu je důkaz o skutečném stavu světa vyjádřený bayesovskou pravděpodobností. Jednou z hlavních myšlenek je, že pravděpodobnost je řádný názor, a že závěr z údajů není nic jiného, než revizí tohoto názoru s ohledem na příslušné nové informace (Gelman & Shalizi, 2013).

Již více než 50 let je bayesianská statistika mnoha vědci viděna jako správná cesta (Edwards, Lindman, & Savage, 1963). Hlavní překážkou pro její použití byla její složitost. Moderní aplikace mohou tyto překážky již překonat, nicméně frekvenční analýzy jsou velice často v povědomí výzkumných pracovníků velmi hluboko.

PVP jednak ztrácejí čas výzkumných pracovníků a hlavně oddalují vědecký pokrok, protože vědci marně usilují o rozšíření účinků, které nejsou skutečné, a tedy nereplikovatelné. PVP tedy vyvolávají otázky o věrohodnosti výsledků výzkumu a ohrožují integritu výzkumu produkcí nereálně elegantních výsledků. To může u dalších studií vést k potřebě ještě více zkreslovat data, aby výsledky byly ještě elegantnější. Pokud by zmíněné reformy účinně snížily výskyt PVP a tedy i tlak na výzkumníky k produkci nerealisticky elegantních výsledků, vědecká integrita by se posílila.

Částečným vysvětlením pochybných výzkumných praktik je fakt, že se většinou publikují jen signifikantní výsledky. Dochází tak k fenoménu publikačního zkreslení, který je popsán v následující kapitole.

3. Publikační zkreslení

Představu o poznatcích ve vědním oborů si děláme na základě publikované literatury. Bohužel se však ukazuje, že v literatuře se v nadměrné míře objevují pozitivní, statisticky významné výsledky. Rosenthal (1979) tento fenomén nazval problémem šuplíku (ang. *file drawer problem*; termín odkazující na představu, že nepublikované studie se hromadí ve výzkumníkově šuplíku). Sacket (1979) jej ve svém seznamu zkreslení postihujících výzkum popisuje jako zkreslení pozitivními výsledky (ang. *positive results bias*). Nejvíce se však v akademické obci používá termín publikační zkreslení (ang. *publication bias*). Právě publikační zkreslení motivuje k p-hackingu popsaném výše, protože vědec přirozeně usiluje

o to, aby jeho výzkumy byly publikovány. Problém je také v samotné úvaze testování hypotéz. Jako bychom říkali: „Tady je zajímavá teorie, která by mohla vysvětlit autismus, ale vlastně se ukázalo, že tato teorie neplatí“, což není zrovna inspirativní poselství, není-li teorie tak prominentní a obecně přijímána, že by zjištění nulových výsledků bylo překvapující. A pokud tomu tak je, je pak velmi pravděpodobné, že jediná studie nebude dostatečně přesvědčivým důkazem, aby svrhla současný převažující názor. Bylo zjištěno, že problém s názvem "šuplíkový efekt" vede k deformaci odborné literatury a vytváří dojem, že pozitivní výsledky jsou mnohem významnější, než ve skutečnosti jsou (Rosenthal, 1979). Nezveřejňování negativních výsledků totiž může vést k mylné generalizaci či přeceňování významu výsledků publikovaných studií, právě tím, že velká část problematiky zůstává nepublikována.

Cohen (1962) provedl meta- analýzu studií publikovaných v prominentním časopise. Na základě průměrné velikosti vzorku, Cohen odhadnul, že průměrná statistická síla studií je kolem 60%. Zdá se tedy, že mnoho studií v psychologii má malou statistickou sílu a tedy i vysokou pravděpodobnost chyby typu II. V důsledku toho by se dalo očekávat, že by se ve spoustě článků mělo uvádět, že se ve studii nepodařilo najít podporu pro zkoumaný efekt (Cohen, 1962).

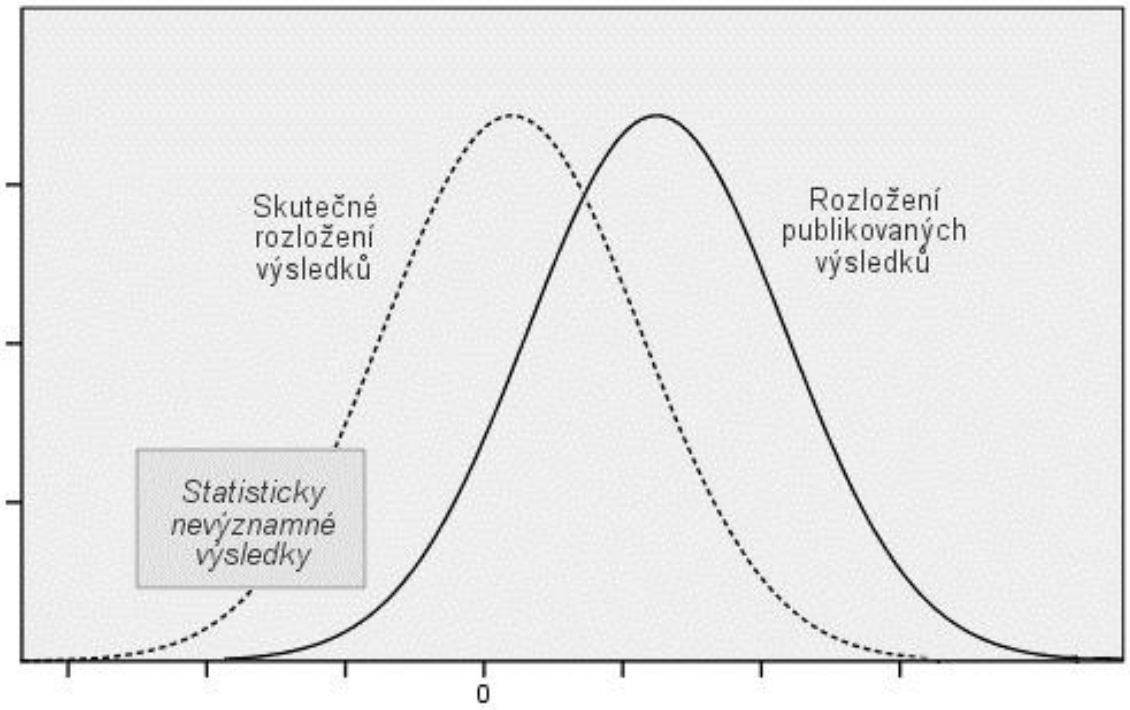
Cohen (1962) také upozornil na problém ve způsobu, jakým psychologové uvažují nad svým výzkumem. Cohen si myslí, že výzkumníci většinou ignorují fakt, že by významný výsledek mohl podlehnout chybě II druhu a testem signifikance "se chrání" pouze vůči falešně pozitivním výsledkům. Cohen si všiml, že tento druh chyby je spojen s velikostí vzorku (a velikostí efektu). Čím je vzorek (nebo efekt) větší, tím je pravděpodobnost výskytu chyby menší. Výzkumníci mohou kontrolovat pravděpodobnost výskytu této chyby cíleným plánováním velikosti vzorku. Těžko říct, kde by dnes byla psychologie, kdyby tenkrát výzkumníci uposlechli Cohena a jeho rady, jak se této chybě vyhnout. Tisíce hodin plýtvání energie a peněz na poddimenzované studie končící neprůkaznými výsledky by možná byly ušetřeny. Naopak při faktu, že malých studií můžeme udělat hned několik oproti např. jedné velké, je pravděpodobné, že se v literatuře objeví velká část falešně pozitivních. Důsledkem pak bývá, že výsledky není možné replikovat. Současná replikační krize, o níž je pojednáno v dalších kapitolách, by Cohena zřejmě vůbec nepřekvapila (Cohen, 1962).

Nicméně v roce 1959 Ted Sterling podal přehledovou zprávu o publikační činnosti čtyř hlavních psychologických žurnálů, a zjistil, že až 97% článků bylo založeno na výsledcích podporující zkoumanou hypotézu (Sterling, 1959) – po „neúspěšných“ studiích, jichž by mělo být téměř 40%, tak nebylo ani stopy. Bylo prokázáno, že statisticky významné výsledky mají větší pravděpodobnost, že budou publikovány ve srovnání s články, které předkládají výsledky negativní (Dickersin, Chan, Chalmersx, Sacks, & Smith, 1987). Přestože studie, jež neprokáží efekt, jsou samozřejmě často stejně tak kvalitní jako ty, které efekt prokáží (Easterbrook, 1991). O třicet šest let později Sterling znovu mapoval literaturu a došel ke stejnému problému - negativní výsledky byly stále cenzurovány (Sterling, Rosenbaum, & Weinkam, 1995). Patnáct let po tom tuto hypotézu Daniele Fanelli z University of Edinburgh potvrdil ještě jednou (Fanelli & Scalas, 2010).

Existují dvě hlavní vysvětlení rozporu mezi „úspěšností“ publikovaných psychologických studií a nedostatečností statistické síly. Jedním z možných vysvětlení je, že výzkumní pracovníci provedou několik studií a publikují pouze úspěšné pokusy a nesignifikantní studie zůstávají v „šuplíku“ (Rosenthal, 1979). Druhým vysvětlením je, že výzkumníci se k signifikantním výsledkům dostanou díky PVP a výsledky výzkumů jsou tedy falešně pozitivní (John, Loewenstein, & Prelec, 2012). Oba postupy mají samozřejmě nepříznivý vliv na věrohodnost a reprodukovatelnost výsledků publikovaných v psychologických časopisech. Dále John Ioannidis píše o silné konkurenci v získávání finančních prostředků a silného tlaku pro získávání významných výsledků (Ioannidis, 2005).

Jednoduchým řešením tohoto rozporu by bylo zvýšení statistické síly studie. Nicméně meta-analýzy a sekce metod v publikovaných článcích opakovaně poukázaly na to, že psychologové neberou v úvahu statistickou sílu při plánování svých studií, a že jsou studie nadále poddimenzované (Schimmack, 2012)

Dalším spoluvinníkem nereprezentativní situace ve vědě, potažmo publikačního zkreslení, je zřejmě akademická incentivní struktura, která se řídí pravidlem “publish or perish“ – publikuj nebo zahyň, které podporuje publikaci signifikantních efektů. V některých vědeckých institucích, může být "produktivita" práce nastavena třeba tak, že výzkumník musí publikovat tři články ročně, včetně jednoho v prestižním časopise s impakt faktorem nejméně pět (Colquhoun, 2011). Vzhledem k tomu, že pozitivní výsledky mají větší pravděpodobnost být publikovány a akademický pracovník povětšinou publikovat musí, výzkum je pak zřejmě s větší pravděpodobností vystaven PVP.



Graf II. *S laskavým dovolením převzato od Ondřeje Nováka (2016).*

Mezi nejvýznamnější zdroje publikačního zkreslení patří selektivní poskytování článků k publikaci autory. Někdy dochází k částečnému publikování výsledků, kdy autoři do svého článku úmyslně některé výsledky nezahrnou, samozřejmě se většinou jedná o nesignifikantní výsledky. Takové chování je některými výzkumníky pokládáno za neetické a je také zařazováno do PVP, zvláště v případech klinických studií, kde například neposkytnutí dat nepodporujících určitou léčbu zkresluje vnímání její účinnosti (Antes & Chalmers, 2003; Chalmers, 1990).

Protože časopisy většinou negativní výsledky nebo přímé replikace nepublikují, výzkumníci se často dopouštějí PVP. Může se také zdát, že falešně pozitivní výsledky plýtvají zdroji, například když proběhne investice do neplodného výzkumného programu. Navíc oblast, která bude publikovat nepřítomnost zkoumaného efektu, riskuje, že obecně ztratí svou důvěryhodnost (Simmons, Nelson, & Simonsohn, 2011)

Přítomnost publikačního zkreslení byla rovněž zkoumána i v meta-analýzách. Největší studie, jež byla doposud provedena, pojednává o hodnocení léčebných postupů z Cochrane Library. Studie ukázala, že pozitivní statisticky významné nálezy mají o 27% větší pravděpodobnost, že budou zahrnuty v meta-analýzách. Další zajímavostí je, že pokud výsledky dokazují nulové nepříznivé účinky, mají o 78% vyšší pravděpodobnost, že budou

zahrnuty do meta-analýzy, než když statisticky významné výsledky ukazují, že nepříznivými účinky léky disponují (Kicinski, Springate, & Kontopantelis, 2015).

Zkreslením podléhá i samotná meta-analýza, v posledních letech se zkoumaly její limity a užitečnost. Řada vědců se na toto téma specializovala a navázala spolupráci se statistiky na zdokonalení výpočetních postupů. Současně se ukázalo, že meta-analýza má mnoho omezení, na která si meta-analytik musí dát pozor (Flather, Farkouh, Pogue & Yusuf, 1997). Výzkumník si musí příslušné studie vyhledat, což může být jednak velmi náročné, ale také pochybné. Problematické také bývá stanovení si kritérií, podle kterých bude různé studie zahrnovat do meta-analýzy. Následkem toho je, že výzkumník vybírá z toho, co je již publikováno, což je vzhledem k publikačnímu zkreslení nereprezentativní.

Ačkoliv v minulosti byly vytvořeny analytické techniky, které měly za úkol v meta-analýze toto zkreslení odstranit, výsledky jejich aplikace nebyly zcela uspokojivé. Důvodem byla především skutečnost, že nikdo přesně neví, jak velké publikační zkreslení je a tudíž ani jak velká korekce je potřebná. Objevila se však nová metoda, která v meta-analýzách s publikačním zkreslením počítá. Jmenuje se p-křivka (p-curve metoda), protože je založena na analýze distribuce statisticky významných p-hodnot pro sadu studií. P-křivka vychází ze skutečnosti, že p-hodnoty mají v případě platnosti nulové hypotézy uniformní rozdělení. To znamená, že každá z možných p-hodnot je stejně pravděpodobná. Pokud křivka vytvořená z p-hodnot zkoumaných studií neodpovídá uniformnímu rozdělení, můžeme dle tvaru zešikmení usuzovat o velikosti efektu, popřípadě o použití PVP v provedených studiích (především p-hacking a šuplíkový efekt) (Simonsohn, Nelson, & Simmons, 2014). Nicméně jde o velice novou metodu, která zatím není dostatečně prozkoumaná. Nejlepším a nejspolehlivějším řešením by tedy stejně bylo publikační zkreslení odstranit.

Dalším zajímavým nápadem je web *Curate Science*, který organizuje a sumarizuje výsledky nezávislých přímých replikací meta-analyticky, takže výzkumníci mohou vidět, jak si jejich výsledky stojí oproti ostatním replikacím a celkově kalibrují svou víru v dané empirické poznatky podle nezávislých důkazů. Výstup meta-analýzy je formou sumy replikací – je zde jednoduše znázorněn tzv. forest plotem, neboli sérií intervalů spolehlivosti replikovaných studií vedle intervalu spolehlivost studie původní. Dále jsou hned vedle dostupné veškeré články a data (“Curate Science”, 2016).

Dalším problémem je také fakt, že kvalita časopisu je měřena počtem citací v ostatních článcích za dané období, což publikační zkreslení jen zvyšuje, protože články s pozitivními

výsledky budou citovány spíše než ty s výsledky negativními (Boor, 1982). Citační impakt navíc znevýhodňuje časopisy, které vydávají články, jež se příliš neuplatňují nebo jsou citovány jen zřídka. Thorne (1977) uvádí, že některé články, například Freuda a dalších časných psychoanalytiků, byly citovány jen zřídka během let bezprostředně po jejich vydání, ale ve velké míře byly citovány během následujících desetiletí. Kdyby byly hodnoceny podle citačních indexů počítaných během deseti let po jejich zveřejnění, kvalita časopisů by se snížila. Thorne (1977) došel k závěru, že posouzení vědecké zásluhy nebo kvality časopisu založený na citačních indexech je jakousi pseudovědeckou popularitou. Dalším argumentem proti citačnímu indexu je fakt, že se nehodnotí, v jakém kontextu je studie citována. Příkladem může být již zmíněná studie Bema (2011), která je často v publikacích propírána za špatnou metodologii. Roku 2013 se Akademie věd ČR připojila k Sanfranciské deklaraci (z roku 2012), která impakt faktor, který měl původně sloužit ke knihovnickému hodnocení časopisů, nedoporučuje používat na hodnocení výzkumu (Akademie věd ČR, 2013).

Problematickým faktorem je také to, že vědu dělají lidé. Často se nemohou smířit s negativním výsledkem. Zapomínají na to, že věda je především o hledání pravdy a vnášejí do ní svoje ambice a přání a mnohdy zapomínají na to, že něco skutečně prokázat je velice náročná a někdy také dlouhá cesta, plná slepých uliček, jež většinou pochopitelně vnímají jako neúspěch, namísto toho aby je spatřovali jako cenné poznání. Jak uvedl Schwartz ve své eseji (2008), výzkumníci se často setkávají s pocitem vlastní neschopnosti a selhání. Zkrátka nejsou vychováváni k tomu, aby pochopili, jak těžké je udělat výzkum. Co ho dělá obtížným je ponoření se do neznáma. Výzkumník si nikdy nemůže být úplně jist, tím to dělá, zda se ptá na správnou otázku, a následně na ni designuje správný experiment, dokud nedostaneme výsledek. Schwarz píše o tzv. „produktivní hlouposti“, být hloupým z vlastní vůle - protože zaměření na důležité otázky staví výzkumníka velice často do nepříjemné situace neúspěchu, a je tedy potřeba s tím v této kariéře počítat. Jedině tím, že výzkumníka nezraní fakt, že jeho hypotéza se nepotvrdila, a bude zvídavě v té samé oblasti bádat dál, se dostane hlouběji do neznáma a pravděpodobně časem něco objeví. Je tedy důležité si tohoto aspektu všímat a ptát se, jak mohou negativní emoce spojené s možným neúspěchem potencionálně ovlivnit výzkum a nejlépe si tuto přirozenou hloupost přiznat a využít jí k hlubší znalosti (Schwartz, 2008).

Někteří vědci, mající pochybnosti o výsledcích psychologického výzkumu, jakoby si této hlouposti byli vědomi. Rozhodli se tedy výsledky psychologického výzkumu podrobit

kontrole jedním ze základních nástrojů vědecké metody – replikací, o které je pojednáno v následující kapitole.

4. Replikace

Replikace je termín označující opakování výzkumné studie, obecně v jiných, ale podobných podmínkách, aby se zjistilo, zda základní poznatky zjištěné v původní studii lze zobecnit i pro ostatní účastníky a situace. Statisticky významné výsledky ve studii jsou většinou úspěchem. Když ale tytéž postupy a metody mohou být replikovány s různými účastníky se stejnými nebo podobnými výsledky, výsledek studie pak vykazuje větší spolehlivost zjištění. Znamená to, že je více pravděpodobné, že tyto výsledky lze zobecnit na širší populaci (Wicherts, Borsboom, Kats, & Molenaar, 2006).

Termín reprodukovatelného výzkumu se vztahuje k myšlence, že konečný produkt akademického výzkumu je článek, který obsahuje veškerá data, výpočty a postupy, jež lze použít jednak k reprodukci výsledků, tak pro inspiraci k výzkumu novému (Fomel et al, 2009). Trvalo téměř dvacet let, než vědci začali brát Cohenovy rady ohledně velikosti vzorku vážně (Gomes & McCullough, 2015). Poprvé v historii disciplíny, začali psychologové spoléhat na replikace a řešit statistickou sílu studie (Klein et al, 2014).

Pokud zjistíme nějaký efekt pomocí výzkumu, měl by ho být schopný zopakovat kterýkoli výzkumník, který má k dispozici originální studii (Simons, 2014). Podle Schoolera (2014) je přímá replikace několika laboratořemi rozmístěnými po celém světě možností, jak zredukovat chyby, kterých se jako výzkumníci dopouštíme, a dobrat se tak spolehlivějších výsledků. Vzhledem k tomu, že reprodukovatelnost experimentů je nezbytnou součástí vědecké metody, má potenciálně velmi vážné důsledky pro mnoho vědních oborů, v nichž jsou významné teorie nereplikovatelné.

Lykken (1968) navrhl tři druhy replikace - přesnou (literal), funkční (operational) a konceptuální (constructive). Schmidt (2009) vyřadil doslovnou, protože v podstatě vyžaduje původní výzkumníky, a redukoval druhy replikací na přímou a konceptuální. V přímé replikaci nový výzkumný tým v podstatě pouze používá experimentální postupy původního výzkumu podle metodické části z původního článku s novou skupinou účastníků. V konceptuální replikaci nejsou původní metody kopírovány, ale záměrně pozměněny, aby jiným způsobem testovaly hypotézu. Zatímco přímé replikace zkoumají spolehlivost původních výsledků, koncepční replikace testují existenci samotného konstruktu (Schmidt, 2009).

Nabízí se ovšem otázka, zdali je úspěšná replikace opravdu měřítkem prokázání efektu. Názory na tuto problematiku se totiž v psychologii dost liší. Existuje mnoho argumentů proti replikacím v psychologii, například, že v psychologii není replikace tak jednoduchá jako například v biologii, kde se stačí podívat vícekrát na různé vzorky pod mikroskopem a efekt je oproti psychologii evidentní a hmatatelný. Psychologické výzkumy jsou často podhodnoceny, co se velikosti vzorku týče, a experiment může ovlivňovat více faktorů, podle některých autorů například i maličkosti jako je den v týdnu či osvětlení. Někdy také dochází k tomu, že jsou experimentální studie navrženy způsobem, který prakticky zaručuje pozitivní výsledky. Jakmile jsou tyto výsledky publikovány a několik výzkumníků experiment zopakuje přesně tak, jak byl publikován, tak pokud byl původní experimentální výsledek dosažen pochybnými výzkumnými metodami, je pravděpodobné, že replikace vyjde, přestože je výsledek falešně pozitivní. Dalším způsobem, jak prověřit hypotézu o efektu, je provést namísto toho replikaci konceptuální, která testuje podobnou hypotézu pomocí různých metod (Yong, 2012).

Pozitivní výsledky se v psychologii někdy chovají jako zvěsti. Snadno se šíří, ale obtížně rozptylují. Pozitivní výsledky ovládají většinu časopisů. Výsledek se pak nezreplikuje, nicméně jeho publikace už není tak jednoduchá, jak možno vidět na příkladu replikace Stéphana Doyena (2012).

Nepodařilo se mu zreplikovat klasický experiment od Johna Bargha (1996), který ukazoval, že lidé chodí pomaleji, pokud byli nevědomě naprimováni deseti slovy, které souvisejí s vyšším věkem. Po několika odmítnutích vydání tohoto výzkumu byla studie vydána v *PLoS ONE* (Doyen et al, 2012). Tato kauza celkově poukázala na pochybnosti o existenci robustních behaviorálních efektu priminigu. Daniel Kahneman například vyzýval výzkumníky primingu, aby se podívali na robustnost zjištěného efektu a otevřeně pochyboval o tomto konstruktu, jehož považuje za *dítko pochybnosti důvěryhodnosti psychologického výzkumu* (Kahneman, 2012).

Jak je možné, že tak zavedený koncept jako je priming nebyl zreplikován? Původní studie je ke dni 19. 3. 2016 citována podle Google Scholar 3529 krát. Bargh v ní navrhnul úkol, ve kterém měli účastníci studie vytvořit větu z různých slov. Po splnění úkolu měli účastníci odejít směrem k výtahu. Experimentátor tajně stopkami měřil čas, který uběhl, než se dostali k němu. Bargh měl hypotézu, že lidé, co řešili slovní úlohy obsahující slova související se stářím, by měli mít v mysli aktivován koncept stáří. V této studii by to

znamenalo, že když by věta v experimentálním úkolu obsahovala slovo, jež souviselo se stářím, účastníci by opouštěli laboratoř pomaleji, než účastníci studie, kteří tato slova neměli. Bargh se samozřejmě přesvědčil o tom, že účastníci experimentu neprohlídli hypotézu a že si sice ničeho zvláštního na slovech nevšimli. Nicméně jejich chování se změnilo. Na vzorku 34 studentů tedy zjistil signifikantní rozdíl času mezi skupinami a usoudil tedy, že chování člověka může být nevědomě naprimováno (Bargh, 1996).

Stephane Doyen s kolegy se pokusili Barghovu studii zopakovat. Snažili se, aby odpovídala studii původní, ale vylepšili metodologii – místo stopek pro přesnost použili infračervený senzor, jenž přesně změřil čas, za nějž účastníci prošli sledovaný úsek, zdvojnásobili počet účastníků a rekrutovali čtyři experimentátory, kteří provedli studii, ale neznali zkoumané hypotézy.

Tentokrát slova neměla žádný vliv na rychlost chůze účastníků. Doyen měl podezření, že výsledky v původní studii byly způsobeny pouze díky efektu experimentátora (Rosenthal, 1966). Doyen tedy zopakoval svůj pokus s 50 novými účastníky výzkumu a 10 novými experimentátory. Tentokrát experimentátoři opět měřili stopkami. Doyen řekl polovině z nich, že někteří účastníci budou chodit pomaleji na základě primingu a ostatní budou mít chůzi rychlejší (Doyen, 2012).

Zjistil, že účastníci studie se pohybovali pomaleji pouze tehdy, když byli testováni experimentátory, kteří věděli, v jaké experimentální skupině se účastníci nacházejí a na základě toho ovládali stopky. Doyen tedy dokázal zreplikovat Barghův experiment jedině tehdy, řekl-li experimentátorům, co mohou očekávat. V Barghově studii, experimentátor dostal obálky se dvěma různými slovními úlohami (buď slova související se stářím, nebo slova neutrální). S příchodem každého dobrovolníka experimentátor náhodně vybral obálku, odvedl ho do zkušební místnosti, dal mu instrukce k úkolu a nechal ho pracovat na úkolu. V průběhu doby trvání studie experimentátor mohl vidět, který úkol účastník dostal, a podle toho mohl změnit své chování. Doyen si nemyslí, že by to byl úmyslná manipulace. Napsal: *„Tato možnost byla v podstatě neformálně potvrzena v naší studii, protože jsme zjistili, že bylo velmi snadné neúmyslně zjistit skupinu, ve které účastník byl, stačil jen jeden pohled na materiál“*(Doyen, 2012).

Bargh se na svém blogu velice bránil, že by experimentátoři mohli vědět, v jaké skupině se účastníci nacházeli. Šel i do osobního útoku a popsal Doyen a jeho tým jako "neexpertní výzkumníky". Později vzal tato slova zpět s tím, že reagoval tak silně, částečně

proto, že viděl rostoucí skepsi vůči nevědomým myšlenkovým procesům, které jsou dle něj důležité (Yong, 2012).

Velká část zmatků v tomto sporu je kvůli podrobnostem v použitých metodách – Co přesně Bargh et al. provedli v původní studii, a jak úzce Doyen et al. reprodukovali postup? Bargh později napsal článek o metodách, v němž popsal další podrobnosti o použité metodice ve výzkumu, kterou prý Doyen et al. nenásledovali. V době publikování původní studie nebyly dostatečně rozepsány, zřejmě z nedostatku místa, nicméně v dnešní době neomezených internetových úložišť, neexistuje žádný důvod, proč by veškerá data a metodiky studií nemohly být publikovány.

Nicméně bychom měli být opatrní v dělání závěrů o „úspěšné či neúspěšné replikaci". O tom, co je pokládáno za úspěšnou replikaci by se dalo dlouze diskutovat. Existuje teorie, že pokud je soud založen na základě rozdílu $p < .05$ a $p > .05$, nemůžeme jen kvůli tomu, že původní studie může zamítnout nulovou hypotézu a replikace ne, říct, že se replikace od původní studie signifikantně liší. Pokud chceme říct, že replikace byla neúspěšná, musíme provést test rozdílu mezi testy použitých v obou studiích. Vypadá to tak, že u každé studie vypočítáme velikost účinku a pak porovnáme studie pomocí Z-testu signifikance rozdílnosti mezi dvěma korelačními koeficienty. Rozdíl korelačních koeficientů byl v Barghově kauze signifikantní (Srivastava, 2012). Problematické však na tom je, že když je vzorek v původní studii malý, je odhad efektu velmi nepřesný. Níže zmíněný Dan Gilbert úspěšnou replikaci vidí tak, že výsledky replikace se nachází v intervalu spolehlivosti studie původní. To ale může být problematické, například když má původní studie široký interval spolehlivosti, tak je poté v podstatě pokaždé replikovatelná. Autoři projektu *Reproducibility Project* přizvávají, že kritérium úspěšné replikace není jasné, sami ho nastavili tak, že není-li signifikantní efekt v tom směru jako studie původní, pak replikace není úspěšná (Open Science Collaboration, 2015).

Opětovné zkoumání tohoto konstruktu a ukázání slabin výzkumu nás nutí se ptát, jak jsou na tom ostatní „prokázané" efekty. Dalším příkladem teoreticky ukotveného psychologického efektu, na nějž se pohled díky replikacím poněkud změnil, je verbal overshadowing effect (Schooler & Engstler-Schooler, 1990). Jedná se o negativní vliv slovního popisu při vizuálním rozpoznávání, jehož aplikace v oblasti policejních vyšetřovacích postupů by byla zásadní. Pozdější studie, které se snažily efekt znovu prozkoumat, se ve velikosti efektu lišily (Schooler, 2011; Lehrer, 2010). Ukázalo se, že efekt

je sice replikovatelný, nicméně se hodně diskutuje o velikosti efektu. Protože stále není jasné, zda slovní hodnocení zastiňuje jakousi část v rozhodovacím procesu nebo jde o sníženou rozlišitelnost paměti. Navíc se ukazuje, že výsledky jsou zřejmě silně ovlivněny načasováním úkolů v experimentu, a jeho důsledky jsou tedy poněkud nadhodnoceny a je zde mnoho prostoru pro další výzkum (Alonga et al, 2014).

První masivní replikační projekt v psychologii, Studie *Many Labs Replication, Project* proběhl v roce 2014. Velká skupina výzkumníků zreplikovala 15 studií z psychologické literatury v různých laboratořích po celém světě. Výzkum je společně s daty zveřejněn v rámci projektu *Open Science Framework*. Z těchto 15 efektů se jich podařilo zreplikovat 13 a 2 nikoliv. U některých experimentálních replikací byly nalezeny větší efekty než v původních studiích. Studie se mohly pochlubit vysokým počtem účastníků, který přesáhl počet 6000 – na větším vzorku se dá lépe ověřit existence efektu. Replikace tedy není synonymem pro ukazatel špatných výzkumů, jak by se z nepovedených kauz mohlo zdát, naopak může díky velkému vzorku a pohledů mnoha špičkových výzkumníků původní studii podpořit (Klein et al, 2014). Problémem může být, že si výzkumnici výsledek replikace mohou vzít osobně, jako již zmíněný Bargh jehož blog byl plný útočných připomínek.

Další replikační projekt s názvem *Reproducibility Project* proběhl v roce 2015. Projektu se zúčastnilo 270 vědců z celého světa, kteří zreplikovali 100 korelačních či experimentálních empirických studií ze tří špičkových psychologických časopisů - *Psychological Science, Journal of Personality and Social Psychology,* a *Journal of Experimental Psychology: Learning, Memory, and Cognition,* publikovaných v roce 2008. Výzkumné projekty a strategie analýz v původních výzkumech se samozřejmě lišily. Na základě konzultací s původními autory, získávání původních materiálů a vnitřního přezkumu byli v replikacích použity původní podmínky. Všechny analýzy pak byly převedeny na společnou velikost účinku, srovnávací koeficient a intervaly spolehlivosti. Kritériem pro závěry o reprodukovatelnost výzkumu byly velikosti efektů originální a replikované studie. Výsledný otevřený datový soubor poskytl první odhad reprodukovatelnosti psychologie a korelaci dat pro podporu hypotézy o důležitosti reprodukovatelnosti (Open Science Collaboration, 2015). Průměrné efekty v replikaci ($r = 0{,}197$, $SD = 0{,}257$) byly poloviční velikosti originálních efektů ($r = 0{,}403$, $SD = 0{,}188$), což představuje podstatný pokles. Z vybraných původních studií mělo 97 významné výsledky ($p < .05$). V této replikaci těchto výsledků dosáhlo 36%. (Open Science Collaboration, 2015). Výsledky studie byly rozhodně překvapivé. Nicméně to neznamená, že nezreplikované efekty neexistují, vzhledem k tomu, že

se autoři studie snažili o statistickou sílu 0,9. To znamená, že i za ideálních podmínek by se dalo čekat, že 10 efektů ze 100 se nezreplikuje.

Proč je tedy replikace v psychologii tak obtížná?

Autoři projektu tvrdí, že existují tři hlavní důvody, proč by původní závěry studie nemohly být úspěšně replikovány. Buď je to tím, že původní výsledky byly falešně pozitivní, nebo jsou replikované výsledky falešně negativní. Také je zde možnost, že obě studie byly správné, ale lišily se kvůli neznámým rozdílům v experimentálních podmínkách či metodik. Největší prediktorem úspěchu replikace byla velikost původních efektů. (Open Science Collaboration, 2015).

Zdá se, že problém malé reprodukovatelnosti v psychologii má na svědomí také zneužívání p-hodnoty. Hodnoty p se běžně používají k testování nulové hypotézy, která obecně uvádí, že neexistuje žádný rozdíl mezi dvěma skupinami nebo že neexistuje žádná korelace mezi dvojicí charakteristik. Čím menší p-hodnota je, tím méně je pravděpodobné, že by k pozorovanému souboru hodnot došlo náhodou, za předpokladu, že nulová hypotéza je pravdivou. P-hodnota 0,05 nebo méně obecně vzato znamená, že nález je statisticky významný a zaručuje publikování článku. Ale to nemusí být nutně pravdou. P-hodnota 0,05, neznamená, že existuje 95% šance, že daná hypotéza je správná. Znamená to, že v případě, že je nulová hypotéza pravdivá, a všechny ostatní předpoklady jsou platné, je 5% šance na získání výsledku alespoň tak extrémního jako ten pozorovaný. P-hodnota nemůže ukázat důležitost efektu: například, léčivo může mít statisticky významný vliv na hladinu glukózy v krvi pacientů, aniž by mělo skutečný terapeutický účinek. Nedorozumění o tom, jaké informace obsahuje p-hodnota, vznikají často již v učebnicích statistiky (Baker, 2016).

Někteří vědci jsou proti pesimistickým závěrům o reprodukovatelnosti v psychologické vědě, jak mnozí učinili na základě výsledků replikačních projektů. Výsledky replikační studie re-analyzoval Daniel Gilbert a jeho tým, který prohlásil, že je z mnoha důvodů chybná. Argumentuje rozličnými podmínkami, jako například, že jsou prováděny v jiných zařízeních, v různém počasí, s různými experimentátory, s různými počítači a v různých jazycích. Podle Daniela Gilberta se studie často od originálu lišily značně, a to nejen zdánlivě zanedbatelnými aspekty, ale i závažnějšími rozdíly, například, že se efekty testovaly na jiných populacích a tím se závažně měnily podmínky studie a byly tedy pak metodologicky odlišné od studií původních. Například jedna z původních studií byla v Izraeli, kde se bavili o

tom, že mají vojáci odejít na povinnou službu, a když to replikovali v Americe, tak byl odchod přerámován na líbánky, protože něco jako je povinná vojenská služba zkrátka v Americe není. Dále Gilbert říká, že do výběru zkoumaných 100 studií byly záměrně vybrány ty s malou statistickou sílou. Podle Gilberta tedy autoři OSF projektu podléhají jakémusi zkreslení a dělají velké závěry o psychologické vědě na základě jedné replikační studie, kde bylo prozkoumáno pouze 100 vybraných efektů (Gilbert, King, Pettigrew, & Wilson, 2016). Hlavním výmluvným argumentem proti Gilbertovi od Joachima Vandekercha je: *„Nezajímá mě, že byl efekt vyzkoumán pouze ve vaší laboratoři, když byla teplota přesně 69 stupňů, bylo pravé poledne, a pátá středa v měsíci. Takové zjištění není zajímavé. Zajímavé je vysvětlení principů lidského chování od jeho základů.“* (Palmer, 2016).

Dalším důležitým faktem je, že u těchto velkých replikačních studií jsou postupy rozebírány s původními autory. Jestliže se replikační projekty nepotkají se spoluprací původních autorů, pak se do ní raději nepouštějí, zkrátka proto, že nemají dostatečně mnoho informací k tomu replikaci uskutečnit. Příklad kritické připomínky Gilberta – podmínky vojenské služby a líbánek, byl například prodiskutován, a i když se zdánlivě jedná o velice odlišné podmínky, v designu studie to bylo vedlejší, protože šlo jen o to zdůvodnit v experimentu krátkodobou nepřítomnost v práci (Nosek et al, 2015). Navíc v těchto replikačních studiích, vždy při změně podmínky udělají pretest a ukázalo se, že lidé vnímají obě podmínky stejně.

Gilbertova kritika se mi tedy zdá poněkud tendenční, protože ukazuje na aspekty, které při prvním pohledu můžeme považovat za problémy, nicméně pak, když se člověk podívá na postupy replikace, zjistí, že jsou tyto aspekty ve skutečnosti ošetřeny.

Další podle mě oprávněnou kritikou replikace je také fakt její časové a finanční náročnosti. Dobrým řešením se zdá provádět replikace v rámci výuky metodologie. Student by měl pak možnost projít si celým procesem experimentu a výsledek by byl pak skutečně využit v praxi. To se většinou u školních experimentů, vzhledem k málo znalostem a velmi omezenému času na práci většině studentů nepovede, což může vést spíše k jejich demotivaci. Autoři článku si také od tohoto postupu slibují jakousi výchovu mladých výzkumníků, které budou tímto způsobem od začátku učit hodnotám jako je otevřenost a poctivost (Frank & Saxe, 2012).

Zatímco někteří by mohli na základě výsledků replikačních projektů být v pokušení dívat se na psychologii jako pavědu, tyto nálezy ve skutečnosti pomáhají psychologické vědě

stát se silnější. Je třeba si uvědomit, že lidské myšlení a chování je mimořádně jemný a stále měnící se předmět studia, takže změny lze očekávat zvláště pak například při pozorování různých národností. Některé výsledky výzkumu mohou být špatné, ale kopat hlouběji a poukázat na jejich nedostatky a navrhovat lepší experimenty je určitě dobrou cestou. Cílem replikací pak rozhodně není negativně poukazovat na původní studie, ale podívat se na to, které z nich jsou prostě spolehlivější. Jestliže u studie vidíme, že je jde bez problémů zreplikovat, pak jejímu efektu můžeme více důvěřovat.

Stěžejním mezníkem v diskuzi o přínosech replikací, byla publikace zářijového čísla *American Psychologist* (stěžejní publikace American Psychological Association) v roce 2015. Toto vydání bylo věnováno replikační krizi v psychologii. Články se pokoušeli identifikovat faktory, které znehodnocují psychologický výzkum (Ferguson, 2015; Maxwell, Lau, & Howard, 2015). Nicméně problémy v psychologii nejsou žádnou novinkou. Již v sedmdesátých letech Cronbach, prominentní metodik, poznamenal obecnou tendenci efektů v psychologii rozkládat se v čase, neboli vlastnost efektů ztrácet signifikanci s časem (Cronbach, 1975). Později Jonathan Schooler získal značnou pozornost článkem v časopise *Nature*, ve kterém popsal všudypřítomnost problému replikace v psychologii, a nastínil některé faktory, které jsou za to potenciálně odpovědné (Schooler, 2011).

Co si vzít z toho, když replikace nevychází? Znamená to pak, že efekt neexistuje? Rozhodně tomu tak být nemusí. Nepovedená replikace pouze znamená, že jeden konkrétní experimentální design zkoumající daný fenomén nemůže efekt spolehlivě vysledovat. Zvolené úkoly v konkrétním experimentu mohou být velice specifické, případně špatně zvolené. Každopádně problémem není nezreplikovaný experiment, ale spíše důvěryhodnost publikací. Jestliže existuje tolik nepovedených replikací, co to vypovídá o prestižních žurnálech, které publikovaly prvotní výzkumy? Jak můžeme důvěřovat tomu, že článek, který zrovna čteme v odborném časopise, můžeme použít jako základ pro svůj vlastní výzkum, nebo jeho obsah použít přímo v praxi? Nepovedené replikace nejsou problémem, nýbrž jen symptomem. Cestou k řešení problému se zdá být předregistrace výzkumů.

5. Předregistrace

Zdali jsou pochybné výzkumné praktiky vědomé či nevědomé, je jedna otázka, důležitější otázkou však je, jak se s nimi vyrovnat? Zdá se, že nejlepší by bylo jim předejít, a to je možné pomocí předregistrace.

Myšlenka předběžné registrace studie je jednoduchá: autoři, kteří se rozhodnou pro výzkum, uvedou své hypotézy, metody a plány analýzy dříve, než začnou data skutečně sbírat (Cumming, 2014). Zamezí se tak dodatečnému lovení souvislostí v datech a tomu, že by někdo své výsledky nepublikoval.

V předregistraci je důležitá hranice mezi konfirmační analýzou dat a dodatečnou explorativní analýzou. Přestože je studie se všemi postupy předregistrovaná a hypotéza nevyjde, může výzkumník ve svém článku kromě předregistrovaného konfirmačního výsledku uvést, že provedl explorační analýzu proměnných a objevil možný vztah mezi jinými proměnnými. Ty může zkoumat v další studii, jež bude na této nové hypotéze postavena. Tento model použili například Bahník et al. (2015) nebo Lurquin et al. (2016) při studiu fenoménu ego deplece, kdy zkoumali existenci tohoto fenoménu. Efekt nevyšel signifikantně, výzkumníci ale dodávají, že na základě další analýzy zjistili, že by některé individuálními charakteristiky jedinců mohly souviset s maximalizací výskytu efektu ego deplece. Další možností je také cíleně rozdělit dataset na polovinu a v jedné půlce volně hledat vztahy a nalezené vztahy pak konfirmovat na půlce druhé, a tím je podpořit či vyvrátit (Lurquin et al., 2016).

Registrace může pomoci v případech, kdy dochází ke střetu zájmů, nebo se výzkumníci mohou z nějakého důvodu zdráhat nulové výsledky publikovat. V roce 2007 Cylharova et al. například publikovali studii týkající se stavu hladiny mastných kyselin ve vztahu s dyslexií u dospělých. Tato výzkumná skupina se zdá v tomto tématu zaujatá, protože hlavní autor článku již dříve napsal publikaci na toto téma. Výzkumníci tvrdili, že poměr omega 3 a omega 6 mastných kyselin se mezi dyslektiky a kontrolní skupinou lišila, zároveň dospěli k závěru, že pokud chtějí jasně prokázat účinky omega 3 mastné kyseliny, musí výsledky studie potvrdit krevní biochemickou analýzou před a po doplnění této kyseliny. V diskuzi pak uvádí, že tento výzkum již probíhá. Ani čtyři roky po vydání tohoto článku, nebyly publikovány žádné výsledky a žádosti o informace byly ignorovány. Kdyby byl tento výzkumný projekt registrován, autoři by museli oznámit výsledky, nebo by alespoň museli zdůvodnit, proč tak učinit nemohou.

Například medicína vzala problém šuplíkového efektu na vědomí a je běžnou praxí, že se klinické studie předregistrují a časopisy se zavazují, že budou zveřejňovat výsledky metodologicky správných studií bez ohledu na výsledek. V posledních letech byly publikovány dvě rané intervenční studie s nulovými výsledky, o autismu (Green et al, 2010) a o pozdním vývoji řeči (Wake et al, 2011). Samozřejmě není uspokojující, že tolik úsilí

nevedlo k revolučnímu zjištění, ale právě tato znalost zabraňuje falešným nadějím a plýtvání finančními i lidskými zdroji na věci, které nejsou účinné. Přesto je nepravděpodobné, že by takovéto studie našly místo ve vysoce impaktovaném časopise.

Projekt OSF (*Open Science Framework*) se snaží všechny zmíněné problémy řešit či spíše jim předcházet. OSF pracuje pod *The Center for Open Science*, což je nezisková organizace působící především prostřednictvím OSF webové aplikace, jejímž cílem je otevřenost, integrita a reprodukovatelnost vědeckého výzkumu. Projekt podporuje efektivní a otevřený výzkumný pracovní postup. Výzkumníci pomocí OSF spolupracují, sdílejí data a hlavně registrují výzkumné projekty, kde se ještě před tím než dojde ke sběru dat, mohou dočkat cenné zpětné vazby. Webové prostředí výzkumníkům ulehčuje a zpřehledňuje práci, nabízí svým uživatelům také praktická datová úložiště (Center for Open Science, 2015).

Tento projekt podporuje myšlenku předregistrace jako otevřeného výzkumného postupu, který některé psychologické žurnály implementují do svých běžných postupů. Výzkumníci mají buď možnost či jsou povinni předložit svoje hypotézy, design a analytické strategie do časopisu pro přezkoumání před zahájením studie. Toto tvrzení může podstoupit odmítnutí a revizi stejně tak jako v typických publikačních procesech. Bude-li v této fázi studie přijata, vědci pak mohou začít své studium s vědomím, že to byl článek v principu přijat (Center for Open Science, 2015).

Jakmile je tato studie dokončena, celý článek je poslán do časopisu pro druhé kolo názorů. Článek teď nemůže být odmítnut na základě negativních výsledků. Existují pochopitelné okolnosti, za kterých se článek nepublikuje, jako jsou například nevhodná data pro testování dané hypotézy nebo například existují modifikace předem registrovaných výzkumných záměru bez dostatečného zdůvodnění těchto změn (Center for Open Science, 2015).

Empirická část

1. Měření prevalence pochybných výzkumných praktik

Simmons, Nelson a Simonsohn (2011) v sérii experimentů ukázali, jak PVP zvyšují pravděpodobnost nalezení podpory nepravdivé hypotézy. Ukazuje se, že tyto praktiky mohou být v dlouhodobém horizontu horší než úplný podvod (John, Loewenstein & Prelec, 2012). Bylo by tedy zajímavé zmapovat, kolik výzkumníků v České republice se v psychologii PVP dopouští.

1.1.Výzkumná otázka

Jaká je prevalence pochybných výzkumných praktik v psychologickém výzkumu v České republice?

1.2.Výzkumné metody

Tato studie má meziskupinový design a probíhala by prostřednictvím dotazníku, který by se zasílal výzkumníkům na email. Účastníci by anonymně odpovídali na to, zda se dopustili činností, které mohou být z hlediska metodologie a statistiky problematické v souvislosti se zvýšením pravděpodobnosti výskytu falešně pozitivního výsledku. Jednalo by se o 10 PVP, podle vzoru Johna, Loewensteina a Prelece (2012) a to –

1. neuvedení všech měřených závislých měření; zastavení sběru dat na základě průběžné analýzy;
2. nezveřejnění všech výzkumných podmínek,
3. zastavení sběru dat dříve, než bylo plánované,
4. protože bylo dosažen požadovaný výsledek,
5. zaokrouhlování p-hodnoty;
6. selektované reportování těch pokusů, které vyšly signifikantně;
7. rozhodnutí o tom zda nechat konkrétní údaje ve studii po tom, co provedu statistickou analýzu,
8. psaní o náhodně objeveném vztahu mezi proměnnými, jako kdyby byl jejich vztah předvídán od začátku;
9. tvrzení, že výsledky nejsou ovlivněny demografickými proměnnými, ačkoliv to tak ve skutečnosti není;
10. falsifikování dat.

Pokud by některou z uvedených praktik prováděli, v další části by pak uváděli, zda si myslí, že jejich činy byly obhajitelné a jak.

Míru pravdivosti odpovědí účastníků studie bych podpořila pomocí tvz. Bayesiánského séra pravdy (ang. *Bayesian truth serum*). Účastníkům by bylo oznámeno, že podle pravdivosti jejich odpovědí se budou dávat peníze charitě. Jak poznat míru pravdivosti

jejich výpovědi? Otázku ohledně PVP bychom cílili jednak na samotné výzkumníky, ale také bychom je poprosili, aby v dané činnosti ohodnotili i své kolegy – to znamená, aby uvedli, jak často se dané činnosti podle jejich názoru dopouštějí jiní. V ideálním stavu by distribuce odpovědí na obě otázky měla být stejná (protože obě otázky se týkají toho samého, jen z jiného úhlu pohledu). Čím rozdílnější jsou, tím větší je možné zkreslení. Toto zkreslení však může být dané nejenom tím, že by o sobě proband neuváděl pravdu, ale také tím, že by měl o svých kolezích horší či lepší mínění, než je tomu ve skutečnosti. John, Loewnstein a Prelec ukazují ve své studii, jak tato metoda funguje. U účastníků, u kterých byla metoda použita, se u některých položek sebehodnocení zhoršilo, zatímco hodnocení kolegů zůstalo stejné (John, Loewenstein a Prelec, 2012).

Zdá se tedy, že metoda motivuje k tomu, aby účastníci výzkumu byli v sebehodnocení upřímnější.

Návratnost dotazníků v původní studii byla relativně nízká, z oslovených vědců průzkum dokončilo pouze 36%. Dalších 33% vědců v průběhu průzkumu ukončilo, zbylí výzkumníci email ani neotevřeli. Je tedy důležité zamyslet se nad tím, jak zvýšit motivaci k vyplnění. Řešením by například mohlo být, že by průzkum musel vyplnit každý v rámci žádosti o grant. Nicméně toto patří mezi ideální řešení, jehož realizace není příliš pravděpodobná. Také by se vzhledem k tématu studie mohla paradoxně snížit pravdivost odpovědí. Další motivací by mohlo být průvodní slovo, v němž by bylo uvedeno, jak je důležité, aby dotazník vyplnili kvůli lepší situaci ve vědě a zdůraznit, že je výzkum anonymní. Anonymita by se dala posílit tím, že by bylo toto slovo obsaženo již v názvu například „Anonymní průzkum“. Mohlo by také pomoci, že explicitně nezazní téma výzkumu –“pochybné výzkumné praktiky“, ale uvede se to například názvem “ Průzkum výzkumného chování v České republice ”. Návratnost by také mohlo posílit větší množství distraktorů cílených na popis výzkumu, které by nebyli přímo spjaté s PVP praktikami, například z jakých zdrojů výzkumníci čerpají účastníky studií nebo jakým způsobem ukládají data apod.

1.3.Populace a výběr vzorku

Základní soubor budou tvořit výzkumní pracovníci v oblasti psychologie v ČR. Probandi do výběrového souboru budou získáváni pomocí kontaktních údajů z Grantové agentury České republiky za dané časové období.

1.4.Průběh výzkumu

Po přijetí mailu a kliknutí na odkaz na výzkum by byl respondent náhodně rozdělen do jedné ze dvou podmínek, které by se lišily v přítomnosti incentiva pro pravdu. Účastníci by byli po vzoru původní studie náhodně rozděleni do skupiny s přítomností incentiva pro pravdu a kontrolní skupiny v poměru 2:1. Respondentům by bylo řečeno, že jim budou prezentovány popisy různých výzkumných postupů. Ohledně každého z nich se jich budeme ptát na tyto otázky:

1) Kolik procent psychologů se podle Vás ve výzkumu dopustilo dané praktiky alespoň jednou? (na škále 0-100%).
2) Jak často se ti výzkumníci, kteří se praktiky dopustili alespoň jednou, dopouštějí? (na škále 0-100%).
3) Dopustil/a jste se někdy této konkrétní praktiky vy? (Ano/Ne)

Účastníkům by bylo sděleno, že tyto tři úhly pohledu pomohou pro přesnější odhad výskytu jednotlivých pochybení. V experimentální skupině s incentivem pro pravdu by bylo účastníkům řečeno, že tři řečené odpovědi zadáme do vzorce, jehož název je Bayesiánské sérum pravdy, jenž určí pravdivost odpovědi a tedy i částku, která se přičte charitě. Bylo by rovněž řečeno, že vlastností vzorce je, že odměňuje pravdivé odpovědi. To znamená, že pravdivé odpovědi týkající se praktiky zvýší hodnotu příspěvku určenou jménem účastníka výzkumu. Také by bylo uvedeno, že pro účely tohoto šetření není nutné, aby účastníci pochopili, jak vzorec funguje (to by jim případně pak mohlo být zasláno na mail, nebo řečeno v debriefingu). Aby bylo zajištěno, že účastníci četli a rozuměli těmto informacím, byli by požádáni o vyplnění kvízu: „Dát pravdivou odpověď v tomto výzkumu ____ množství peněz darovaných na charitu mým jménem“ (s výběrem z možností: nemá žádný vliv, snižuje a zvyšuje). Účastníci, kteří by nesprávně odpověděli, by byli automaticky přesměrováni zpět na stránku s návodem, dokud by na otázku odpověděli správně.

Dále by byli účastníci seznámeni s 10 PVP a pro každou by zodpověděli tři výše uvedené otázky. Respondenti, kteří uvedli, že tuto praktiku někdy provedli, by byli rovněž otázáni, zda si myslí, že to bylo obhajitelné. Pokud by chtěli rozvinout, proč si mysleli, že to bylo (nebo nebylo) obhajitelné, dostali by k tomu prostor. Pořadí 10 PVP by bylo určeno náhodně.

Po poskytnutí tři odpovědí pro každou z položek by účastníci byli dotázáni, zda někdy měli pochybnosti o integritě výzkumu u a) výzkumníky u jiných institucí, b) jiných výzkumnících ve své instituci, c) postgraduálními studenty, d) jejich.

Výzkum by byl ukončen demografickými otázkami na věk, pohlaví, vzdělání.

1.5.Analýza dat

Po vzoru původní studie se použijí tři analýzy. U první a druhé otázky použiji t- test a porovnám experimentální skupinu s kontrolní. Použije se také Bonferroniho korekce kvůli ověření nezávislosti dat, protože pochybných výzkumných praktik je ve studii 10.

U třetí dichotomické otázky použiji chí kvadrát - test nezávislosti. Na základě výsledků těchto analýz se usoudí, zda incentivum pro pravdu fungovalo a posléze na základě četností odpovědí odhadnu prevalenci pochybných výzkumných praktik.

Zajímavé by také bylo srovnat výsledky USA a České republiky. Toto srovnání bych provedla pomocí testu chí-kvadrát dobré shody.

1.6.Diskuze

Problematikou incentiva pro pravdu je, že hodnocení ostatních nemusí být založené na racionálním posouzení, může například docházet k egocentrickému zkreslení (Tanaka, 1993). Také jde o to, jakým způsobem jsou ostatní hodnoceni, zda se hodnotitel bere jako referenční bod, nebo u druhých předpokládá jiné chování než u sebe. Jinak se tato metoda zdá jako zatím jediná vhodná k odhalování různých témat, o kterých účastníci studií z různých důvodů nechtějí hovořit.

Zajímavé by také bylo zmapovat znalost o problematice v psychologickém výzkumu v České republice obecně. Posuzovala by se znalost o replikacích, předregistracích, publikačních zkresleních, PVP a ostatních relevantních tématech. Ve studii by pak účastníci mohli vyjádřit názor na tato témata. Také by bylo zajímavé zmapovat, nakolik výzkumníci publikují v Čechách a nakolik v zahraničí.

Závěr

Pochybné výzkumné praktiky v psychologii, s nimi spojené publikační zkreslení vrcholící nereplikovatelností výzkumů vede k oprávněné skepsi ohledně důvěryhodnosti psychologických poznatků v publikované literatuře. Mluvením a psaním o pochybných výzkumných praktikách a vyvíjením postupů, jak se jim vyvarovat, je cestou, jak tyto chyby minimalizovat či odstranit. Je také důležité o svých výzkumných výsledcích a zjištěních ostatních pochybovat a připouštět alternativní vysvětlení a kritiky. Výzkumní pracovníci by měli být posuzováni na základě kvality, nikoliv kvantity jejich práce. Finanční podpora replikací a jejich uznání by snížila překážky pro jejich publikování. Informace o výsledcích replikace by pak měly být připsány k původnímu článku.

Ve vědě je potřeba rozvíjet hodnotový systém, kde nebude pochybením nepotvrzená hypotéza, nýbrž to bude další zjištění, které nám může napovědět správnou cestu. Na vysokých školách po celém světě by byly samozřejmé replikační laboratoře, v nichž by se budoucí výzkumníci učili jednak výzkumným postupům, ale také „výzkumným mravům“.

Vědecká metoda se snaží zkoumaný objekt nebo fenomén co nejvíce poznat, a to je velmi obtížné. Věda si zaslouží respekt právě proto, že její zjištění nejsou jednoduchá, ne proto, že dostane všechno správně na první pokus. Nejistota, která teď v psychologickém výzkumu panuje, neznamená, že jí nemůžeme věřit a používat její zjištění k důležitým rozhodnutím. Znamená to jen, že bychom měli zůstat obezřetní a měli bychom být otevření změnám. Není náhoda, že každý dobrý článek na konci obsahuje frázi, že je potřeba problematiku více prozkoumat. Je potřeba se stále učit. Naštěstí si všech zmíněných problémech jsou výzkumní psychologové povětšinou vědomi, a dříve či později budou muset všichni přistoupit na přísnější pravidla.

Cílem práce bylo popsat vybrané problematické aspekty v psychologickém výzkumu, které jsem směřovala hlavně na pochybné výzkumné praktiky a publikační zkreslení. Druhým cílem bylo podat přehled metodologických opatření, které by mohly psychologický výzkum zkvalitnit. Za tato opatření jsem v této práci považovala replikaci a předregistraci.

Seznam použité literatury

Akademie věd ČR se připojuje k Sanfranciské deklaraci hodnocení výzkumu. (2013). *Akademie věd České republiky*. Retrieved from: http://data.avcr.cz/sd/novinky/hlavni-stranka/130603-sanfranciska-vyzva.html

Alogna, V. K., Attaya, M. K., Aucoin, P., Bahník, Š., Birch, S., Birt, A. R., … & Zwaan, R. A. (2014). Registered replication report: Schooler & Engstler-Schooler (1990). *Perspectives on Psychological Science*, 9, 556–578.

Antes, G., & Chalmers, I. (2003). Under-reporting of clinical trials is unethical. *The Lancet, 361*(9362), 978-979.

Bahník, Š., Vranka, M., & Dlouhá, J. (2015). X good things in life: Processing fluency effects in the "Three good things in life" exercise. *Journal of Research in Personality, 55*, 91–97.

Bargh, J. A., Chen, M. & Burrows, L. (1996). Automaticity of social behavior: Direct effects of trait construct and stereotype activation on action. *Journal of Personality and Social Psychology*, 71(2), 230–244.

Baker, M. (2016). Statisticians issue warning over misuse of P values. *Nature*, 531(7593), 151-151.

Bem, D. J. (1987). Writing the Empirical Journal Article. Retrieved from http://dbem.ws/WritingArticle.pdf

Bem, D. J. (2011). Feeling the future: experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology*, 100(3), 407–426.

Boor, M. (1982). "The citation impact factor: Another dubious index of journal quality."*American Psychologist*, 37:975-977.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *The Journal Of Abnormal And Social Psychology*, 65(3), 145-153.

Colquhoun, D. (2011). Publish-or-perish: Peer review and the corruption of science. *Theguardian*. Retrieved from https://www.theguardian.com/science/2011/sep/05/publish-perish-peer-review-science

Cumming, G. (2014). The new statist ics: Why and how. *Psychological Science*, 25, 7–29.

Curate Science. (2016). Curate Science. Retrieved from: http://curatescience.org/

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. American Psychologist, 30(2), 116-127.

Cyhlarova, E., Bell, J., Dick, J., MacKinlay, E., Stein, J., & Richardson, A. (2007). Membrane fatty acids, reading and spelling in dyslexic and non-dyslexic adults Eu*ropean Neuropsychopharmacology*, 17 (2), 116-121.

Dickersin, K.; Chan, S.; Chalmers, T. C.; et al. (1987). Publication bias and clinical trials. *Controlled Clinical Trials*, 8 (4): 343–353.

Doyen, S., Klein, O., Pichon, C. L. & Cleeremans, A. (2012). Behavioral priming: it's all in the mind, but whose mind? *PloS One*, 7(1), e29081.

Easterbrook, P. J. (1991) Publication bias in clinical research. *Lancet* 337 (8746): 867–872.

Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference in psychological research. *Psychological Review*, 70, 193–242. doi:10.1037/h0044139

Enserink, M (2011). Dutch university sacks social psychologist over faked data *Science.*

Fanelli, D., & Scalas, E. (2010-4-7). "Positive" Results Increase Down the Hierarchy of the Sciences. *Plos One*, 5(4), e10068-.

Fomel, Sergey; Claerbout, Jon (2009). "Guest Editors' Introduction: Reproducible Research". *Computing in Science and Engineering* 11 (1): 5–7.

Fiedler, K., & Schwarz, N. (2015). Questionable Research Practices Revisited. *Social Psychological And Personality Science*, 7(1), 45-52.

Flather, M. D., Farkouh, M. E., Pogue, J. M., & Yusuf, S. (1997). Strengths and limitations of meta-analysis: larger studies may be more reliable. *Controlled clinical trials*, 18(6), 568-579.

Francis, G. (2013). Replication, statistical consistency, and publication bias (with discussion).*Journal of Mathematical Psychology* 57, 153–169.

Frank, M. C., & Saxe, R. (2012). Teaching Replication. *Perspectives On Psychological Science*, 7(6), 600-604.

Gelman, A., & Loken, E. (2013). The garden of forking paths: Why multiple comparisons can be a problem, even when there is no "fishing expedition" or "p-hacking" and the research hypothesis was posited ahead of time. Unpublished manuscript.

Gelman, A., & Shalizi, C. R. (2013). Philosophy and the practice of Bayesian statistics. *British Journal Of Mathematical And Statistical Psychology*, 2013(66), 8–38.

Gilbert, D. T., King, G., Pettigrew, S., & Wilson, T. D. (2016). Comment on "Estimating the reproducibility of psychological science". *Science*, 351(6277), 1037-1037.

Gomes, C. M., & McCullough, M. E. (2015). The effects of implicit religious primes on dictator game allocations: A preregistered replication experiment. *Journal of Experimental Psychology: General*, 144(6), e94–e104.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015-3-13). The Extent and Consequences of P-Hacking in Science. *Plos Biology*, 13(3), e1002106-.

Ioannidis, J. P. A. (2005). Why Most Published Research Findings Are False. *Plos Medicine*, 2(8), e124.

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. *Psychological Science*, 23(5), 524–532.

John, L. K., Loewenstein, G. & Prelec, D. (2012). Measuring the Prevalence of Questionable Research Practices With Incentives for Truth Telling. (Supplementary Materials)

Kahneman, D.(2012) "A proposal to deal with questions about priming effects" *Nature.*

Kicinski, M., Springate, D. A., & Kontopantelis, E. (2015). Publication bias in meta-analyses from the Cochrane Database of Systematic Reviews. *Statistics In Medicine*, 34(20), 2781-2793.

Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Bahnik, S., Bernstein, M. J., … Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), 142–152.

Lehrer, J. (2010). The truth wears off: Is there something wrong with the scientific method? *The New Yorker*, 52–57.

Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70, 151–159.

Lurquin, J. H., Michaelson, L. E., Barker, J. E., Gustavson, D. E., von Bastian, C. C., Carruth, N. P., et al. (2016) No Evidence of the Ego-Depletion Effect across Task Characteristics and Individual Differences: A Pre-Registered Study. Retrivered from: http://doi.org/10.1371/journal.pone.0147770

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?: 980-1037. *American Psychologist*, 70(6), 487-498.

Mazar, N., Amir, O., & Ariely, D. (2008). The Dishonesty of Honest People: A Theory of Self-Concept Maintenance. *Journal Of Marketing Research*, 45(6), 633-644.

Mazar, N., & Ariely, D. (2015). Dishonesty in scientific research. *Journal Of Clinical Investigation*, 125(11), 3993-3996.

Novák, O. (2016). *Publikační zkreslení v meta-analýzách: Jak ho poznat a jak se mu vyhnout.* (Seminární práce) Univerzita Karlova, Praha.

Open Science Collaboration (2015). "Estimating the reproducibility of Psychological Science". *Science* 349 (6251)

Palmer, K. (2016). Psychology Is in Crisis Over Whether It's in Crisis. Retrieved from: http://www.wired.com/2016/03/psychology-crisis-whether-crisis/

Ranstam, J. (2012). Why the P-value culture is bad and confidence intervals a better alternative. *Osteoarthritis And Cartilage*, 20(8), 805-808.

Rohrer, D., Pashler, H., & Harris, C. R. (2015). Do subtle reminders of money change people's political views? *Journal Of Experimental Psychology*: General, 144(4), e73-e85.

Rosenthal R. Experimenter effects in behavioral research. New York, NY: *Appleton-Century-Crofts*, 1966. 464 p

Rosenthal R. (1979) File drawer problem and tolerance for null results. *Psychol Bull*;86:638-41.

Sackett, D. L. (1979). Bias in analytic research. *Journal of chronic diseases*, 32(1), 51-63.

Shalvi, S., Gino, F., Barkan, R., & Ayal, S. (2015). Self-Serving Justifications: Doing Wrong and Feeling Moral. *Current Directions In Psychological Science*, 24(2), 125-130.

Schimmack, U. (2012). The ironic effect of significant results on the credibility of multiple-study articles. *Psychological Methods*, 17(4), 551-566

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology*, 13, 90–100.

Schooler, J. W., & Engstler-Schooler, T. Y. (1990). Verbal overshadowing of visual memories: Some things are better left unsaid. *Cognitive Psychology*, 22, 36–71.

Schooler, J. W. (2011). Unpublished results hide the decline effect. *Nature*, 470, 437.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, 515(7525), 9-9.

Schwartz, M. A. (2008). The importance of stupidity in scientific research. *Journal Of Cell Science*, 121(11), 1771-1771.

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Pasquale, Aust, F., Awtrey, E. C., … Nosek, B. A. (2015). Many analysts, one dataset: Making transparent how variations in analytical choices affect results. Retrieved from: osf.io/gvm2z

Simmons, J. P., Nelson, L. D. & Simonsohn, U. (2011). False-positive psychology undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, 22(11), 1359–1366.

Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal Of Experimental Psychology*: General, 143(2), 534-547.

Sterling, T. D. (1959). Publication Decisions and their Possible Effects on Inferences Drawn from Tests of Significance or Vice Versa. *Journal Of The American Statistical Association*, 54(285), 30-34.

Sterling, T. D., Rosenbaum, W. L., & Weinkam, J. J. (1995). Publication Decisions Revisited: The Effect of the Outcome of Statistical Tests on the Decision to Publish and Vice Versa. *The American Statistician*, 49(1), 108-112.

Srivastava, S. (2012). Some reflections on the Bargh-Doyen elderly walking priming brouhaha. The hardest science. Retrivered from: https://hardsci.wordpress.com/2012/03/12/some-reflections-on-the-bargh-doyen-elderly-walking-priming-brouhaha/

Tanaka, K. 'ichiro. (1993). Egocentric bias in perceived fairness: Is it observed in Japan? *Social Justice Research*, 6(3), 273-285.

"The Statistical Power Of Abnormal-Social Psychological research: A Rewev" by Jacob Cohen. (2015). *Replicability index*. Retrieved from https://replicationindex.wordpress.com/2015/09/22/the-statistical-power-of-abnormal-social-psychological-research-a-revew-by-jacob-cohen/

Thorne, F. C. The citation index: Another case of spurious validity. *Journal of Clinical Psychology*, 1977, 33, 1157-1161.

Wagenmakers, E., Wetzels, R., Borsboom, D. & van der Maas, H. (2011). Why psychologists must change the way they analyze their data: the case of psi: comment on Bem (2011). *Journal of Personality and Social Psychology*, 100(3), 426–432.

Wicherts, J. M.; Borsboom, D.; Kats, J.; Molenaar, D. (2006). "The poor availability of psychological research data for reanalysis". *American Psychologist* 61 (7): 726–728.

Yong, E. (2012). Bad Copy: In The wake of high- profile controversies, psycholpgist are facing up to problems with replication. *Nature*, 298-300.