

Replikační krize ve vědě a její přesah do kinantropologie: úvod do problematiky

Ladislav Baloun* a Jiří Mališ

Fakulta tělesné kultury, Univerzita Palackého v Olomouci, Olomouc, Česká republika

Copyright: © 2023 L. Baloun & J. Mališ. Toto je open access článek vydaný pod Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>).

Východiska: V posledních deseti letech otrásá světovou vědeckou komunitou (zejména v sociálně-vědním a biomedicinském výzkumu) neschopnost zopakovat výsledky dříve provedených vědeckých studií. Pro tento fenomén se vžil název replikační krize. **Cíle:** Cílem této eseje je uvedení čtenáře do problematiky takzvané replikační krize ve vědě a popsat její možné příčiny. V první části zmíníme důležitost replikací při vzniku vědeckých poznatků, rozebereme základní terminologii sledovaného fenoménu a představíme si mediálně nejznámější příklady neúspěšných replikací v psychologickém a biomedicinském výzkumu. V další části se zaměříme na možné příčiny replikační krize, jako jsou pochybné vědecké praktiky v analýze dat se zaměřením na nesprávné použití p hodnot. V poslední části uvedeme výzvy, které upozorňují na problematiku replikační krize v kinantropologii. **Závěry:** Přesto, že aktuálně nemáme přímé důkazy o existenci replikační krize v kinantropologii, existují pochybnosti o vědecké praxi při aplikaci inferenční statistiky, zejména v souvislosti s použitím p hodnot. Z tohoto důvodu je důležité problematiku sledovat a zahájit širokou diskusi. Naší snahou není podat detailní přehled o tématu, ale seznámit čtenáře se základními myšlenkami a zdroji z oblasti replikační krize tak, aby byl schopen si v případě zájmu najít více relevantních informací.

Klíčová slova: kognitivní zkreslení, HARKing, p hodnota, p -hacking, pochybné vědecké praktiky, testování hypotéz

Úvod

Aristoteles tvrdí, že je lidskou přirozeností toužit po pochopení světa kolem nás (Purtill, 1970). Tato přirozená touha po poznání je základním kamenem vědy, která se snaží systematicky sbírat data a tato data dále zpracovat způsobem, který ji pomůže poznat svět takový, jaký opravdu je (Rudner, 1961). Existují různé přístupy, jak svět kolem nás poznávat. Vědecký přístup je specifický tím, že jde o systematickou činnost, která vytváří a organizuje znalosti ve formě ověřitelných vysvětlení o světě (Wilson, 1999). Vědecký přístup nám tedy pomáhá pochopit svět takový, jaký opravdu je.

Nezdar a vyvracení zastaralých vědeckých „objevů“ je součástí vědeckého vývoje, pokud se ale vědci uchylují k problematickým praktikám a jejich apriorní motivací není snaha pochopit svět a jeho fungování, pak se hovoří o tzv. „špatné vědě“ (Goldacre, 2009). Mnozí vědci se domnívají, že způsob, jakým je systémově nastaven svět vědy, tento špatný přístup přímo podporuje

(Horton, 2015; Smaldino & McElreath, 2016). To má mnohé neblahé důsledky, mezi které patří: plýtvání finančními zdroji, nedůvěra ve vědu, potenciálně zdraví škodlivé dopady na člověka či životní prostředí (Goldacre, 2009; Ritchie 2021).

Jedním z důsledků „špatné vědy“ je fenomén, kterým se budeme v tomto článku podrobně zabývat a tím je „replikační krize“. Jedná se o problém opětovného ověření dřívějších vědeckých výsledků. Pokud se výsledky nedaří úspěšně potvrdit či ověřit, jedná se o selhání verifikace, a ta je nedílnou součástí dobré vědecké práce (Schlick, 1936).

1. Replikační krize

Žádný zkoumaný efekt nemůžeme přijmout jako průkazný, pokud jej není možné opakovat (Dunlap, 1926). Replikace je nezbytnou součástí falzifikace zkoumané hypotézy. Pokud je nějaký efekt pouze nalezen, ale není následně zopakován, neměl by být považován za vědecký objev (Popper, 1997). Ne náhodou bylo motto první evropské vědecké společnosti Accademia del Cimento, založené v 17. století ve Florencii, „provando e riprovando“, neboli „dokaž a dokaž znovu“, či „otestuj a otestuj znovu“ (Harris, 2017). Právě neúspěšná re-

* Korespondenční adresa: Mgr. Ladislav Baloun, Ph.D., Katedra aplikovaných pohybových aktivit, Fakulta tělesné kultury, Univerzita Palackého v Olomouci, tř. Míru 117, 779 00 Olomouc, Česká republika, email: ladislav.baloun@upol.cz

plikace vědeckých hypotéz nezávislími vědci je fenomén, na který je upozorňováno minimálně od počátku devatenáctého století (Babbage, 1830), ale do širšího povědomí se dostal až v posledním desetiletí, a to díky celé řadě bizarních studií, které se staly mediálně známými (Ritchie, 2021).

Emblematickou studií je ta, kterou publikoval Daryl Bem (2011), jejíž výsledky televizní moderátor označil jako „Time travelling porn“ (The Colbert Report, 2011). V této studii se dochází k závěru, že lidé jsou schopni předpovídat, jaký obrázek jim počítač náhodně vygeneruje v případě, že se jedná o pornografické obrázky. Studie byla přijata v prestižním vědeckém časopise „Journal of Personality and Social Psychology“. Problém byl nejen v tom, že přinesl velice překvapivé výsledky, které byly podepřeny pouze slabými velikostmi účinku, ale také v tom, že časopis odmítl přijmout replikační studii, která výsledky původní studie vyvrátila, a tím přímo přispěl k šíření takto problematických výsledků, které bychom mohli bez nadsázky označit za potvrzení o existenci paranormálních jevů (Ritchie, 2021). Bemovou studií to bohužel nekončí. Mezi ty mediálně nejznámější patří studie na základě zcela vymyšlených výsledků od Stapela¹ a Lindenberg (2011), za kterou byla hlavnímu autorovi pozastavena profesura na univerzitě (Levelt Committee, 2012), nebo studie, jejíž výsledky se následně ukázaly s výrazně nižším efektem, jako byla Carlo Dweck² (2006) a její „mind set“ (Sisk et al., 2018), či několikrát nepotvrzené výsledky Amy Cuddy³ (Carney et al., 2010) jejíž videa s tzv. „Power posing“ vidělo přes 67 milionů lidí (Cuddy, 2012). Tyto kauzy přispěly ke zviditelnění replikační krize tak, jako titulní strana „How science goes wrong“ prestižního týdeníku „The Economist“ (19. října, 2013). V průzkumu prestižního časopisu Nature celkem 90 % vědců (z 1576 respondentů) souhlasí s tvrzením, že

¹ Diederik Stapel je nizozemský sociální psycholog, který byl propuštěn z Tilburské univerzity za publikování vědeckých závěrů na základě fiktivních dat. V textu je zmíněna studie, ve které píše o údajném vlivu prostředí, jako je rozbitý chodník, odpadky na ulici, na zvýšenou diskriminační činnost lidí, kteří jsou takovému prostředí vystaveni.

² Psycholožka Carol Dweck ze Stanfordské univerzity se proslavila knihou *Mindset: The New psychology of success*. V této knize rozděluje lidi na ty s tzv. „fixed mindset“ (myšlení zabrděné) a „growth mindset“ (myšlení růstové). Její teorie zjednodušeně řečeno vyzdvihuje lidi s „growth mindset“, kteří jsou ti úspěšní, a naopak ti s „fixed mindset“ jsou ti méně úspěšní. Sama pak prodává školám program zaměřený na rozvoj „growth mindset“. Funkčnost programu i samotná teorie „fixed mindset“ vs. „growth mindset“ byla mnohokrát zpochybněna (Chivers, 2017; Li & Bates, 2019).

³ Psycholožka Amy Cuddy se proslavila prezentováním tzv. „Power Posing“ neboli vlivu postavení těla na vnitřní chemické reakce a sebevědomí. Autoři následných studií na toto téma považovali zkoumaný efekt za zanedbatelný (Ranehill et al., 2015; Smith & Apicella, 2017).

existuje replikační krize a pouze 3 % se domnívalo, že žádná taková krize neexistuje (Baker, 2016).

Je potřeba dodat, že samotná neschopnost replikovat vědecké výsledky nemusí mít vždy negativní konotace (manipulace, nekompetentnost či podvod), ale může být součástí procesu, který povede k poznatkům nezbytným pro pokrok vědy. Například jde o situace, kdy dochází k měření určitého jevu různými metodami, případně když se metody postupně zpřesňují (Milton & Possolo, 2020). Jak uvádí Redish et al. (2018), „úspěch vyžaduje selhání“ a věda potřebuje čas, aby „doladila“ své výsledky. Selhání při procesu zopakování výsledků je pak kritickou součástí této cesty.

V našem příspěvku chceme poukázat na celou řadu problematických praktik vědců, kteří jsou pod systémovým tlakem na jejich publikační činnost, jejímž důsledkem je snaha vědce publikovat co nejvíce prací v co nejkratším čase. Tento publikační tlak se tak negativně odráží v problémech replikovat výsledky celé řady studií. Problém, který nastiňujeme, se týká především kvantitativního přístupu za využití inferenční statistiky a *p* hodnot, které výsledky dělí na statisticky signifikantní a statisticky nesignifikantní.

2. Replikace nebo reprodukce vědeckých výsledků a falešně pozitivní zjištění: definice pojmů

Pro pochopení kontextu replikační krize je dobré si udělat pořádek v termínech, které se v této souvislosti nejčastěji používají a zároveň, jak budeme zmíněné termíny chápat v tomto článku. Někteří zahraniční autoři používají termín replication crisis – replikační krize (Amrhein, et al., 2019; Loken & Gelman, 2017), jiní zase reproducibility crisis – krize reprodukovatelnosti (Baker, 2016). Často jsou oba pojmy replikace a reprodukce v souvislosti s krizí zaměňovány a používány jako synonyma. Důležité je poznamenat, že oba pojmy (replikace a reprodukce) jsou v různých vědeckých oborech různě chápány, jiné to je v informatice a jiné v sociálních vědách (Fidler & Wilcox, 2018). Přes některé snahy o sjednocení terminologie napříč vědními obory (Goodman et al., 2016) budeme v tomto článku využívat následující terminologii.

A. Reprodukce (jako užší termín): Opakování výsledků předchozí studie za použití stejných materiálů a postupů, jaké použil původní výzkumník. Tedy máme stejný datový soubor a využíváme stejné výpočetní operace, stejné analýzy (např. použito v National Academies of Sciences, Engineering, and Medicine, 2019; Nosek & Errington, 2020).

B. Replikace (jako širší pojem): Opakování předchozí studie, pokud jsou dodrženy stejné postupy, ale jsou shromažďována nová data. Replikace se dá dále dělit na přímou a koncepční replikaci.

B1) *Přímá replikace*: Opakování předchozí studie identickým postupem jako původní výzkumník za využití nového datového souboru.

B2) *Koncepční replikace*: Testování stejné hypotézy jako původní výzkumník za použití rozdílných metod, např. rozdílný způsob měření, rozdílný charakter výzkumného souboru jako věk a pohlaví (např. použito ve Fidler & Wilcox, 2018; Schmidt, 2009).

V souvislosti s termínem replikační krize ve vědě je také potřeba definovat pojem *falešně pozitivní výsledky/zjištění* (*false-positive results/findings*). V případě využití statistické inference v biomedicinském, sociálně-vědním, kinantropologickém či na ekonomii zaměřeném výzkumu je nejčastěji používanou metodou Null hypothesis significance testing^{4,5} (NHST; Ioannidis, 2005; Ivarsson et al., 2015; Nickerson, 2000; Ziliak & McCloskey, 2008). V rámci metody NHST je pojem nulová hypotéza používán nejčastěji k popisu předpokladu, že neexistuje žádný rozdíl mezi dvěma intervenčními skupinami, či neexistuje měřitelný účinek, či žádný vztah mezi proměnnými. V zásadě odpovídáme na binární otázku ANO/NE (např. funguje experimentální lék lépe než placebo?). Falešně pozitivní výsledek je pak ten, kdy nesprávně zamítneme nulovou hypotézu na základě stanovené prahové úrovně tzv. *p* hodnoty (konvenčně nejčastěji $p \leq 0,05$; $p \leq 0,01$; $p \leq 0,001$), i když je tato nulová hypotéza pravdivá. Ve vědecké terminologii bývá toto chybné chování označováno jako chyba typu I (Forstmeier et al., 2017; Ioannidis, 2005; National Academies of Sciences, Engineering, and Medicine, 2019).

Detailní popis problematiky geneze falešně pozitivních výsledku v biomedicinském výzkumu naleznete ve vědecké esaji od Ioannidise (2005)⁶, jím představený teoretický model je široce uplatnitelný jak ve společensko-vědním výzkumu, tak i v oblasti kinantropologie. Ioannidisova (2005) argumentace je zásadní pro pochopení pravděpodobnosti výskytu falešně pozitivních zjištění⁷ (Forstmeier et al., 2017). Vybrané aspekty rozebereme podrobněji dále v článku.

⁴ Do češtiny nejčastěji překládáno jako *statistické testování nulových hypotéz* nebo jen *testování nulových hypotéz*.

⁵ Metoda NHST se výrazně podílí na hodnocení replikovatelnosti studií, a proto je potřeba věnovat pozornost jejím limitům a adekvátnosti aktuálního využití (podrobně např. v Lakens, 2022; National Academies of Sciences, Engineering, and Medicine, 2019).

⁶ Samotný model prošel četnou kritikou a řada výzkumníků upozorňovala (Goodman & Greenland, 2007a; Goodman & Greenland, 2007b; Jager & Leek, 2014; Schimmack, 2019) na příliš katastrofické závěry a tvrzení o tom, že většina publikovaných vědeckých výsledků je falešně pozitivní. Přes nesouhlas s některými závěry, ti samí autoři potvrzují premisu ohledně nevhodného užívání metody NHST a *p* hodnot ve vědě.

⁷ Ve většině případů nekoresponduje s 5 % hladinou statistické významnosti alfa, tedy $p \leq 0,05$, ale výrazně ji překračuje.

Ve vztahu k pojmu replikační krize je potřeba také doplnit, že jsou i autoři (Fanelli, 2018), kteří se tématu replikovatelnosti a integrity vědeckého bádání věnují dlouhodobě a jsou veřejně proti používání narativu replikační krize ve vědě, jelikož to může zvýšit nedůvěru daňových poplatníků ve vědecké metody a postupy.

3. Když se replikace nedaří

Od dob publikování článku Bema (2011) a kauzy vymyšlených výsledků psychologa Stapela (podrobně popsáno v Ritchie, 2021) v roce 2011 se podařilo provést celou sérii replikačních studií v psychologii např. série s názvem Many labs 1-5 (Ebersole et al., 2016, 2020; Klein et al., 2014, 2018, 2022) či The Reproducibility Project: Psychology (Open Science Collaboration, 2015), v sociálních vědách (Camerer et al., 2018), v ekonomii (Camerer et al., 2016). Autoři těchto replikací vybrali na základě předem definovaných pravidel významné vědecké studie, které byly publikovány v prestižních časopisech pro daný obor, a poté se je pokusili zreplikovat. Replikace byly ve většině případů předem registrovány (metodika práce byla zveřejněna před sběrem dat) a záměrně se stanovil výrazně vyšší rozsah výzkumného vzorku, než měly originální studie. Společným charakterem všech výsledků těchto přímých replikací je nalezení nižšího počtu statisticky signifikantních výsledků s výrazně nižšími velikostmi účinků (effect size) oproti původním studiím. Výjimku tvořily replikace Many labs 1 (Klein et al., 2014), kde se dařilo nacházet i vyšší velikosti účinku, než měly originální studie.

Podrobněji rozebereme jeden příklad za všechny. Skupina 270 výzkumníků pod hlavičkou Open Science Collaboration (2015) vybrala 100 experimentálních a korelačních studií (97 % z nich mělo statisticky signifikantní výsledek) především z oblasti sociální a kognitivní psychologie, které byly publikovány v roce 2008 ve třech významných oborových časopisech. Při přípravě protokolu metodiky se snažili kontaktovat a konzultovat vše potřebné s autory originálních studií. Takto připravené protokoly veřejně předem registrovali, a poté provedli samotné replikace. Ve výsledku mělo statisticky signifikantní výsledek jen 36 % replikací a velikost účinku se snížila o cca 50 % oproti originálním studiím.

O reprezentativnosti, „přínosu“, významu a dopadu těchto replikací proběhla rozsáhlá diskuse (např. Stroebe, 2019). Nehledě na to, že existují i úspěšné replikace (Zwaan et al., 2018), obecně bychom asi očekávali vyšší míru replikovatelnosti studií prezentovaných ve špičkových zahraničních odborných časopisech.

V biomedicinském výzkumu způsobilo pozdvižení zjištění farmaceutických korporací Amgen (Begley &

Ellis, 2012) a Bayer (Prinz et al., 2011). Výzkumný tým pod vedením Glenna Begley (Amgen) měl za úkol vyhledávat akademické články, které prezentovaly výsledky tzv. landmarkových (přelomových) preklinických studií neboli nadějných kandidátů na lék na onkologická onemocnění. V následujícím kroku se pak zaměstnanci Amgenu pokoušeli tyto studie replikovat, a i přes velkou snahu dosáhnout úspěchů v podobě pozitivního výsledku (do svých laboratoří si dokonce pozvali některé autory originálních studií, aby zopakovali svůj postup i výsledky) byli schopni úspěšně replikovat výsledky pouze v 6 z 53 studií, to je 11 % (Harris, 2017). O trochu vyšší míru replikovatelnosti (20-25 %) popsali i zaměstnanci Bayeru (Prinz et al., 2011).

Na základě překvapivě nízké míry replikovatelnosti v preklinickém výzkumu vznikla iniciativa s názvem „The Reproducibility Project: Cancer Biology“ (Errington et al., 2014; Mullard, 2021). Cílem bylo provést přímou replikaci celkem 193 experimentů z 53 článků z oblasti preklinického výzkumu onkologických onemocnění. Záměrně byly vybrány články s největším dopadem z prestižních vědeckých časopisů. Přes obrovskou časovou a finanční náročnost se nakonec podařilo znovu provést jen 50 experimentů z 23 článků. Sami autoři (Errington et al., 2021a) přiznávají složitost zjišťování míry replikovatelnosti, ale i tak se o to pokusili. U experimentů s pozitivním účinkem byla u replikací mediánová hodnota velikosti účinku o 85 % nižší než mediánová hodnota velikosti účinku v původních experimentech, navíc byly velikosti účinků u replikací nižší než v originále v 92 % případů (Errington et al., 2021b). Mezi další důležitá zjištění patří neschopnost získat data z originálních studií ani po kontaktování samotných autorů, což se povedlo jen v 38 % případů (Errington et al., 2021a; Rodgers & Collings, 2021). Podle některých konzervativních odhadů (Freedman et al., 2015) může být 50 % nákladů (cca 28 mld. USD) spojených s preklinickým výzkumem v USA vynaloženo na studie, které nepůjde zreplikovat.

Vysoká míra falešně pozitivních studií, a tedy neschopnost replikovat výsledky, se projevila a je dobře zdokumentována i v oblasti genomiky, konkrétně při zjišťování tzv. kandidátních genů⁸ u vybraných onemocnění jako např. u schizofrenie a deprese (Border et al., 2019; Farrell et al., 2015; Plomin, 2020; Ritchie, 2021).

Předkládané publikace popisující neschopnost replikovat předchozí zjištění s tendencí výrazně se snižující velikostí účinků jsou jen výběrem těch mediálně nejpropíranějších. Podle našeho názoru je tedy otázka

„zda má věda problém s replikacemi?“ neadekvátní, ale spíše bychom měli pátrat „jak moc velký tento problém je?“. Zároveň se nám formuluje otázka navazující, „co se skrývá za důvodem této neschopnosti úspěšně replikovat prestižní vědecké studie?“.

4. Pochybné vědecké praktiky a strategie vedoucí ke zkreslování výsledků

Steneck (2006) definuje ve svém rámci tři typy chování vědeckých pracovníků:

a. Odpovědně prováděný výzkum, což chápeme jako ideální chování, které je žádoucí;

b. Fabrikaci (vymyšlení výsledků), falzifikaci (záměrnou úpravu či vědomou manipulaci výsledků) a plagiátorství (společně bývají označovány jako FFP), což chápeme jako nežádoucí chování, které je prováděno záměrně, a proto by se mělo perzekuovat;

c. Pochybné vědecké praktiky (z anglického Questionable Research Practices, dále jen QRPs), které se nachází v tzv. šedé zóně, někde mezi ideálním a nežádoucím chováním a často jsou projevem neznalosti než záměrné manipulace.

Mezinárodní definice QRPs aktuálně neexistuje, ale běžně se označují jako postupy a praktiky porušující metodologické zásady, které falešně zvyšují pravděpodobnost nalezení důkazů pro podporu zkoumané hypotézy. Zjednodušeně výzkumník selektivně aplikuje libovolné metody, tak aby dosáhl konkrétního (mnohdy předem daného) výsledku. Co je důležité, podstata metody nemusí být pochybná, nebezpečnou se stane až v případě selektivní aplikace⁹. Jako příklad uvádíme selektivní zveřejňování výsledků (cherry picking), vyloučení datových bodů na základě post hoc kritérií, stanovení hypotézy až po zjištění výsledků neboli HARKing (z anglického *hypothesizing after the results are known*), p-hacking (podrobněji popíšeme níže) či zatajování střetu zájmů. V silně kompetitivním prostředí boje o vědecké prostředky (nejčastěji finance) pak tyto QRPs fungují jako steroidy, které uměle zvyšují výkon. Zmíněné praktiky nemusí být nutně nesprávné a v některých situacích je lze legitimně využít, ale vyvolávají otázky. Často jsou prováděny z důvodu neznalosti a vědci mnohdy ani netuší, že jde o špatný postup (Bechhofer & Dattani, 2021; Bouter et al., 2016; Gerrits et al., 2019; John et al., 2012; Xie et al., 2021).

Data z posledních let ukazují (Agnoli et al., 2017; Fraser et al., 2018; Gerrits et al., 2019; John et al., 2012) překvapivě vysokou míru výskytu QRPs u aka-

⁸ Termínem kandidátní gen se označuje gen, o němž se předpokládá, že souvisí s určitým znakem, jako je nemoc nebo fyzická vlastnost. Vzhledem ke svému genomickému umístění nebo známé funkci se předpokládá, že gen hraje roli v daném znaku, a proto je kandidátem na další studium (National Human Genome Research Institute, 2023).

⁹ Pokud nezahrnu do analýzy odlehle hodnoty jen proto, že chci cíleně snížit p hodnotu na méně než 0,05, a prohlásit můj výsledek za statisticky signifikantní nález, pak jde o p-hacking. Pokud dopředu jasně komunikuji (ideálně v prerigistračním formuláři), že data umístěná za úroveň 2. či 3. směrodatné odchylky nebudou zahrnuta do analýzy, pak jde o legitimní postup.

demických pracovníků, oproti tomu závažné prohřešky ve formě FFP přiznává pouze nepatrná část vědců. U vybraných QRPs je přiznalo i více jak 50 % dotazovaných vědců. Fanelliho (2009) starší metaanalýza uvádí, že téměř 34 % dotazovaných vědců přiznalo alespoň jednu QRP, novější metaanalýza (Xie et al., 2021) uvádí aspoň jednu přiznanou QRP u 12,5 % dotazovaných výzkumníků.

Dle Fiedlera a Schwarze (2016) je prevalence těchto praktik chybně nadhodnocena. Důvodem této nepřesnosti je dle jejich názoru skutečnost, že se při prezentaci výsledků nerozlišovalo, zda šlo o jednu provedenou QRP v průběhu celé vědecké kariéry či jen o ojedinělý výstřelek, který se neopakoval. Zároveň dotazování vědci neměli často prostor k dovysvětlení kontextu praktikování QRPs.

Prozatím asi největší, nejkomplexnější a metodologicky nejpropracovanější šetření tohoto typu bylo provedeno v Nizozemí (Gopalakrishna et al., 2022) na vzorku 6813 vědeckých pracovníků, přičemž 51,3 % uvedlo, že se často zapojuje alespoň do jedné QRPs. Zde je určitě potřeba dodat, že v tomto šetření zahrnovala definice QRPs mnohem širší portfolio chování, oproti starším článkům (Agnoli et al., 2017; Fraser et al., 2018; Fiedler & Schwarz, 2016; John et al., 2012), kde se zaměřili primárně na manipulace při analýze dat.

Je pravděpodobné, že velká část vědeckých pracovníků často ani netuší, že se dopouští něčeho špatného, především jde o QRPs, které se vztahují ke zpracování, analýze dat a interpretaci výsledků inferenční statistiky. Tuto domněnku podporují výsledky šetření, kde jsou akademici dotazováni na jejich interpretaci statistických testů v podobě p hodnot. Nedostatečná znalost správné interpretace výsledků na základě použití p hodnot byla zjištěna například na vzorku dotazovaných psychologů (Badenes et al., 2015, 2016), sociálních vědců a ekonomů (Lyu et al., 2020), či lékařů (Andreu et al., 2021).

Není tajemstvím, že mnoho výzkumníků (Cohen, 1994; De Groot, 2014; Rosenthal, 1979; Rosnow & Rosenthal, 1989) varovalo již v průběhu 20. století před nevhodným užíváním p hodnot ve vědecké komunitě. Ziliak a McCloskey (2008) pak důsledně zmapovali, jak „kult statistické významnosti“, tedy odklon od hledání věcné významnosti a soustředění se na významnost statistickou, poškozuje mnoho vědních oborů od ekonomie, přes psychologii až po medicínu.

Důkazy o chybovosti při prezentování výsledků podala skupina nizozemských výzkumníků Nuijten et al. (2016). Pro hodnocení prevalence výskytu chyb v souvislosti s používáním p hodnot využili algoritmus s názvem *Statcheck* (podrobnosti o aplikaci jsou dostupné na <https://michelenuijten.shinyapps.io/statcheck-web/>), který funguje jako kontrola „pravopisu“ pro statistiku. Konkrétně zjišťuje, zda p hodnoty odpovídají jejich doprovodné testovací statistice a stupňům

volnosti. Do programu *Statcheck* vložili více jak 250 tisíc p hodnot z 16 tisíc odborných psychologických článků, kde byly prezentovány potřebné statistiky, tak aby identifikovali možné chyby. Algoritmus funguje, jelikož mnoho čísel ve statistických testech je na sobě závislých, takže znalost jen některých z nich vždy umožňuje reprodukovat ostatní. Stejně jako když používáme Pythagorovu větu a chceme vypočítat přeponu trojúhelníku pomocí délky ostatních dvou stran (Ritchie, 2021). Výsledky ukázaly, že téměř 50 % sledovaných článků mělo aspoň jednu číselnou nesrovnalost. Dle autorů ale byla většina těchto chyb menšího rozsahu. Velké nesrovnalosti¹⁰ pak byly nalezeny u přibližně 12 % analyzovaných článků. Taková nesrovnalost může vzniknout z mnoha důvodů, jako například překlep, nepozornost při přepisu až po vědomé podvody. Zajímavé ale je, že větší pravděpodobnost vzniku takové chyby je u situace, kdy se nesignifikantní výsledek stane „omylem“ výsledkem signifikantním. To dle autorů naznačuje systematické zkreslování ve prospěch statisticky signifikantních výsledků.

S rostoucím povědomím o replikační krizi v odborných kruzích popsala své znepokojení britská psycholožka Dorothy Bishop (2019) a formulovala termín *Čtyři jezdci replikační apokalypsy*. Dle jejího názoru jsou hlavními čtyřmi vědeckými neřestmi, které vedou k nereplikovatelnému výzkumu:

1. Publikací zkreslení (publication bias) – Vědci publikují méně článků s výsledky, které neprokazují žádný účinek, oproti článkům, které účinek našly. Zároveň editoři vědeckých časopisů je méně často přijímají. To znamená, že vytváříme předsudky vůči nulovým hypotézám.
2. Nízká síla statistických testů (low statistical power) – Pokud má studie malý výzkumný vzorek (sample size) a hledaný účinek (či souvislost) je malý, pak je pravděpodobné, že reálný účinek nezjistíte, i když v realu existuje.
3. P-hacking – Jde o strategie, kdy hledáte jen statisticky významné výsledky a ty pak publikujete. P-hacking má mnoho podob, např. mnohočetné testování hypotéz, neboli na datovém souboru provedu testování mnoha hypotéz, ale prezentuji pak jen některé.
4. HARKing – Stanovení hypotéz až poté, kdy je znám výsledek datové analýzy. Obdobně jako p-hacking je natolik rozšířený, že tento způsob práce výzkumníci často považují za dobrou praxi.

Na základě výše uvedených problémů s neúspěšnými replikacemi a četných kritik v odborném i veřejném prostoru na téma „nevhodné používání p hodnot

¹⁰ Jako velkou označují autoři tu nesrovnalost, která změní statistický závěr, kdy se díky nepřesnosti stane výsledek signifikantní, místo aby byl nesignifikantní, nebo naopak (např. místo správné hodnoty $p=0,06$ bylo uvedeno $p=0,049$).

či *p-hackingu*” (Nuzzo, 2014; Simmons et al., 2011) se rozhodla Americká asociace statistiků (American Statistical Association; ASA) vydat prohlášení s názvem *ASA Statement on statistical significance and p-values* (Wasserstein & Lazar, 2016). V tomto dokumentu ASA vyjadřuje znepokojení nad zneužíváním a dezinterpretací *p* hodnot ve vědecké komunitě a dává šest doporučení jak (ne)pracovat s *p* hodnotami:

1. *P* hodnoty mohou naznačovat, nakolik jsou data neslučitelná se zadaným statistickým modelem.
2. *P* hodnoty neměří pravděpodobnost, že zkoumaná hypotéza je pravdivá, ani pravděpodobnost, že data vznikla pouze náhodou.
3. Vědecké závěry a obchodní nebo politická rozhodnutí by neměla být založena pouze na tom, zda *p* hodnota překročí určitou hranici.
4. Správné vyvození závěrů vyžaduje úplné podávání zpráv a transparentnost.
5. *P* hodnota neboli statistická významnost neměří velikost účinku ani důležitost výsledku.
6. Sama o sobě *p* hodnota neposkytuje dobrou míru důkazu týkajícího se modelu nebo hypotézy.

Greenland et al. (2016) pak vytvořili doplňující materiál k prohlášení ASA (Wasserstein & Lazar, 2016), který podrobně rozebírá 25 nejčastějších dezinterpretací, která jsou spojována s používáním *p* hodnot, konfidenčních intervalů a konceptu statistické síly (statistical power). Klíčový problém podle Greenlanda et al. (2016) je ve skutečnosti, že neexistují jednoduché intuitivní výklady pojmů spojených s inferenční statistikou (*p* hodnota, konfidenční interval atp.), naopak je pro pochopení těchto konceptů potřebná pozornost na detaily a trpělivost (Greenland et al., 2016).

Celková snaha ASA o řešení problému s nevhodným užíváním statistické inference, především pak rozebíraných *p* hodnot vyústila v roce 2019 ve vydání speciálního čísla odborného časopisu The American Statistician (TAS; Wasserstein et al., 2019). V tomto volně dostupném čísle najdete 43 článků, které z různých pohledů popisují problematiku analýzy dat a statistické inference. Cílem ASA bylo vytvořit monotematické číslo TAS, ve kterém bude problematika podrobně představena i nestatistikům, kteří využívají statistické nástroje ve vědě, obchodě či v politice. Nejzásadnějším sdělením je pak doporučení ASA odstoupit při prezentování výsledků od používání termínů „statisticky významný”, tedy dichotomizace výsledků na základě prahové hodnoty $p < 0,05$ (či jakékoli jiné). Žádná *p* hodnota nemůže odhalit věrohodnost, pravdu nebo důležitost asociace či účinku. Označení statistické významnosti tedy neznamená ani nenaznačuje, že asociace nebo účinek je vysoce pravděpodobný, skutečný, pravdivý nebo důležitý. *P* hodnoty by dle vyjádření ASA měli být prezentovány jako spojitá veličina v kontextu dalších údajů, jako velikost účinku, velikost

výzkumného vzorku či jak pravděpodobná byla vaše hypotéza před provedením výzkumu (Wasserstein et al., 2019).

Pro snížení pravděpodobnosti získání falešně pozitivních zjištění, a tím pádem zvýšení replikovatelnosti výsledků ve vědě, může výzkumník realizovat řadu postupů jako preregistrace studie před samotným provedením výzkumu, otevřený a transparentní přístup založený na principech Open science (podrobně ve Forstmeier et al., 2017; Munafó et al., 2017). Neméně důležitým krokem je kvalitní výuka a školení v oblasti analýzy dat a principů inferenční statistiky (Dienes, 2008; Gigerenzer, 2018; Munafó et al., 2017). Různé přístupy k analýze dat přináší odlišné odpovědi (Dienes, 2008). Jednou z možností je využití Bayesovského přístupu k analýze dat, kdy výzkumník dostává odpověď na otázku, která ho často skutečně zajímá¹¹ („jak pravděpodobná je moje hypotéza za předpokladu dat, které mám k dispozici”). Podstatou Bayesovského přístupu je stanovení tzv. apriorní pravděpodobnosti, která se reviduje na základě pozorovaných dat pomocí Bayesova vzorce (Świątkowski & Carrier, 2020; Wong et al., 2022).

V psychologii začíná být hojně využíván směr s názvem New Statistic (Calin-Jageman & Cumming, 2019; Cumming & Calin-Jageman, 2016), který je založen na odhadech velikostí účinků, konfidenčních intervalech a meta-analytickém myšlení. Podrobný popis všech zmíněných alternativ si zaslouží vlastní příspěvek a i proto se jím nebudeme dále zabývat.

5. Replikační krize a její přesah do kinantropologie

Kinantropologie jako vědní obor vychází výrazně z metod a postupů využitých v sociálně-vědním či biomedicínském výzkumu, proto není překvapením, že se výše zmíněné QRP a nesprávné využívání *p* hodnot vedoucí k potenciální replikační krizi může týkat i našeho oboru. Prozatím jsme nezaznamenali žádné cílené pokusy o preregistrované replikační studie ve stylu Open Science Collaboration (2015) či Reproducibility Project: Cancer Biology (Errington et al., 2014; Mullard, 2021).

Jedna z prvních vlaštovek přinášející empirická data o možnosti systematického zkreslení publikovaných

¹¹ Frekventistický přístup (nejčastěji reprezentovaný metodou NHST), založený na *p* hodnotách nám bohužel tuto otázku nemůže zodpovědět (Cohen, 1994). Odpovídá nám na otázku, „jak pravděpodobný je výskyt našich dat (či extrémnějších) za předpokladu platnosti nulové hypotézy“. Odpůrci této metody dokonce zastávají názor, že nejde o testování hypotéz (Schwab et al., 2011), případně dávají metodě pejorativní název *Nil Null Hypotheses*, ve volném překladu „nicotná nulová hypotéza“ (Nickerson, 2000).

výsledků z oblasti kinantropologie¹² přinesla analýza Twomey et al. (2021). Předmětem analýzy bylo posoudit míru pozitivních výsledků uváděných hypotéz. Vzorek tvořilo celkem 300 odborných článků ze tří předních vědeckých časopisů (Q1 a Q2) z oblasti „sport science“. Výsledky ukázaly, že 81 % prezentovaných hypotéz bylo hodnoceno pozitivně (hodnoceny jako statisticky významné). Pokud tyto výsledky dáme do kontextu analýzy Abt et al. (2020)¹³ docházíme k závěru, že jde o míru nepřiměřeně vysokou. Ve zkratce, pokud má studie malý rozsah výzkumného souboru, je zde velký předpoklad, že bude mít i nízkou statistickou sílu testu (např. 40 %). To znamená, že pokud by všechny testované hypotézy v oblasti kinantropologie byly v reálu pravdivé, tak bychom dokázali správně určit pouze 40 % těchto hypotéz. Zbylé hypotézy (60 %) byly tzv. falešně negativní. Z tohoto důvodu tedy nalezená 81% míra pozitivních hypotéz vzbuzuje jistou nedůvěru a naznačuje systematické zkreslení v oblasti kinantropologie.

V souvislosti s medializací fenoménu replikační krize vyzvala řada autorů (výzkumníků, editorů) z oblasti kinantropologie k velké obezřetnosti při plánování výzkumných studií a analyzování datových souborů. Borg et al. (2020) a Sainani et al. (2021) rozebírají na jednotlivých reálných příkladech (publikované články) konkrétní chyby při analýze dat. Halperin et al. (2018) upozorňují na základní metodologické nedostatky v oblasti „sport science“ vedoucí k nízké validitě výzkumných postupů. Konkrétně zmiňují nízký počet longitudinálních a replikačních studií, omezené publikování tzv. nulových (statisticky nesignifikantních) a triviálních výsledků či nedostatečná vědecká transparentnost. Obecná doporučení pak uvádí nutnost: a) úzké spolupráce kinantropologů a statistiků, metodologů; b) transparentního sdílení dat; c) výpočtu potřebného rozsahu výzkumného vzorku před provedením studie; d) vizualizace dat; e) zaznamenávat a uvádět počet provedených analýz na jednom datovém souboru (tzv.

¹² V tomto příspěvku jsou pod pojmem kinantropologie zahrnuty vědní obory používané v anglicky psané literatuře jako *Exercise Science*, *Sport Science* či *Kinesiology*.

¹³ V této analýze autoři náhodně vybrali ve stanoveném tříletém období 120 článků z časopisu *Journal of Sports Sciences*. Cílem bylo zjistit, jaká je mediánová hodnota rozsahu výzkumného vzorku v tomto prestižním časopise, který publikuje především experimentální studie. Výsledná mediánová hodnota byla 19 participantů na studii. Zároveň se analýza zajímala, zda autoři u 120 vybraných článků provedli výpočet pro stanovení minimálního rozsahu výzkumného souboru. Pouze 13 článků obsahovalo nějakou formu odpovídajícího postupu. K samotnému výpočtu statistické síly je nejčastěji potřeba stanovit hladinu alfa (α), použitý statistický test a očekávanou minimální velikost účinku. Pokud mají studie malý rozsah výzkumného vzorku (např. 19), jsou schopné odhalit jako statisticky významné pouze velké velikosti účinku. Zároveň se zvyšuje pravděpodobnost nalezení falešně pozitivních zjištění (Button et al., 2013; Mesquida et al., 2022).

mnohočetném testování hypotéz souboru; Sainani & Chamari, 2022). V kinantropologii už není výjimkou ani kritika používání metody statistického testování nulových hypotéz (NHST), v jehož středobodu stojí p hodnoty (Harrison et al., 2020; Ivarsson et al., 2015), či výzvy možného využití bayesovského přístupu při analýze dat (Hecksteden et al., 2022).

Závěr

Téma replikační krize je spojeno s bouřlivým obdobím současné vědy. Z dostupné evidence (část byla prezentována v tomto příspěvku), která byla v posledních letech nashromážděna pro danou problematiku, vyvozujeme, že problém má hluboké kořeny v samotném nastavení vědeckého systému.

Je důležité zdůrazňovat, že veškeré bádání je dynamickým zkoumáním reality a chyby jsou nedílnou součástí pokroku. Poznání je neustále se vyvíjející cestou bez konce, proto nemůžeme vědecké články přijímat jako konečnou odpověď na otázku, ale pouze jako postup, a to ne vždy tím správným (pravdivým) směrem. Replikace nám v hledání správného směru napomáhají. Je to efektivní způsob, jak ověřovat výsledky bádání, proto je zapotřebí, aby samotní vědci nebrali neúspěšné replikace svého výzkumu za nezdar a profesní selhání, ale jako přirozenou a podstatnou část vědecké praxe. K tomu by měla mimo jiné napomáhat také edukace zaměřená nejen na metodologii, ale také na etickou stránku věci.

Domníváme se, že vědecká společnost dostatečně nereflktuje závažnost tohoto problému, a pro jakoukoliv změnu bude nutná určitá forma sebereflexe. Jak píše britský psycholog Chambers (2019), jako vědci žijeme v konfliktu mezi tím, co je nejlepší pro vědu, a tím, co je nejlepší pro vědce. Tento konflikt je zapotřebí odstranit prostřednictvím změn pobídek a reorganizací celého systému, jejímž cílem by bylo dosažení souladu pobídek s kvalitou výzkumu.

V prezentovaném příspěvku jsme se snažili čtenáře uvést do fenoménu replikační krize ve vědě a jejího přesahu do kinantropologie. Přestože nemáme přímé důkazy o neschopnosti replikovat vědecké studie v oblasti kinantropologie, vyvozujeme, že i naše oblast je zasažena sledovaným problémem. Naše dedukce vychází především z těchto premis:

1. Prezentované neúspěšné replikační studie jsou především produktem biomedicínského či sociálně-vědního výzkumu (zejména psychologie). Kinantropologie jako multidisciplinární obor pak ve svých výzkumných metodách a datových analýzách vychází především z těchto oborů.

2. Není důvod očekávat, že by kinantropologové analyzovali svá data obezřetněji než psychologové či vědci z biomedicínského prostředí, proto je i v kinan-

tropologii pravděpodobný výskyt čtyř jezdců replikační apokalypsy (p-hacking, HARKing, publikační zkreslení, nízká statistická síla testů; Abt et al., 2020).

3. Problematika replikační krize je v posledním desetiletí bouřlivě diskutována vědeckou komunitou zejména na Západě (USA, Británie, Německo, Nizozemí...). V české kinantropologii se bohužel až na pár výjimek problém nedebatuje. Samotné důvody neznáme, můžeme se jen domnívat.

Kromě informování kinantropologické obce o sledovaném fenoménu je naším cílem právě zahájení věcné diskuse. V chystaném navazujícím článku bychom se pak chtěli zaměřit a podrobně rozebrat faktory a tlaky, které se podílejí na evoluci špatné vědy. Pokud čtenáře prezentované téma zaujalo, doporučujeme publikace Ritchie (2021), Chambers (2019), Harris (2017) a National Academies of Sciences, Engineering, and Medicine (2019). Komplexní a sumarizující pohled na problematiku replikační krize v kinantropologii přináší Mesquida et al. (2022).

Referenční seznam

- Abt, G., Boreham, C., Davison, G., Jackson, R., Nevill, A., Wallace, E., & Williams, M. (2020). Power, precision, and sample size estimation in sport and exercise science research. *Journal of Sports Sciences*, 38(17), 1933-1935. <https://doi.org/10.1080/02640414.2020.1776002>
- Agnoli, F., Wicherts, J. M., Veldkamp, C. L., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PloS One*, 12(3), Article e0172792. <https://doi.org/10.1371/journal.pone.0172792>
- Andreu, M. F., Ballve, L. D., Verdecchia, D. H., Monzón, A. M., & Carvalho, T. D. (2021). Is the p-value properly interpreted by critical care professionals? Online survey. *Revista Brasileira de Terapia Intensiva*, 33(1), 88-95. <https://doi.org/10.5935/0103-507X.20210009>
- Amrhein, V., Trafimow, D., & Greenland, S. (2019). Inferential statistics as descriptive statistics: There is no replication crisis if we don't expect replication. *The American Statistician*, 73(1), 262-270. <https://doi.org/10.1080/00031305.2018.1543137>
- Babbage, C. (1830). *Reflections of the decline of science in England, and on some of its causes*. R. Clay Bread-Street-Hill, Cheapside.
- Badenes, L., Frias, D., Monterde, H., & Pascual, M. (2015). Interpretation of the p value: A national survey study in academic psychologists from Spain. *Psicothema*, 27(3), 290-5. <https://doi.org/10.7334/psicothema2014.283.290-295>
- Badenes, L., Frias, D., Iotti, B., Bonilla, A., & Longobardi, C. (2016). Misconceptions of the p-value among Chilean and Italian Academic Psychologists. *Frontiers in Psychology*, 7, Article 1247. <https://doi.org/10.3389/fpsyg.2016.01247>
- Baker, M. (2016). Reproducibility crisis. *Nature*, 533(26), 353-66. <https://doi.org/10.1038/533452a>
- Begley, C. G., & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature*, 483(7391), 531-533. <https://doi.org/10.1038/483531a>
- Bechhofer, N., & Dattani, S. (2021). The speed of science. *Works in Progress*. <https://worksinprogress.co/issue/the-speed-of-science>
- 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), 407-425. <https://doi.org/10.1037/a0021524>
- Bishop D. (2019). Rein in the four horsemen of irreproducibility. *Nature*, 568(7753), Article 435. <https://doi.org/10.1038/d41586-019-01307-2>
- Bouter, L. M., Tjink, J., Axelsen, N., Martinson, B. C., & Ter Riet, G. (2016). Ranking major and minor research misbehaviors: Results from a survey among participants of four World Conferences on Research Integrity. *Research Integrity and Peer Review*, 1(1), 1-8. <https://doi.org/10.1186/s41073-016-0024-5>
- Border, R., Johnson, E. C., Evans, L. M., Smolen, A., Berley, N., Sullivan, P. F., & Keller, M. C. (2019). No support for historical candidate gene or candidate gene-by-interaction hypotheses for major depression across multiple large samples. *American Journal of Psychiatry*, 176(5), 376-387. <https://doi.org/10.1176/appi.ajp.2018.18070881>
- Borg, D. N., Lohse, K. R., & Sainani, K. L. (2020). Ten common statistical errors from all phases of research, and their fixes. *PM & R*, 12(6), 610-614. <https://doi.org/10.1002/pmrj.12395>
- Button, K. S., Ioannidis, J. P., Mokrysz, C., Nosek, B. A., Flint, J., Robinson, E. S., & Munafò, M. R. (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews. Neuroscience*, 14(5), 365-376. <https://doi.org/10.1038/nrn3475>
- Calin-Jageman, R. J., & Cumming, G. (2019). The new statistics for better science: Ask how much, how uncertain, and what else is known. *The American Statistician*, 73(Suppl. 1), 271-280. <https://doi.org/10.1080/00031305.2018.1518266>
- Camerer, C. F., Dreber, A., Forsell, E., Ho, T. H., Huber, J., Johannesson, M., ... Wu, H. (2016). Evaluating replicability of laboratory experiments in economics. *Science*, 351(6280), 1433-1436. <https://doi.org/10.1126/science.aaf0918>
- Camerer, C. F., Dreber, A., Holzmeister, F., Ho, T. H., Huber, J., Johannesson, M., ... Wu, H. (2018). Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nature Human Behaviour*, 2(9), 637-644. <https://doi.org/10.1038/s41562-018-0399-z>
- Carney, D., Cuddy, A. J., & Yap, A. J. (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science*, 21(10), 1363-8. <https://doi.org/10.1177/0956797610383437>
- Chambers, C. (2019). *The seven deadly sins of psychology: a manifesto for reforming the culture of scientific practice*. Princeton University Press.
- Chivers, T. (2017). A mindset "revolution" sweeping Britain's classes may be based on shaky science. *BuzzFeed*, 14(1). https://www.buzzfeed.com/tomchivers/what-is-your-mindset?utm_term=.000Razv2n#.ht5JOoZv9
- Cohen, J. (1994). The earth is round ($p < .05$). *American Psychological Association*, 49(12), 997-1003.
- Cuddy, A. (2012). Your body language may shape who you are [Video]. TEDGlobal. https://www.ted.com/talks/amy_cuddy_your_body_language_may_shape_who_you_are
- Cumming, G., & Calin-Jageman, R. (2016). *Introduction to the new statistics: Estimation, open science, and beyond*. Routledge, Taylor & Francis Group.
- De Groot, A. D. (2014). The meaning of "significance" for different types of research [translated and annotated by Eric-Jan

- Wagenmakers, Denny Borsboom, Josine Verhagen, Rogier Kievit, Marjan Bakker, Angelique Cramer, Dora Matzke, Don Mellenbergh, and Han L. J. van der Maas]. *Acta Psychologica*, 3(148), 188-194. <https://doi.org/10.1016/j.actpsy.2014.02.001>
- Dienes, Z. (2008). *Understanding psychology as a science: an introduction to scientific and statistical inference*. Palgrave Macmillan.
- Dunlap, K. (1926). The experimental methods of psychology. In C. Murchison (Ed.), *Psychologies of 1925* (pp. 331–351). Clark University Press.
- Dweck, C. S. (2006). *Mindset: The new psychology of success*. Random House.
- Ebersole, C. R., Atherton, O. E., Belanger, A. L., Skulborstad, H. M., Allen, J. M., Banks, J. B., ... & Nosek, B. A. (2016). Many Labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology*, 67, 68-82. <http://dx.doi.org/10.1016/j.jesp.2015.10.012>
- Ebersole, C. R., Mathur, M. B., Baranski, E., Bart-Plange, D. J., Buttrick, N. R., Chartier, C. R., ... Szecsi, P. (2020). Many Labs 5: Testing pre-data-collection peer review as an intervention to increase replicability. *Advances in Methods and Practices in Psychological Science*, 3(3), 309-331. <https://doi.org/10.1177/25152459209586>
- Errington, T. M., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021a). Reproducibility in cancer biology: Challenges for assessing replicability in preclinical cancer biology. *Elife*, 10, Article e67995. <https://doi.org/10.7554/eLife.67995>
- Errington, T. M., Iorns, E., Gunn, W., Tan, F. E., Lomax, J., & Nosek, B. A. (2014). An open investigation of the reproducibility of cancer biology research. *Elife*, 3, Article e04333. <https://doi.org/10.7554/eLife.04333>
- Errington, T. M., Mathur, M., Soderberg, C. K., Denis, A., Perfito, N., Iorns, E., & Nosek, B. A. (2021b). Investigating the replicability of preclinical cancer biology. *Elife*, 10, e71601. <https://doi.org/10.7554/eLife.71601>
- Fanelli, D. (2009). How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data. *PloS One*, 4(5), Article e5738. <https://doi.org/10.1371/journal.pone.0005738>
- Fanelli, D. (2018). Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences*, 115(11), 2628-2631. <https://doi.org/10.1073/pnas.1708272114>
- Farrell, M. S., Werge, T., Sklar, P., Owen, M. J., Ophoff, R. A., O'Donovan, M. C., ... Sullivan, P. F. (2015). Evaluating historical candidate genes for schizophrenia. *Molecular Psychiatry*, 20(5), 555-562. <https://doi.org/10.1038/mp.2015.16>
- Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, 7(1), 45-52.
- Fidler, F., & Wilcox, J. (2018). Reproducibility of scientific results. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy*. <https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility/>
- Forstmeier, W., Wagenmakers, E. J., & Parker, T. H. (2017). Detecting and avoiding likely false-positive findings – A practical guide. *Biological Reviews of the Cambridge Philosophical Society*, 92(4), 1941–1968. <https://doi.org/10.1111/brev.12315>
- Fraser, H., Parker, T., Nakagawa, S., Barnett, A., & Fidler, F. (2018). Questionable research practices in ecology and evolution. *PloS One*, 13(7), Article e0200303. <https://doi.org/10.1371/journal.pone.0200303>
- Freedman, L. P., Cockburn, I. M., & Simcoe, T. S. (2015). The economics of reproducibility in preclinical research. *PLoS Biology*, 13(6), Article e1002165. <https://doi.org/10.1371/journal.pbio.1002165>
- Gerrits, R. G., Jansen, T., Mulyanto, J., van den Berg, M. J., Klazinga, N. S., & Kringos, D. S. (2019). Occurrence and nature of questionable research practices in the reporting of messages and conclusions in international scientific Health Services Research publications: A structured assessment of publications authored by researchers in the Netherlands. *BMJ Open*, 9(5), Article e027903. <https://doi.org/10.1136/bmjopen-2018-027903>
- Gigerenzer, G. (2018). Statistical rituals: The replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2), 198-218. <https://doi.org/10.1177/2515245918771329>
- Goldacre, B. (2009). *Bad science*. Fourth Estate.
- Goodman, S. N., Fanelli, D., & Ioannidis, J. P. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), Article 341ps12. <https://doi.org/10.1126/scitranslmed.aaf5027>
- Goodman, S., & Greenland, S. (2007a). Assessing the unreliability of the medical literature: A response to “Why most published research findings are false.”. *Johns Hopkins University, Dept. of Biostatistics Working Papers*. Working Paper 135. <https://biostats.bepress.com/cgi/viewcontent.cgi?article=1135&context=jhubiostat>
- Goodman, S., & Greenland, S. (2007b). Why most published research findings are false: Problems in the analysis. *PLoS Medicine*, 4(4), Article e168. <https://doi.org/10.1371/journal.pmed.0040168>
- Gopalakrishna, G., Ter Riet, G., Vink, G., Stoop, I., Wicherts, J. M., & Bouter, L. M. (2022). Prevalence of questionable research practices, research misconduct and their potential explanatory factors: A survey among academic researchers in the Netherlands. *PloS One*, 17(2), Article e0263023. <https://doi.org/10.1371/journal.pone.0263023>
- Greenland, S., Senn, S. J., Rothman, K. J., Carlin, J. B., Poole, C., Goodman, S. N., & Altman, D. G. (2016). Statistical tests, p values, confidence intervals, and power: A guide to misinterpretations. *European Journal of Epidemiology*, 31(4), 337-350. <https://doi.org/10.1007/s10654-016-0149-3>
- Halperin, I., Vigotsky, A. D., Foster, C., & Pyne, D. B. (2018). Strengthening the practice of exercise and sport-science research. *International Journal of Sports Physiology and Performance*, 13(2), 127-134. <https://doi.org/10.1123/ijssp.2017-0322>
- Harris, R. F. (2017). *Rigor mortis: How sloppy science creates worthless cures, crushes hope, and wastes billions*. Basic Books.
- Harrison, A. J., McErlain-Naylor, S. A., Bradshaw, E. J., Dai, B., Nunome, H., Hughes, G. T., ... Fong, D. T. (2020). Recommendations for statistical analysis involving null hypothesis significance testing. *Sports Biomechanics*, 19(5), 561-568. <https://doi.org/10.1080/14763141.2020.1782555>
- Hecksteden, A., Forster, S., Egger, F., Buder, F., Kellner, R., & Meyer, T. (2022). Dwarfs on the shoulders of giants: Bayesian analysis with informative priors in elite sports research and decision making. *Frontiers Sports Active Living*, 4, Article 793603. <https://doi.org/10.3389/fspor.2022.793603>
- Horton, R. (2015). Offline: What is medicine's 5 sigma. *Lancet*, 385(9976), Article 1380. [https://doi.org/10.1016/S0140-6736\(15\)60696-1](https://doi.org/10.1016/S0140-6736(15)60696-1)

- Ioannidis J. P. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), Article e0020124. <https://doi.org/10.1371/journal.pmed.0020124>
- Ivarsson, A., Andersen, M. B., Stenling, A., Johnson, U., & Lindwall, M. (2015). Things we still haven't learned (so far). *Journal of Sport and Exercise Psychology*, 37(4), 449-461. <https://doi.org/10.1123/jsep.2015-0015>
- Jager, L. R., & Leek, J. T. (2014). An estimate of the science-wise false discovery rate and application to the top medical literature. *Biostatistics*, 15(1), 1-12. <https://doi.org/10.1093/biostatistics/kxt007>
- 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. <https://doi.org/10.1177/0956797611430953>
- Klein, R. A., Cook, C. L., Ebersole, C. R., Vitiello, C., Nosek, B. A., Hilgard, J., . . . Ratliff, K. A. (2022). Many Labs 4: Failure to replicate mortality salience effect with and without original author involvement. *Collabra: Psychology*, 8(1), Article 35271. <https://doi.org/10.1525/collabra.35271>
- Klein, R. A., Ratliff, K. A., Vianello, M., Adams, R. B., Jr., Bahník, Š., Bernstein, M. J., . . . Nosek, B. A. (2014). Investigating variation in replicability: A "many labs" replication project. *Social Psychology*, 45(3), 142-152. <http://doi.org/10.1027/1864-9335/a000178>
- Klein, R. A., Vianello, M., Hasselman, F., Adams, B. G., Adams Jr, R. B., Alper, S., . . . Sowden, W. (2018). Many Labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science*, 1(4), 443-490. <https://doi.org/10.1177/2515245918810225>
- Lakens, D. (2022). *Improving your statistical inferences*. Retrieved from https://lakens.github.io/statistical_inferences/ <https://doi.org/10.5281/zenodo.6409077>
- Levitt Committee. (2012). *Flawed science: The fraudulent research practices of social psychologist Diederik Stapel* [English translation], 28 Nov. 2012; <https://osf.io/eup6d>
- Li, Y., & Bates, T. C. (2019). You can't change your basic ability, but you work at things, and that's how we get hard things done: Testing the role of growth mindset on response to setbacks, educational attainment, and cognitive ability. *Journal of Experimental Psychology: General*, 148(9), 1640-1655. <https://doi.org/10.1037/xge0000669>
- Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, 355(6325), 584-585. <https://doi.org/10.1126/science.aal3618>
- Lyu, X. K., Xu, Y., Zhao, X. F., Zuo, X. N., & Hu, C. P. (2020). Beyond psychology: Prevalence of p value and confidence interval misinterpretation across different fields. *Journal of Pacific Rim Psychology*, 14, Article e6. <https://doi.org/10.1017/prp.2019.28>
- Mesquida, C., Murphy, J., Lakens, D., & Warne, J. (2022). Replication concerns in sports and exercise science: A narrative review of selected methodological issues in the field. *Royal Society Open Science*, 9(12), Article 220946. <https://doi.org/10.1098/rsos.220946>
- McCrum, C., van Beek, J., Schumacher, C., Janssen, S., & Van Hooren, B. (2022). Sample size justifications in Gait & Posture. *Gait & Posture*, 92, 333-337. <https://doi.org/10.1016/j.gaitpost.2021.12.010>
- Milton, M. J., & Possolo, A. (2020). Trustworthy data underpin reproducible research. *Nature Physics*, 16(2), 117-119. <https://doi.org/10.1038/s41567-019-0780-5>
- Mullard, A. (2021). Half of top cancer studies fail high-profile reproducibility effort. *Nature*, 600(7889), 368-369. <https://doi.org/10.1038/d41586-021-03691-0>
- Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., du Sert, N. P., Simonsohn, U., Wagenmakers, E. J., Ware, J. J., & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour*, 1, Article 0021. <https://doi.org/10.1038/s41562-016-0021>
- National Academies of Sciences, Engineering, and Medicine (2019). *Reproducibility and Replicability in Science*. The National Academies Press. <https://doi.org/10.17226/25303>
- National Human Genome Research Institute (2023). *Candidate gene*. <https://www.genome.gov/genetics-glossary/Candidate-Gene>
- Nickerson, R. S. (2000). Null hypothesis significance testing: A review of an old and continuing controversy. *Psychological Methods*, 5(2), 241-301. <https://doi.org/10.1037/1082-989X.5.2.241>
- Nosek, B. A., & Errington, T. M. (2020). What is replication? *PLoS Biology*, 18(3), Article e3000691. <https://doi.org/10.1371/journal.pbio.3000691>
- Nuijten, M. B., Hartgerink, C. H., Van Assen, M. A., Epskamp, S., & Wicherts, J. M. (2016). The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, 48(4), 1205-1226. <https://doi.org/10.3758/s13428-015-0664-2>
- Nuzzo, R. (2014). Statistical errors: P values, the „gold standard“ of statistical validity, are not as reliable as many scientists assume. *Nature*, 506(7487), 150-152. <https://doi.org/10.1038/506150a>
- Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, 349(6251), Article aac4716. <https://doi.org/10.1126/science.aac4716>
- Plomin, R. (2020). *Kód života: Jak z nás DNA dělá to, kým jsme*. Jan Melvil Publishing.
- Popper, K. R. (1997). *Logika vědeckého bádání*. OIKOYMENH.
- Prinz, F., Schlange, T., & Asadullah, K. (2011). Believe it or not: How much can we rely on published data on potential drug targets? *Nature Reviews Drug Discovery*, 10(9), Article 712. <https://doi.org/10.1038/nrd3439-c1>
- Purtill, R. (1970). The purpose of science. *Philosophy of Science*, 37(2), 301-306. <https://doi.org/10.1086/288303>
- Ranchill, E., Dreber, A., Johannesson, M., Leiberg, S., Sul, S., & Weber, R. (2015). Assessing the robustness of power posing: No effect on hormones and risk tolerance in a large sample of men and women. *Psychological Science*, 26(5), 653-656. <https://doi.org/10.1177/0956797614553>
- Redish, A. D., Kummerfeld, E., Morris, R. L., & Love, A. C. (2018). Reproducibility failures are essential to scientific inquiry. *Proceedings of the National Academy of Sciences*, 115(20), 5042-5046. <https://doi.org/10.1073/pnas.1806370115>
- Ritchie, S. (2021). *Science fictions: Exposing fraud, bias, negligence and hype in science*. Vintage.
- Rodgers, P., & Collings, A. (2021) Reproducibility in Cancer Biology: What have we learned? *eLife*, 10, Article e75830. <https://doi.org/10.7554/eLife.75830>
- Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin*, 86(3), 638-641.
- Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist*, 44(10), 1276-1284.
- Rudner, R. (1961). An introduction to simplicity. *Philosophy of Science*, 28(2), 109-119.

- Sainani, K. L., Borg, D. N., Caldwell, A. R., Butson, M. L., Tenan, M. S., Vickers, A. J., Vigotsky, A. D., Warmenhoven, J., Nguyen, R., Lohse, K. R., Knight, E., & Bargary, N. (2021). Call to increase statistical collaboration in sports science, sport and exercise medicine and sports physiotherapy. *British Journal of Sports Medicine*, 55(2), 118-122. <https://doi.org/10.1136/bjsports-2020-102607>
- Sainani, K., & Chamari, K. (2022). Wish list for improving the quality of statistics in sport science. *International Journal of Sports Physiology and Performance*, 17(5), 673-674. <https://doi.org/10.1123/ijsp.2022-0023>
- Schimmack, U. (2019). *Ioannidis (2005) was wrong: Most published research findings are not false*. Replicability-Index. <https://web.archive.org/web/20200919152240/https://replicationindex.com/2019/01/15/ioannidis-2005-was-wrong-most-published-research-findings-are-not-false/>
- Schlick, M. (1936). Meaning and verification. *The Philosophical Review*, 45(4), 339-369. <https://doi.org/10.2307/2180487>
- 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(2), 90-100. <https://doi.org/10.1037/a0015108>
- Schwab, A., Abrahamson, E., Starbuck, W. H., & Fidler, F. (2011). Perspective researchers should make thoughtful assessments instead of null-hypothesis significance tests. *Organization Science*, 22(4), 1105-1120.
- Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological Science*, 29(4), 549-571. <https://doi.org/10.1177/0956797617739704>
- 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. <https://doi.org/10.1177/0956797611417632>
- Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, 3(9), Article 160384. <https://doi.org/10.1098/rsos.160384>
- Smith, K. M., & Apicella, C. L. (2017). Winners, losers, and posers: The effect of power poses on testosterone and risk-taking following competition. *Hormones and Behavior*, 92, 172-181. <https://doi.org/10.1016/j.yhbeh.2016.11.003>
- Stapel, D. A., & Lindenberg, S. (2011). Coping with chaos: How disordered contexts promote stereotyping and discrimination. *Science*, 332(6026), 251-253. <https://doi.org/10.1126/science.1201068> (Retraction published Stapel, D. A., & Lindenberg, S. (2011, December 2). *Science*, 334(6060), 1202. <https://doi.org/10.1126/science.334.6060.1202-a>)
- Steneck N. H. (2006). Fostering integrity in research: Definitions, current knowledge, and future directions. *Science and Engineering Ethics*, 12(1), 53-74. <https://doi.org/10.1007/pl00022268>
- Stroebe, W. (2019). What can we learn from many labs replications? *Basic and Applied Social Psychology*, 41(2), 91-103. <https://doi.org/10.1080/01973533.2019.1577736>
- Świątkowski, W., & Carrier, A. (2020). There is nothing magical about Bayesian statistics: An introduction to epistemic probabilities in data analysis for psychology starters. *Basic and Applied Social Psychology*, 42(6), 387-412. <https://doi.org/10.1080/01973533.2020.1792297>
- The Colbert Report. (2011). *Time travelling porn – Daryl Bem*. <http://www.cc.com/videoclips/bhf8jv/the-colbert-report-time-traveling-porn-daryl-bem>
- Twomey, R., Yingling, V., Warne, J., Schneider, C., McCrum, C., Atkins, W., Murphy, J., Medina, C. R., Harlley, S., & Caldwell, A. (2021). The nature of our literature: A registered report on the positive result rate and reporting practices in kinesiology. *Communications in Kinesiology*, 1(3), 1-17. <https://doi.org/10.51224/cik.v1i3.43>
- Wasserstein, R. L., & Lazar, N. A. (2016). The ASA statement on p-values: Context, process, and purpose. *The American Statistician*, 70(2), 129-133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wasserstein, R. L., Schirm, A. L., & Lazar, N. A. (2019). Moving to a world beyond “p < 0.05”. *The American Statistician*, 73(Suppl. 1), 1-19. <https://doi.org/10.1080/00031305.2019.1583913>
- Wilson, E. O. (1999). *Consilience: The unity of knowledge*. Vintage Books.
- Wong, T. K., Kiers, H., & Tendeiro, J. (2022). On the potential mismatch between the function of the bayes factor and researchers' expectations. *Collabra: Psychology*, 8(1), Article 36357. <https://doi.org/10.1525/collabra.36357>
- Xie, Y., Wang, K., & Kong, Y. (2021). Prevalence of research misconduct and questionable research practices: A systematic review and meta-analysis. *Science and Engineering Ethics*, 27(4), 1-28. <https://doi.org/10.1007/s11948-021-00314-9>
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice, and lives*. University of Michigan Press.
- Zwaan, R. A., Pecher, D., Paolacci, G., Bouwmeester, S., Verkoeijen, P., Dijkstra, K., & Zeelenberg, R. (2018). Participant nonnaiveté and the reproducibility of cognitive psychology. *Psychonomic Bulletin & Review*, 25(5), 1968-1972. <https://doi.org/10.3758/s13423-017-1348-y>

The replication crisis in science and its overlap into kinanthropology: an introduction

Background: Over the past decade, the global scientific community (particularly in social science and biomedical research) has been shaken by the inability to replicate the results of previously conducted scientific studies. This phenomenon has come to be known as the replication crisis. **Objective:** The aim of this essay is to introduce the reader to the so-called replication crisis in science and to describe its possible causes. In the first part, we describe the importance of replication in the emergence of scientific knowledge, discuss the basic terminology of the phenomenon under study, and present the most well-known examples of failed replications in psychological and biomedical research. In the next section, we will look at possible causes of the replication crisis, such as questionable research practices in data analysis with a focus on the misusing of p-values. Finally, we conclude with challenges that highlight the replication crisis in kinanthropology. **Conclusions:** Despite the fact that we do not currently have direct evidence for the existence of a replication crisis in kinanthropology, there are doubts about the scientific practice of applying inferential statistics, particularly in relation to the use of p-values. For these reasons, it is important to monitor the issue and initiate a broad discussion. Our aim is not to give a detailed overview of the topic, but to introduce the reader to the basic ideas and resources in the field of replication crisis, so that he/she will be able to find more relevant information if interested.

Keywords: cognitive bias, HARKing, p-value, p-hacking, questionable research practices, hypothesis testing