JUSSI VALTONEN & NELLI HANKONEN

Mitä toistettavuuskriisistä pitäisi ajatella? Johdanto tutkijoiden haastatteluihin

”Eikö olekin coolein torakkatutkimus, jonka olette ikinä nähneet?” Näin kaksi New Yorkin yliopiston (NYU:n) professoria esitteli viime vuonna tutkimusartikkelin, joka kaikkien sosiaalipsykologian jatko-opintoseminaarilaisten tuli lukea.

Zajoncin, Heingartnerin ja Hermanin (1969)[1] klassikko ”Social enhancement and impairment of performance in the cockroach” saattaa olla maailman tunnetuin sosiaalipsykologian alaan kuuluva torakkatutkimus. NYU:n tutkijakoulutettavia ei vaadittu perehtymään siihen siksi, että New Yorkiin muuttanut joutuu torakoiden kanssa tekemisiin kylpyhuoneessaan lähes varmasti, vaan siksi, että Zajoncin (1965) viettiteorian on ajateltu kertovan jotakin olennaista myös ihmisestä. Tärkeä osa teoriaa on sosiaalisena helpontumisena (*social facilitation*) tunnettu ilmiö, jonka mukaan yksilön suoritus paranee hyvin hallituissa tilanteissa, jos paikalla on muita, vaikka vaikeissa tehtävissä muiden läsnäolo heikentää suoritusta. Tutkijakoulutettavat pantiin pohtimaan torakkakokeen tuloksia Gordon Allportin, Solomon Aschin ja Leon Festingerin klassikkotutkimusten ohella siksi, että Zajoncin ja kollegojen (1969) tutkimusta on puolen vuosisadan ajan siteerattu laajasti oppikirjoissa ja katsausartikkeleissa, ja se on inspiroinut valtavan määrän jatkotutkimusta.

”Ei sekin!” yksi jatko-opiskelijoista parkaisi torakkakokeesta keväällä. Zajoncin klassikosta oli helmikuussa julkaistu esirekisteröity toistoyritys, jossa päätulosta ei onnistuttu toistamaan. Halfmann, Bredehöft ja Häusser (2020) toistivat kokeen kolme kertaa alkuperäistä suuremmalla koe-eläinjoukolla, ja ratkaiseva yhdysvaikutus, klassiseksi luultu sosiaalisen helpontumisen efekti, jäi toistumatta.

NYU:n väitöskirjantekijä ei kuitenkaan parahtanut angstista ainoastaan siksi, että kyseenalaistetuksi tuli kenties sosiaalipsykologian historian tunnetuin torakkakoe. Saman kohtalon on kuluneen vuosikymmenen aikana kokenut lukuisa joukko myös muita psykologisia tutkimustuloksia.

Eikä ainoastaan psykologiassa: eri tieteenalojen tutkijoista jopa 90 prosenttia oli *Nature*-lehden kyselyssä sitä mieltä, että tieteessä on meneillään toistettavuus- eli replikaatiokriisi (*replication crisis*; Baker, 2016). Keskustelua tutkimusten toistettavuuteen ja luotettavuuteen liittyvistä kysymyksistä on käyty viime vuosina aktiivisesti esimerkiksi biolääketieteissä (Begley & Ellis, 2012; Glasziou ym., 2014; Prinz, Schlange & Asadullah, 2011) ja myös muilla aloilla (esim. Feilden, 2017). Vaikka monet esiin nousseista kysymyksistä ovat monille empiirisille tieteenaloille yhteisiä (ks. esim. Arguello, 2019), osa tutkijoista arvelee psykologian olleen erityisen altis kritiikille (Gelman, 2016b).

Tämä artikkeli ja haastattelukooste syntyivät havainnosta, jonka mukaan Suomessa ei tieteellisissä lehdissä juuri ole julkaistu pohdiskelua psykologian replikaatiokriisistä, vaikka kansainväliset tieteelliset lehdet ovat omistaneet aiheelle useita katsauksia (esim. Nelson, Simmons & Simonsohn, 2018; Shrout & Rodgers, 2018) ja erikoisnumeroita (Chambers, 2019; Nosek & Lakens, 2014; Pashler & Wagenmakers, 2012) sekä julkaisseet painokkaita puheenvuoroja siitä, kuinka tutkimus-, julkaisu-, rahoitus- ja urakehityskäytäntöjä tulisi muuttaa (esim. Benjamin ym., 2018; Inzlicht, 2019; Nosek ym., 2015, 2019; Nosek & Lakens, 2014; Spellman, 2015; Vazire, 2018). Toistettavuuskriisinä tunnettu keskustelu

tunnetaan myös tieteenfilosofien parissa (Fidler & Wilcox, 2018).

Tässä *Psykologia*-lehden haastattelukokonaisuudessa kysytään, mitä keskustelua seuranneet kotimaiset psykologitutkijat tilanteesta ajattelevat. Taustoitamme haastatteluja lyhyellä yhteenvedolla tapahtumista, jotka olivat vaikuttamassa kansainvälisen keskustelun alkamiseen, ja sen jälkeen annamme puheenvuoron kymmenelle kotimaiselle tutkijalle.

TOISTETTAVUUSKRIISIN LYHYT HISTORIA

Vaatimusta tulosten toistettavuudesta pidetään kvantitatiivisia menetelmiä käyttävillä empiirisillä tieteenaloilla yhtenä tieteellisen tiedon kriteereistä. Tieteenfilosofi Karl Popperin (1959) mukaan tieteellisesti osoitettu ilmiö on sellainen, ”jonka pystyy tuottamaan säännönmukaisesti uudelleen kuka tahansa, joka tekee asianmukaisen kokeen määritellyllä tavalla” (s. 23–24). Kuten Halfmann ja kollegat (2020) toteavat, on epäselvää, kuinka hyödyllinen esimerkiksi Zajoncin (1965) sosiaalisen helpontumisen teoria on, jos tutkimustuloksia, joihin teoria perustuu, osoittautuu mahdottomaksi toistaa.

Mahdollisuudesta, jonka mukaan psykologeilla saattaisi olla tieteenalan yleisissä käytännöissä korjaamisen varaa, alettiin keskustella laajasti vajaa vuosikymmen sitten. Niihin aikoihin sijoittui monta toisistaan näennäisesti riippumatonta tapahtumaa, jotka yhdessä käynnistivät tutkimusmenetelmiä koskevan keskustelun, jota on käyty psykologiassa ja naapurialoilla siitä saakka.

Tieteellistä parapsykologiaa?

Yksi keskustelun käynnistäjistä oli sosiaalipsykologi Daryl Bemin yhdeksän parapsykologisen kokeen sarja, joka julkaistiin *Journal of Personality and Social Psychology* -sarjassa 2011. Bemin (2011) tulokset näyttivät monen mielestä käsittämättömiltä. Tulosten mukaan tulevat tapahtumat, joista koehenkilöt eivät voineet tietää, olisivat vaikuttaneet heidän käyttäytymiseensä. Arvostetussa sarjassa julkaistu koesarja herätti monet niin alan sisällä kuin sen ulkopuolellakin pohtimaan, miten oli mahdollista, että pinnallisesti oikeaoppisen näköinen koesarja, jossa ei ollut yhtä päivänselvää teknistä virhettä, saattoi päätyä niin merkillisiin tuloksiin (esim. Yarkoni, 2011).

Bemin (2011) raflaavat tulokset eivät toistuneet muiden laboratorioissa (Galak, LeBoeuf, Nelson & Simmons, 2012), mikä oli monen mielestä odotettavaa. Tätä hälyttävämpänä osa tutkijoista piti sitä, mitä koesarja tuli paljastaneeksi tutkimusten toistamisesta yleisemmällä tasolla. Itse toistokokeiden tuloksia osoittautui nimittäin huomattavasti vaikeammaksi julkaista kuin niiden tekijät olivat odottaneet: niin *JPSP* kuin usea muukin psykologian alan julkaisusarja kieltäytyi julkaisemasta toistoyritysten tuloksia (French, 2012).

Niin tärkeänä kuin koetulosten toistamista teorian tasolla pidettiinkin, harva oli tajunnut, kuinka epätavallista kokeiden toistotulosten julkaiseminen akateemisessa psykologiassa todellisuudessa oli (Makel, Plucker & Hegarty, 2012). Jopa räikeän virheelliset tutkimustulokset näyttivät liian helposti jäävän julkaistuun tutkimuskirjallisuuteen kummittelemaan vielä senkin jälkeen, kun ne oli toistamalla todettu harhaanjohtaviksi, kun taas luotettavamman, suuremmalla otoskoolla tehdyn toistokokeen tuloksia ei näyttänyt saavan muiden tutkijoiden saataville millään.

Kysymyksiä herää myös muualla

Samoihin aikoihin ilmeni, että myös valtavirran psykologinen tutkimuskirjallisuus saattoi sisältää selvästi enemmän ongelmallisia tuloksia kuin oli yleisesti ajateltu. Barghin, Chenin ja Burrowsin (1996) vaikutusvaltainen sosiaalista viritystä (*social priming*) käsittelevä tutkimus, jota moni oli tottunut opettamaan sosiaalipsykologian johdantokursseilla, asettui yhtenä ensimmäisistä kyseenalaiseksi, kun Doyen, Klein, Pichon ja Cleeremans (2012) epäonnistuivat yrityksessään toistaa sen tulokset. Alkuperäisten tulosten mukaan vanhuuteen liittyvät ärsykkeet olisivat saaneet koehenkilöt kävelemään tavallista hitaammin, mutta laajasti siteerattu ilmiö hävisi kaksi kertaa suuremmalla otoskoolla tehdyssä toistotutkimuksessa kokonaan.

Usean samankaltaisen toistoyrityksen epäonnistuminen johti siihen, että koko sosiaalisena virityksenä tunnettu ilmiöjoukon totuuspohja kyseenalaistettiin, mikä herätti laajasti huomiota myös psykologipiirien ulkopuolella (esim. Bower, 2012; Yong, 2012).

Huijareita, kyseenalaisia käytäntöjä ja luotettavuuskriisi

Kolmas erityisesti sosiaalipsykologien tutkijayhteisöä hätkähdyttänyt tapahtuma oli hollantilaisen Diederik Stapelin jääminen kiinni laajasta tutkimusvilpistä, johon hän oli syyllistynyt vuosikymmenten ajan kenenkään huomaamatta. Stapelin tapaus herätti huomiota, koska hänen sepitettyihin aineistoihin perustuvia tutkimustuloksiaan oli julkaistu kymmenissä artikkeleissa, myös kaikkein vaikutusvaltaisimmissa julkaisusarjoissa. Huijauksen ohella monien huomio kiinnittyi siihen, että tutkimusmenetelmät näyttivät ongelmallisilta myös monissa niistä Stapelin julkaisuista, joissa itse aineistoa ei ollut sepitetty. Epäilyksiä heräsi pian myös muiden tutkijoiden töistä, joita niitäkin arveltiin väärennetyiksi (Simonsohn, 2013).

Samoihin aikoihin julkaistiin useita havainnollisia esimerkkejä siitä, kuinka helppoa täysin rehellisten tutkijoiden on päätyä vääriin positiivisiin havaintoihin, vaikka he eivät tietoisesti tekisi mitään vilpillistä. Simmons, Nelson ja Simonsohn (2011), Vul, Harris, Winkielman ja Pashler (2009), Sullivan (2007) ja Ioannidis (2005) havainnollistivat kukin useita pinnallisesti viattoman oloisia ratkaisuja, joilla tutkija saattoi saada I tyypin virheen[2] mahdollisuuden paisumaan räikeän suureksi aineistoa kerätessään ja analysoidessaan. Ongelmia esiintyi niin psykologiassa, aivotutkimuksessa kuin biolääketieteissäkin. Kyse ei myöskään ollut marginaalisesta ongelmasta: suuri osa psykologeista myönsi kysyttäessä syyllistyvänsä kyseenalaisiin tutkimuskäytäntöihin (*questionable research practices*, QRP) ainakin toisinaan (John, Loewenstein & Prelec, 2012). Johnin ja kollegojen (2012) tutkimuksen viisi yleisimmältä näyttänyttä kyseenalaista käytäntöä olivat 1) kaikkien riippuvien muuttujien raportoimatta jättäminen, 2) lisäaineiston kerääminen sen jälkeen, kun analyysejä on jo tehty, 3) vain niiden tutkimusten valikoiva raportoiminen, jotka ”toimivat”, 4) aineiston käsittely, esimerkiksi äärihavaintojen poistaminen ratkaisujen seurausten tarkastelun jälkeen ja 5) odottamattoman tuloksen väittäminen ennustetuksi.

Kaikki tämä viittasi siihen, että ongelmallisia tutkimustuloksia tuli julkaistuksi huomattavasti useammin kuin moni oli tiedostanut. ”Tajusimme, että kokonaiset tutkimuskirjallisuudet saattoivat koostua vääristä positiivisista tuloksista”, Joseph Simmons kiteytti tilanteen *New York Times Magazinelle* (Dominus, 2017). Yhteen parapsykologiseen koesarjaan liittynyt kohu olikin laajentunut koskemaan koko alaa: ”It was Bem, perhaps, who kicked us all into the realization that bad work could be the rule, not the exception”, tilastotieteilijä Andrew Gelman (2016a) tiivisti tilanteen blogissaan. Yhdessä toistettavuusongelmien ja huijaustapausten kanssa vyyhti herätti laajimmillaan sen huolestuttavan kysymyksen, oliko tiede lainkaan niin itseään korjaavaa kuin oli ajateltu, ainakaan käytännössä (Inzlicht, 2019; Ioannidis, 2012).

Ongelmien ruotimisen ohella alettiin nopeasti keskustella siitä, missä määrin tutkijoiden kannustimet ja tiedeyhteisön omat sosiaaliset normit, kuten paine julkaista mahdollisimman yllättäviä tuloksia mahdollisimman nopeasti, vinouttivat kirjallisuutta väärien positiivisten havaintojen ja muiden ongelmien suuntaan (esim. Bakker, van Dijk & Wicherts, 2012; Gelman, 2011; Giner-Sorolla, 2012). Samalla heräsi aktiivinen keskustelu myös siitä, mitä eri ongelmille olisi mahdollista tehdä (Nosek, Spies & Motyl, 2012; Open Science Collaboration, 2012; Pashler & Wagenmakers, 2012; Spellman, 2012).

Psykologisten tutkimusten toistettavuusprojekti

Kuinka usein kvantitatiiviset tutkimustulokset sitten jäävät psykologiassa toistumatta? Ongelman mittakaavasta saatiin konkreettinen arvio Brian Nosekin johtamassa toistettavuusprojektissa (Open Science Collaboration, 2015). Projektissa yritettiin toistaa sata psykologian alan tutkimustulosta, jotka oli julkaistu kolmessa arvostetussa lehdessä.

Kun kokeet toistettiin, tulokset olivat tilastollisesti merkitseviä enää vain 36 prosentissa tutkimuksista, ja alkuperäiset efektikoot puolittuivat. Aloilla oli myös keskinäisiä eroja: tulokset toistuivat kognitiivisessa psykologiassa keskimäärin paremmin kuin sosiaalipsykologiassa.

Joillekin psykologian aloille kuten kognitiiviseen neurotieteeseen keskustelu ei jostakin syystä ole vielä rantautunut, ainakaan samalla painokkuudella kuin sosiaalipsykologiaan (Chambers, 2019). Osa havainnoista viittaa kuitenkin siihen, että monet samoista kysymyksistä koskevat myös aivotutkimusta (Button ym., 2013; Carp, 2012). Jotkut ovat argumentoineet, että ainakin osa ongelmista olisi kognitiivisessa neurotieteessä jopa muita psykologian osa-alueita suurempia (Szucs & Ioannidis, 2017). Korkeaimpaktisten julkaisusarjojen vaikuttimet eivät ainakaan vaikuttaisi kognitiivisessa neurotieteessä tieteen edistymisen kannalta sen puhtaammilta kuin muuallakaan (Arguello, 2019; Huber, Potter & Huszar, 2019).

Vanhoja ongelmia, tuoreita kiukunpurkauksia

Monet ongelmista olivat toki olleet tiedossa jo pitkään ennen toistettavuuskriisinä tunnettua tuoretta keskustelua (Rodgers & Shrout, 2017). Osa tutkijoista oli varoitellut jo aiemmin ja toistuvasti, että pienillä otoksilla tehtyjen kokeiden tuloksiin ei välttämättä kannattaisi luottaa, varsinkaan jos ilmiöiden efektikokoja ei tiedetä (Cohen, 1962, 1990; Rosnow & Rosenthal, 1989). Samoin on ollut jo pitkään tiedossa, että huolestuttavan suuri osa tutkimustuloksista jää tutkijoiden pöytälaatikkoon (Rosenthal, 1979).

Moni tiedeyhteisössä on silti avoimesti myöntänyt, että ongelmien syvyys tuli heille yllätyksenä (esim. Gelman, 2016a; Lindsay, 2016; Valtonen, 2019). Empiirisen tieteellisen kirjallisuuden oli ajateltu yleisesti olevan selvästi luotettavampaa. Gelman (2016a) kirjoittaa blogissaan: ”So, as of early 2011, there’s a sense that something’s wrong, but it’s not so clear to people *how* wrong things are, and observers (myself included) remain unaware of the ubiquity, indeed the obviousness, of fatal multiple comparisons problems in so much published research. Or, I should say, the deadly combination of weak theory being supported almost entirely by statistically significant results which themselves are the product of uncontrolled researcher degrees of freedom.”

Mutta subjektiivinen muutos tapahtui Gelmanin (2016a) mukaan nopeasti: ”Late 2016: We have now reached the ’emperor has no clothes’ phase. When seemingly solid findings in social psychology turn out not to replicate, we’re no longer surprised.”

Kaikki eivät ole olleet ainoastaan ilahtuneita toistettavuuskriisiä koskevasta keskustelusta. Sekä kriisi-sana että keskusteluun liittyvä uudistusliike ovat herättäneet myös vastalauseita, ja julkinen keskustelu on paikoin ollut tutkijoiden välillä varsin henkilökohtaista (esim. Letzter, 2016). Osa tutkijoista ajattelee, että ongelmia on liioiteltu tai tilanne on tulkittu muuten väärin (Fanelli, 2018; Stroebe & Strack, 2014). Huolta on herättänyt myös se, miten valtamedian suurelle yleisölle suunnattu otsikointi tieteestä, joka ”on rikki” (esim. Engber, 2017), vaikuttaa alan julkiseen kuvaan. Toisten mukaan taas luottamus alaa kohtaan ei koskaan palaudukaan, jos ongelmista ei keskustella avoimesti eikä tarvittavia muutoksia tehdä (esim. Gelman, 2016a; Pashler & de Ruiter, 2017; Shea, 2011; Valtonen, 2019).

Ongelmista ratkaisuihin

Sen lisäksi, että osa tutkijoista on surrut paljastuneiden ongelmien mittakaavaa julkisesti ja henkilökohtaisesti (esim. Inzlicht, 2016), käyty keskustelu on selvästi vaikuttanut alaan lyhyessä ajassa jo monin tavoin.

Sen lisäksi, että kansainvälisten konferenssien luentosalit ovat pullistelleet aihetta käsitelleissä tilaisuuksissa, osa tutkijoista, vertaisarvioijista, lehdistä ja rahoittajista on jo muuttanut toimintatapojaan myös käytännössä (esim. Lindsay, 2016; Nelson, Simmons & Simonsohn, 2018). Uusista käytännöistä erityisesti tutkimusten esirekisteröinti (*preregistration*; Nosek ym., 2019) ja rekisteröidyt raportit (*registered reports*; Nosek & Lakens, 2014) ovat saaneet kansainvälisessä keskustelussa kannatusta, ja esimerkiksi yhä useammat lehdet vaativat nyt, että tutkimusaineistot julkaistaan avoimesti. Moni kansainvälisessä tiedeyhteisössä tuntuukin katsovan, että keskustelu on tehnyt alalle kaiken kaikkiaan

hyvää (Aschwanden, 2018; Rodgers & Shrout, 2017; Vazire, 2018).

Rekisteröidyt raportit näyttäisivät tuovan tasapainoa tutkimustulosten valikoivaan julkaisemiseen: tuore ennakkojulkaisu (Scheel, Schijen & Lakens, 2020) osoittaa, että rekisteröidyissä raporteissa enää 44 prosenttia tutkimuksista tuki koeteltuja hypoteeseja, kun tätä ennen luku on psykologian alan julkaisuissa ollut 96 prosenttia. Kirjoittajat päättelevät, että alan tavalliset käytännöt ovat johtaneet negatiivisten tulosten aliraportointiin tavalla, joka vaarantaa tieteellisten tulosten luotettavan kasautumisen.

Monet ovat todenneet, että replikaatiokriisin ydin ei näyttäisi olevan niinkään tilastotieteessä – vaikka myös tilastotiedettä käytetään ajoittain väärin – vaan tavassa, jolla näyttöön on laajemmin suhtauduttu. Amrhein, Trafimow ja Greenland (2019) toteavat sen unohtuvan helposti, että tilastolliset testit eivät testaa pelkästään yksittäisiä hypoteeseja, vaan myös niiden taustaoletuksia ja koko tutkimuksen kontekstia. Tieteenfilosofit ovat muistuttaneet jo pitkään, että mitään tieteellistä hypoteesia ei voi testata eristettynä sen taustaoletuksista (Duhem–Quine-teesi; ks. Stanford, 2017). Kaikkia lisäoletuksia ei ole

TAULUKKO 1. Munafòn ja kumppanien (2017) manifesti toistettavalle tieteelle, mukaillen suomennettu.

Teema	Ehdotus	Esimerkkejä ratkaisuista
Menetelmät	Kognitiivisilta vääristymiltä suojautuminen	Kaikki ratkaisut alla
		Sokkoutus
	Metodologisen koulutuksen parantaminen	Tilastotieteen ja tutkimusmenetelmien koulutus tuleville tutkijoille ja jatkuva oppiminen tutkijoille
	Riippumaton metodologinen tuki	Metodologien mukaan ottaminen tutkimukseen
		Riippumaton valvonta
	Yhteistyö ja tiimitiede	Monen tutkimusryhmän yhteistutkimus/ aineistonkeruu
		Tiimitiedekonsortiot
Raportointi ja levittäminen	Tutkimusten esirekisteröinti	Rekisteröidyt raportit
		Open Science Framework
	Tutkimusraporttien laadun parantaminen	Tarkistuslistojen käyttö raportoinnissa
		Protokolla-tarkistuslistat
	Eturistiriidoilta suojautuminen	Eturistiriitojen paljastaminen
Uusinnettavuus	Läpinäkyvyyden ja avoimen tieteen edistäminen	Avoin data, materiaalit, ohjelmistot jne.
		Esirekisteröinti
Arviointi	Vertaisarvioinnin monipuolistaminen	Esijulkaisukäytäntö (pre-print)
		Vertaisarviointi etu- ja jälkikäteen
Kannustimet	Avoimista ja uusinnettavista käytännöistä palkitseminen	Ansiomerkit (OSF Badges)
		Rekisteröidyt raportit

mahdollista avata tutkimuksissa edes teoriassa, saati käytännössä.

Myös tilastolliseen päättelyyn liittyy monenlaisia oletuksia, minkä vuoksi se tulisikin Amrheinin ja kollegojen (2019) mukaan mieltää oletusten ja tutkimusaineiston välisten suhteiden epävakaiksi, paikallisiksi kuvauksiksi sen sijaan, että tilastollisen päättelyn ajateltaisiin tarjoavan yleistettäviä johtopäätöksiä. Amrhein ja kollegat (2019) suosittelevat, että tilastollisia tuloksia kohdeltaisiin selvästi epätäydellisempinä ja epävarmempina kuin nykyisin on tapana. Epävarmuuden hyväksyminen voisi vähentää myös houkutusta valikoivaan raportointiin: koska *p*-arvojen koko vaihtelee joka tapauksessa, tutkimuksia ei ole perusteltua valikoida niiden perusteella. He suosittelevat, että tutkimusraporteissa keskityttäisiin epävarmojen johtopäätösten sijasta kuvailemaan tarkasti tutkimusmenetelmät, aineisto, analyysimenetelmät ja niiden perusteet, ja että tutkimuksia arvioitaisiin aineistojen ja menetelmien laadun eikä tulosten tai johtopäätösten perusteella.

Psykologian professori Marcus Munafò ja kollegat julkaisivat vuonna 2017 manifestin uusinnettavan tieteen puolesta *Nature Human Behaviour* -sarjassa. Taulukossa 1 on heidän ehdotuksensa tieteellisten käytäntöjen parantamiseksi luokiteltuna teemojen mukaan ja esimerkkejä konkreettisista ratkaisuista (Munafò ym., 2017).

Osa käytännön kysymyksistä on kuitenkin vielä osin auki (ks. esim. Hardwicke & Ioannidis, 2018), ja myös omat epäviralliset huomiomme viittaavat siihen, että eri tutkimusympäristöissä ja -laboratorioissa toimitaan edelleen varsin vaihtelevin tavoin.

KESKUSTELU AKTIIVISEMMIN VIREILLE MYÖS SUOMESSA?

Tämän haastattelun tarkoitus on koota yhteen eri psykologian ja sosiaalipsykologian alojen tutkijoiden näkemyksiä tutkimustulosten toistettavuudesta ja edistää tulevien kotimaisten suuntaviivojen määrittelyä. Koska alan uudet normit ovat vasta vakiintumassa myös kansainvälisesti, olemme kutsuneet haastateltavaksi suomalaisia tutkijoita, jotka ovat aktiivisesti osallistuneet keskusteluun toistettavuuskriisistä eri foorumeilla, Suomessa tai ulkomailla. Tavoitteena ei ollut luoda kattavaa katsausta eikä edustavaa otosta suomalaisista psykologian alan tutkijoista vaan käynnistää keskustelua. Kirjoittajat ovat vastanneet kysymyksiin kirjallisesti, eivätkä he ole kirjoittaessaan nähneet toistensa vastauksia.

Vastauksista näkee, että yleisesti ottaen haastatellut suhtautuvat tilanteeseen vakavasti ja huolestuneina. Moni heistä oli paitsi törmännyt ongelmiin tekaistuja aineistoja koskevien kohujen vuoksi, myös huomannut kysymysten vaikuttavan suoraan omaan tutkimukseensa. Moni onkin seurannut kansainvälistä keskustelua tiiviisti.

Haastatteluvastauksissa nostetaan esiin monia myös edellä mainittuja kyseenalaisia tutkimuskäytäntöjä – *p*-hakkerointi (*p-hacking*, aineiston käsitteleminen tavalla, joka lisää väärien positiivisten havaintojen mahdollisuutta; Simmons ym., 2011), HARKing (*hypothesising after results are known*, hypoteesien luominen tulosten jo ollessa tiedossa; Bones, 2012; Kerr, 1998), vähäiseen tilastolliseen voimaan tyytyminen ja julkaisuvinouma. Haastateltavat mainitsevat myös monia Munafòn ja kollegojen (2017) mainitsemista ratkaisuehdotuksista: haastattelemamme kotimaiset tutkijat kannattavat hekin rekisteröityjä raportteja ja avoimen tieteen pelisääntöjä kuten aineistojen ja analyysien avointa julkaisemista. Munafòn ja kollegojen (2017) esitykseen verrattuna vähemmän mainintoja puolestaan saavat tiimitiede, tarkistuslistojen hyödyntäminen laadun parantamiseksi ja esijulkaisukäytännöt. Lukija löytää haastatteluvastauksista myös hyödyllisiä lähdeviitteitä tarkempaa perehtymistä varten.

Jos toiveemme toteutuu, tämä kokonaisuus herättää ajatuksia myös *Psykologia*-lehden lukijoissa. Ovatko vetoomukset avoimen tieteen puolesta muuttaneet toimintatapoja Suomessa? Millaisia muutoksia tutkijat ja opettajat ovat Suomessa tehneet omiin käytäntöihinsä? Olisiko kotimaassa toiveita tehdä yhteistyötä eri yliopistojen välillä menetelmäkoulutuksen kehittämiseksi psykologian alalla? *Psykologia*-lehti olisi mielestämme mitä sopivin tämän keskustelun kotimaiseksi foorumiksi myös jatkossa.

Viitteet

[1] Johdannon ja haastattelun yhteinen lähdeluettelo on haastatteluosuuden perässä sivuilla 455–458.

[2] I tyypin virheellä tarkoitetaan hypoteesintestauksessa sitä, että nollahypoteesi hylätään, vaikka se on tosi (esim. väitetään, että ryhmien keskiarvoilla on eroa, vaikka todellisuudessa eroa ei ole). Kun minun pitää arvioida, onko kadulla vastaan tuleva kalpea, kaapuun pukeutunut hahmo vampyyri vai muuten vain kalpea ja kaapuasuinen, teen I tyypin virheen, jos isken seipään hänen sydämeensä eikä hän olekaan vampyyri. (Daniël Lakensin esimerkki.)

Kiitokset

Kiitokset Minttu Palsolalle avusta artikkelin viimeistelyssä ja kaikille haastatelluille.

Mitä toistettavuuskriisistä pitäisi ajatella? Kymmenen suomalaistutkijaa sähköpostihaastattelussa

HAASTATELLUT TUTKIJAT

Matti Heino, VTM, BBA. Tutkimusala: motivaatio, käyttäytymislääketiede, kompleksiset järjestelmät.

Ville-Juhani Ilmarinen, PsT. Tutkimusala: persoonallisuus- ja sosiaalipsykologia.

Sointu Leikas, PsT. Tutkimusala: sosiaali- ja persoonallisuuspsykologia.

Jari Lipsanen, PsM. Erikoisala: psykologian arviointimenetelmät.

Jan-Erik Lönnqvist, PsT. Tutkimusala: sosiaali- ja persoonallisuuspsykologia.

Maria Olkkonen, Dr. rer. nat. Tutkimusala: havaintopsykologia, värihavainto, kognitiivinen neurotiede.

Esa Palosaari, PsT. Tutkimusala: päätöksentekotutkimus ja sosiaalipsykologia.

Tuuli Anna Renvik, VTT. Tutkimusala: ryhmien välisten suhteiden sosiaalipsykologia.

Kaisa Saurio, PsM. Tutkimusala: kehityspsykologia ja tunteiden säätely.

Mia Silfver-Kuhalampi, VTT. Tutkimusala: moraalipsykologia, emootiot, hyvinvointi.

1. Mikä sai sinut aikanaan havahtumaan psykologian replikaatiokriisiin? Mikä oli havainto, tutkimus tai kirjoitus, joka sai sinut oivaltamaan tilanteen ongelmallisuuden?

Matti Heino (MH): Tehdessäni kandityötä positiivisesta (työ)psykologiasta ymmärsin noin puolessa välissä alkaa katsomaan kriittisiä lähteitä, mikä osoittautui pohjattomaksi arkuksi. Seurailin myös Facebookin menetelmäryhmiä sekä Twitteriä, joissa käydyt kiistelyt omaksi ihmetykseksenі saivat tilastotieteen vaikuttamaan… kiinnostavalta? Lisäpontta toi se, ettei silloin kukaan henkilökunnasta opettanut näistä asioista tai pitänyt niitä kovinkaan ihmeellisinä; erityisen mieleenpainuvina muistan keskustelut HARKingista (Kerr, 1998) ja *p*-arvojen tulkinnasta.

Ville-Juhani Ilmarinen (V-JI): Vaikea muistaa ensimmäistä havahduttavaa tekstiä, mutta se on varmaa, että en ottanut alussa sitä niin tosissaan kuin olisi ollut syytä. Oivaltaminen ei siis tullut kuin salama kirkkaalta taivaalta vaan sai alkusysäyksen oman alan tutkijoiden blogeista. Luulen, että Many Labs -projektien tulokset ovat olleet kaikkein keskeisimmässä asemassa omassa havahtumisessani etenkin kriisin levinneisyyteen.

Sointu Leikas (SL): Sain omassa työssäni esivaroituksen ennen varsinaisen replikaatiokriisin puhkeamista. Olin vuonna 2011 laatinut rahoitushakua varten tutkimussuunnitelman, jossa viitattiin sekä Diederik Stapelin kollegoiden tutkimuksiin että ego depletion- ja social priming

-kirjallisuuteen (kolme hutia yhdessä suunnitelmassa!). Stapelin väärinkäytökset paljastuivat pari kuukautta ennen suunnitelman jättämistä. Pian tämän jälkeen tulivat esille myös priming- ja ego depletion -kirjallisuuden ongelmat. Jouduin kirjoittamaan nopeasti uuden suunnitelman. Tämä omakohtainen kokemus todella teki tietoiseksi siitä, miten isoista ongelmista on kysymys, vaikka Stapelin toiminta olikin omaa luokkaansa, eikä aineiston täydellinen fabrikointi ole replikaatiokriisin ydin.

Myöhemmin aloin seurata muun muassa Data Colada- ja Replicability Index -blogeja sekä Andrew Gelmanin blogia, joissa käsiteltiin kriisiä perusteellisesti metodologiselta kannalta.

Itselleni yksi iso juttu oli huomata, että ei-replikoituvia tuloksia ja heikoin metodein tehtyjä tutkimuksia oli julkaistu myös kaikkein arvostetuimmissa lehdissä. Toki tämän olisi pitänyt olla itsestään selvää, mutta kuitenkin ennen kriisiä olin jossain määrin käyttänyt heuristiikkaa, jonka mukaan korkean impaktin lehdissä julkaistuihin tuloksiin voi luottaa. Nythän termistä "prestigious journal" on tullut hieman pilkkanimitys, kun monet kovatasoisissa lehdissä julkaistut raflaavat tulokset – joiden kaltaisia nämä lehdet ovat aiemmin nimenomaan toivoneet – ovat osoittautuneet epäluotettaviksi.

Jari Lipsanen (JL): On vaikea määritellä mitään tiettyä kirjoitusta tai tutkimusta. Lähtökohtaisesti tietyt ongelmat ovat olleet tiedossa aina ja esimerkiksi tilastotieteen opetuksessa Helsingissä on aina pyritty korostamaan niin sanottuja hyviä tutkimuskäytäntöjä, eli *p*-hackausta tai muuta sellaista itsepetosta ei pitäisi harrastaa. Valitettavasti on samanaikaisesti myös aina ollut hyvin tiedossa, että kyllä sitä on harrastettu.

Jan-Erik Lönnqvist (J-EL): Kun Daryl Bemin artikkeli julkaistiin 2011 *JPSP:ssä*, olin melko hämmentynyt. Artikkelissa hän siis esittelee yhdeksän koetta, joiden tulokset viittaisivat siihen, että ihmisillä on psi-voimia, tai tarkemmin, että he kykenevät ennustamaan tulevia tapahtumia, esimerkiksi kolikonheittoja, paremmin kuin sattuma antaisi olettaa. Ajattelin keskustelua seuratessani, että Bem, joka siis oli sosiaalipsykologian supertähti, halusi osoittaa, että julkaisukulttuurissamme on jotain mätää. Mutta kävikin ilmi, että hän aivan vilpittömästi uskoi löydöksiinsä. Kuinka pielessä asiat ovat, aloin tosissani tajuta vasta vuosia myöhemmin, kun en itse eivätkä monet kollegat kyenneet replikoimaan klassisina pidettyjä perustuloksia psykologian eri aloilta. Lopullinen oivallus koitti, kun huomasin, etten enää tiennyt, mihin aiempaan tutkimukseen luottaa, kun suunnittelin uusia tutkimuksia.

Maria Olkkonen (MO): En osaa sanoa yhtä tiettyä kirjoitusta, mutta kiinnitin yleisesti asiaan huomiota, kun siitä alettiin puhua enemmän noin kymmenen vuotta sitten. Ja kun asiasta alettiin puhua, aloin itsekin kiinnittää ongelmiin huomiota lukiessani alan kirjallisuutta.

Esa Palosaari (EP): Olin kuullut tutkimuksista, joissa replikoitui vain alle puolet kokeiden tuloksista, mutta se ei vielä havahduttanut minua miettimään asioita uusiksi. Saatoin ajatella, että replikoimattomuus voisi johtua julkaisuharhasta ja siitä, että *p*-arvokriteerillä .05 saadaan aina väärä positiivinen tulos todennäköisyydellä 5 %. Varsinainen havahtuminen tapahtui, kun huomasin, että *p*-arvotulkintani oli väärä. Olin nähnyt Teemu Pauhan Facebookissa jakaman tekstin tutkijoiden virheellisistä *p*-arvotulkinnoista ja lukenut Colquhounin (2014) artikkelin samasta aiheesta. Colquhoun argumentoi, että *p*-arvolla .05 väärän positiivisen tuloksen todennäköisyys voi olla paljon suurempi kuin 5 %, esimerkiksi kymmeniä prosentteja. Siten tutkimusten replikoitumattomuus olisi vähemmän yllätyksellistä ja replikointi tarpeellisempaa. Ja tilastoanalyysien tekeminen ja tulosten tulkinta olisi suurella todennäköisyydellä pielessä muillakin kuin minulla, koska tiedeyhteisö oli antanut analyyseistäni palkintojakin. Aavistukseni ei ollut väärä: esimerkiksi Nummenmaan tilastollisten menetelmien oppikirjassa on selkeästi virheellinen *p*-arvomääritelmä ja tulkinta, ja 89 prosentissa pohjoisamerikkalaisista psykologian oppikirjoista on jokin väärä *p*-arvotulkinta (Cassidy, Dimova, Giguère, Spence & Stanley, 2019). Jari Lipsanen mainitsi kerran Fisherin ja Neymanin ja Pearsonin kiistoista *p*-arvoihin liittyen, ja minulle on selvinnyt, että varsinainen frekventistinen tilastotiedekään ei näytä vieläkään ratkaisseen kyseisiä kiistoja. Ainakin Helsingin

yliopiston, Aalto-yliopiston ja MIT:n frekventistisen tilastollisen päättelyn kurssien oppimateriaalit näyttävät sisältävän samoja 1900-luvulta peräisin olevia sisäisiä ristiriitoja, joiden ajattelun ritualistisesta välttelystä on syytetty psykologeja (Gigerenzer, 2004). Vaikuttaa siltä, että harvat tietävät, mitä tekevät, mikä johtaa virheelliseen tilastollisten menetelmien käyttöön, tulkintaan, raportointiin ja julkaisupäätöksiin.

Tuuli Anna Renvik (TAR): Muistelen, että alun perin olisin havahtunut asiaan sosiaalisen median verkostoni kautta: erityisesti kokeelliseen (sosiaali)psykologiaan keskittyneet kollegani eri maista alkoivat tuoda ongelmia enenevissä määrin ilmi. Jälkeenpäin ajateltuna tämä on siinä mielessä ”mukava” muisto, että tutkijayhteisö puuttui asiaan sisältä päin ja alkoi käydä keskustelua myös epävirallisempia kanavia pitkin. Melko pian myös opiskelijani alkoivat kysellä asiasta, ja pidin tärkeänä ottaa replikaatiokriisin käsittelyn osaksi opetustani johdantokurssista alkaen.

Kaisa Saurio (KS): Havahduin replikaatiokriisiin omaan väitöskirjatyöhöni liittyvien ongelmien ja virheiden kautta. Olin tietoinen esimerkiksi julkaisuvinoumasta jo perusopintojen aikana. Ajattelin kuitenkin pitkään, että tarvitaan vain pientä hienosäätöä. Replikaatiokriisistä kuulemisen jälkeen tutustuin minulle suositeltuihin materiaaleihin melko hitaasti, sillä aina oli tulossa jokin näennäisesti kiireellisempi deadline. Lopulta selvisi, että lähes kaikki ne konkreettiset ja kiireelliset ongelmat liittyivät tiiviisti replikaatiokriisiin.

Halusin tehdä tutkimukseni tilastolliset analyysit mahdollisimman tarkasti ja oikein. Yritin paniikissa selvittää, miksi kaikki pienetkin muutokset muuttivat tulokset täysin. Tutkimusprojektin aineisto oli hyvin pieni, ja etsin tietoa myös siitä, miten pieni otoskoko tulisi huomioida analyyseissa ja tulosten tulkinnassa. En löytänyt kumpaankaan kysymykseen vastauksia oppikirjoista tai muista yleisesti käytössä olevista materiaaleista. Välillä ajattelin olevani liian tyhmä tekemään väitöskirjaa, ja välillä kuvittelin, että minulla on huijarisyndrooma. Päädyin lopulta käymään Daniël Lakensin Improving your statistical inferences -nettikurssia ja kuulin ensimmäisen kerran tilastollisesta voimasta (https://www.coursera.org/learn/statistical-inferences). Pian selvisi, että omassa tutkimuksessani väärän negatiivisen todennäköisyys läheni sataa prosenttia. Tämä on psykologian tutkimuksessa hyvin tyypillistä (ks. esim. Button ym., 2013; Szucs & Ioannidis, 2017).

Väärän negatiivisen tuloksen suuresta todennäköisyydestä huolimatta olin onnistunut saamaan joissain analyyseissa tilastollisesti merkitseviä tuloksia. Olin eksynyt kyseenalaisten tutkimuskäytäntöjen harmaalle alueelle. Jos olisin vielä valikoinut tulokset tilastollisen merkitsevyyden tai kiinnostavuuden perusteella, kyseessä olisi ollut tyylipuhdas *p*-hakkerointi (Gelman & Loken, 2013; Simmons ym., 2011). *P*-arvot eivät toimi virhekontrollina, jos niitä ei ymmärretä ja käytetä asianmukaisesti. Jostain syystä hyvin suuri osa oppikirjoista määrittelee tilastollisen merkitsevyyden väärin (Cassidy ym., 2019). *P*-arvot eivät esimerkiksi taianomaisesti paljasta, miten todennäköisesti hypoteesi on totta tai miten todennäköisesti saadut tulokset ovat sattumaa (Gigerenzer, 2018).

Mia Silfver-Kuhalampi (MS-K): Vahvimmin on jäänyt mieleen laaja artikkeli Stapelin tapauksesta, olikohan *Helsingin Sanomien Kuukausiliitteessä*? Oli uskomatonta, miten paljon artikkeleita oli julkaistu keksityistä aineistoista, ja tunsin empatiaa niitä tohtoriopiskelijoita kohtaan, joiden työ käytännössä tuhoutui.

2. Mitkä ovat omasta mielestäsi huolestuttavimpia piirteitä replikaatiokriisissä (psykologian alalla tai tieteessä laajemmin) tai sitä koskevassa keskustelussa?

MH: Tuntuu, ettei kaikkien mielestä näistä pitäisi keskustella näkyvästi, sillä ”ihmiset menettävät luottamuksensa tieteeseen”. Mielestäni *huonon* tieteen luottamuspula ei ole ongelmallista, ja ainoa tapa välttää vielä suuremmat ongelmat on läpinäkyvä keskustelu.

V-JI: Se on huolestuttavaa, ettei tieto keskimäärin ole luotettavaa ja että tätä ei kovin avoimesti tunnusteta. Parannusten etenemisnopeuden kannalta on myös huolestuttavaa, että varhaisen

uravaiheen tutkijat ovat olleet yliedustettuna tässä keskustelussa ja kokeneemmat aliedustettuna. Siihen on varmasti useita syitä, mutta tiedon tuottamisprosessin uudistamisen kannalta olisi toivottavaa, että ne, joiden sana painaa alan sisällä tällä hetkellä eniten, olisivat myös kiihdyttämässä tätä muutosta.

SL: Monet tutkijat ovat tarttuneet ongelmiin erittäin rakentavalla tavalla, ja uskon, että suurin osa tiedeyhteisöstä tulee muuttamaan tai on jo muuttanut tutkimuskäytäntöjään. Siinä mielessä nykytilanne ei ole mielestäni hirveän huolestuttava. Ikävin asia on mielestäni se, että luottamus olemassa olevaan kirjallisuuteen on osittain mennyt. Toinen huolestuttava asia on, että jotkut arvostetut lehdet eivät ole ainakaan näkyvästi lähteneet mukaan purkamaan kriisiä, ja jotkut arvostetut tutkijat ovat kokonaan kieltäytyneet näkemästä ongelmia. Ja tietysti on erittäin ikävää, että jotkin alalla vakiintuneet, luotettavina pidetyt ilmiöt eivät näytäkään pitävän paikkaansa.

Kriisin alussa joidenkin alalla vahvassa asemassa olevien tutkijoiden reaktio – vähättely ja kriisistä puhuvien tutkijoiden syyttely – oli jonkinasteinen järkytys, mutta siitä on toivottavasti päästy nyt eteenpäin.

JL: Suurin ongelma replikaatiokriisikeskustelussa on, että sitä tunnutaan valitettavan usein pitävän tilastollisena kriisinä. Sitä se ei nähdäkseni ole, sillä replikaatiokriisin aikana ei ole varsinaisesti esitetty yhtään tilastotieteeseen liittyvää tulosta, joka olisi jotenkin yllättävä. Lähtökohtaisesti kaikki esimerkiksi *p*-arvoihin liittyvät tulkinnat, efektin koon ja tilastollisen voiman arvioinnin tärkeys on esitetty jo vuosikymmeniä sitten. Kysymyksessä on pikemminkin psykologinen kriisi, joka liittyy julkaisujärjestelmään, jossa lähtökohtaisesti halutaan aina ”uusia” tuloksia ja jokaisella artikkelilla pitäisi olla jokin uusi kontribuutio tieteeseen. Tämä näkemys on mielestäni ongelmallinen, sillä lähtökohta on ja on aina ollut, että kaikkia tuloksia on koeteltava. Kuitenkin tutkimuksen ulkopuoliset tavoitteet asettavat haasteita tieteellisille ideaaleille. Samoin virantäytöissä suositaan vahvasti pitkää julkaisuluetteloa. Media ja tiedeyhteisö haluavat tiedeviestintää, jossa nostetaan esiin yksittäisten tutkimusten tuloksia, jotka ovat kiinnostavia ja ”mullistavat” tietyn aihealueen. Ainakin itse näen nämä tekijät replikaatiokriisin suurimpina taustasyinä. Suurin ongelma on aina ollut, että replikaatioita ei tehdä tarpeeksi, ja tämä uhkaa tieteen itseään korjaavaa prosessia.

Valitettavaa on myös se, että varsinkin alkuvaiheessa tehtiin tutkimuksia, joissa yksittäisiä tuloksia replikoitiin ainoastaan yhden kerran ja tämän replikaation perusteella määriteltiin tuloksen totuusarvo. Ei havaittu sitä, että uskomusjärjestelmä ei ole dikotomia, vaan pikemminkin jatkumo, jossa suuren määrän toistoja perusteella uskomuksemme tiettyyn tulokseen siirtyy hitaasti kohti varmempaa uskomusta tuloksen paikkansa pitävyydestä tai pitämättömyydestä. Epävarmuuden määrän pitäisi olla selvää myös puhtaasti psykologisten ilmiöiden matemaattisesta formulaatiosta. Varsin usein nimittäin esityksissä ja jopa tieteellisissä tutkimuksissa näkee hyvinkin monimutkaisen vuokaavion kuvaamassa jotain tiettyä ilmiötä, ja lopulta perusteena esitetään muuttujien välisiä korrelaatioita. Tämä on selkeä yksinkertaistus ja tietää varsin varmasti, että näin se ei ole. Ei flow ole taidon ja haasteen suhde, ei maailma ole automaattisesti normaalisti jakautunut ja niin edelleen. Varsinkin validiteettiongelma alkaa olla jo todellinen uhka, koska samalla kun replikaatiokriisin myötä vaaditaan suurempia aineistoja, tutkimusryhmät vaikuttavat pyrkivän koko ajan lyhyempiin ja nopeammin täytettäviin mittareihin. Kysymys kuuluukin: meneekö lapsi pesuveden mukana, kun suuremman *N*:n saamiseksi uhrataan mahdollisesti osa mittauksen validiteetista?

J-EL: Nyt ehkä harmittaa eniten, että niin monia ilmiöitä on ylipäätään lähdetty psykologisoimaan. Jos esimerkiksi osoittautuu, että stereotypiauhkaa ei ole olemassa tai että IAT:llä mitatut implisiittiset asenteet eivät ennusta rasistista käyttäytymistä, tätä voidaan käyttää lyömäaseena esimerkiksi väitettäessä, että joidenkin ryhmien huonompi koulumenestys johtuukin siis ryhmän ominaisuuksista (koska stereotypiauhkatulokset eivät replikoidu) tai että rasismia ei ole olemassa (koska tietyt IAT-tulokset eivät replikoidu). Jotenkin siis huolestuttaa se, miten tätä kriisiä tullaan käyttämään yhteiskunnallisessa keskustelussa.

EP: Huolestuttavin, vaikkakin inhimillinen ja ymmärrettävä, piirre on ollut kyynisyys sekä epäilynä muista tutkijoista että havaintona heistä. Monetkaan tutkijat eivät usko muiden tutkijoiden etsivän rehellisesti totuutta, vaan olevan erilaisten materiaalisten ja arvostukseen liittyvien kannustimien ohjaamia. Käsitys ei valitettavasti ole kuvitteellinen. Olen kuullut tutkijan sanovan, että kaikki väitteet nykytutkimuksessa olevista tilastollisen voiman ja *p*-arvojen ongelmista pitävät kyllä paikkansa, mutta miksi hänen pitäisi olla ensimmäisten joukossa tekemässä asioita oikein. Ilmeisesti julkaisuja saa lehtiin tekemällä asioita väärin.

TAR: Kriisin tässä vaiheessa, kun monia korjausliikkeitä on jo pyritty tekemään tutkimuskentällä, olen erityisen huolissani tieteen uskottavuudesta ulkokehällä olevien kuten opiskelijoiden tai ylipäänsä tieteestä kiinnostuneiden kansalaisten silmissä. He eivät voi olla kunnolla kartalla asian mittasuhteista, ongelmiin liittyvästä keskustelusta eivätkä askelmerkeistä eteenpäin, joten hämmennys voi ymmärrettävästi olla melkoinen. Tämäkin kriisi voidaan kuitenkin nähdä tieteen mahdollisuutena kehittyä ja korjata itseään, kuten asiaan kuuluu.

KS: Olen itse eniten huolestunut siitä, miten vähäistä replikaatiokriisiä koskeva keskustelu on Suomessa. Kuulen jatkuvasti tutkijoilta ja opiskelijoilta, että aiheesta ei juuri puhuta. Suomenkielistä materiaalia on olemassa hyvin vähän, tai se on hyvin vaikeasti löydettävissä. Uusia tutkimuksia suunnitellaan ja rahoitetaan kuin mitään ei olisi tapahtunut. Tiede on rikki, missä ovat hätäkokoukset ja selvitykset? Suomen kilpailukyvystä ja yliopistojen asemasta kansainvälisissä vertailuissa puhutaan paljon. Tulevaisuudessa tulemme jäämään muista pahasti jälkeen, jos emme ota replikaatiokriisiä vakavasti ja pian.

MS-K: Huolestuttavinta on ollut havaita se, miten paljon heikkouksia julkaisuprosesseissa on ollut. Monet ovat tukeutuneet työssään sittemmin kelvottomiksi havaittuihin lähteisiin.

Välillä olen miettinyt, onko rehellisellä ihmisellä mitään mahdollisuuksia pärjätä kilpailussa.

3. Missä määrin tiedeyhteisön keskustelu on mielestäsi keskittynyt olennaisimpiin asioihin?

MH: Mielestäni ”tiedeyhteisö” ei ole homogeeninen entiteetti; sen sisällä on käyty monenlaisia keskusteluja, ja onkin tärkeää, että ihmiset tekevät paljon erilaisia asioita tutkimuksen luotettavuuden parantamiseksi.

V-JI: Jos seuraa keskustelua parannusehdotuksista, niin siellä mielestäni puhutaan olennaisista asioista. Toki siihen olennaisesti liittyy ymmärrys siitä, mitä aikaisemmin ja vielä nykyäänkin laajalti tehdään väärin, mutta painotus voisi olla enemmän parannuksiin keskittyvä. Voi kyllä olla, ettei näitä kahta oikein voidakaan erottaa toisistaan. On myös vähemmän olennaisia teemoja, kuten minkälaisella sävyllä huonosti toistuvia tuloksia tulisi kritisoida. Retoriikka ei tietenkään näyttele mitätöntä roolia, mutta standardien luominen tieteellisen keskustelun sävyyn tuntuu tässä vaiheessa resurssien haaskaukselta.

SL: Suuressa määrin. Olen positiivisesti yllättynyt ja vaikuttunut siitä, miten moni on lähtenyt aktiivisesti selvittämään ja muuttamaan tilannetta ja on avoin muutokselle. Lisäksi konkreettisia muutoksia, jotka ratkaisevat joitakin ongelmia, on tehty.

Yksi asia, joka on mietityttänyt minua keskustelussa, on menetelmäopetuksen parantamisen korostaminen perus- ja jatko-opinnoissa. Menetelmäopetus on tietenkin äärimmäisen tärkeää sinänsä, mutta itse en pidä siihen liittyviä tekijöitä kovin olennaisena osana replikaatiokriisiä. Huipputasonkin koulutuksen saaneet oppivat tutkimuksen tekemisen käytännöt vasta itse tekemällä sitä. Toisekseen metodologisiin virheisiin ja aineiston viilaamiseen ovat julkisuuden tietojen perusteella syyllistyneet myös erittäin kokeneet tutkijat, joiden menetelmäosaamisessa tuskin on suuria puutteita. On syytä olettaa, että tulosten kalastelua on jopa opetettu joissain tutkimusryhmissä nuorille tutkijoille. Taidoista ei ole apua tällaisessa ilmapiirissä. Tämän vuoksi asenteiden ja tutkimuskäytäntöjen muutos on mielestäni keskeisempi asia.

JL: Keskustelussa tunnutaan valitettavan usein rajoittuvan siihen, miten saadaan varmistettua, että yksittäisen julkaistun artikkelin tulos on ”totta”. Tämä näkökulma on yleisesti tieteessä varsin kestämätön, ja varsinkin se on kestämätön nuorissa tieteenaloissa kuten psykologiassa. Replikaatiokriisikeskustelussa on rajoituttu valitettavan paljon keskusteluun hypoteesin testauksesta (*p*-arvot), joka on kuitenkin tutkimuksessa varsin kapea osa-alue.

Huomattavasti tärkeämpää olisi keskustella otannasta, eli onko meillä oikeasti edustavia otoksia siitä perusjoukosta, johon väitämme tuloksia yleistävämme? Vastaavasti pitäisi huomattavasti enemmän puhua käytetyistä mittareista. Liian usein artikkeleissa mittari otetaan varsin annettuna ja ainoastaan sen nimen ja mahdollisesti Cronbachin alfan mainitseminen riittää julkaisufoorumille osoittamaan sen pätevyyden. Kuitenkin mittaamme varsin abstrakteja asioita ja pitäisi keskustella huomattavasti enemmän siitä, arvioimmeko mittareillamme oikeasti persoonallisuutta, persoonallisuushäiriöitä, kiintymyssuhdetta tai mitä ikinä tutkimuksessa ollaankaan arvioimassa. Replikaatiokriisin sijasta haluaisin itse siirtää keskustelun pohdinnaksi siitä, onko meillä (tai pitäisikö olla) pikemminkin otantakriisi ja validiteettikriisi.

Replikaatiokriisin ratkaisuksi on esitetty muun muassa suurempia otoskokoja, joista ei ole erityisesti hyötyä, jos otokset ovat valikoituneita. Vastaavasti *Naturessa* ja *Sciencessä* julkaistut jutut ovat tyypillisesti nykyisin psykologian alalla juttuja, joissa tutkimusten *N* on satojatuhansia tai jopa miljoonia, jolloin voi jopa keskustella, mikä merkitys hypoteesin testauksella on, vaikka *p*-arvot tai vastaavat juttuihin laitetaan. Haasteita siis on, kun yritetään löytää tilastollista keinoa vahvistaa tulosten replikoitumista. Kyllä niitäkin on, mutta ne ovat normaaleja hyviä käytäntöjä ja silti nekään eivät sataprosenttisella varmuudella takaa tulosta.

MO: Mielestäni ainakin havaintopsykologian ja kognitiivisen neurotieteen alalla keskustelu on liittynyt tärkeimpiin asioihin kuten posiitivisen julkaisemisen harhaan (*positive publication bias*) ja siihen liittyen vaikeuteen julkaista nollatuloksia ja replikaatioita.

EP: Tutkijat ovat kiinnittäneet huomiota sekä vinoutuneeseen, virheelliseen kirjallisuuteen johtaviin kannustimiin että tietoon. Molemmat ovat olennaisia. Keskustelu on myös johtanut toimenpiteisiin. Daniël Lakens on kouluttanut netissä kymmeniä tuhansia tulkitsemaan *p*-arvoja, luottamusvälejä ja tilastollista voimaa oikein sekä ymmärtämään, mitä ongelmia voi esimerkiksi olla siinä, että kerää lisää aineistoa, kunnes saa tilastollisesti merkitsevän tuloksen (https://www.coursera.org/learn/statistical-inferences). Brian Nosekin ja kollegoiden työ läpinäkyvämmän tieteen infrastruktuurin ja kannustimien parantamiseksi on edesauttanut sitä, että tällä hetkellä yli 200 lehteä tarjoaa mahdollisuuden julkaisuun rekisteröidyn raportin muodossa (https://osf.io/rr/). Rekisteröidyissä raporteissa vertaisarviointi tehdään käsikirjoitukselle ennen aineiston keräämistä ja tuloksia, ja jos käsikirjoitus hyväksytään, se julkaistaan aineiston keräämisen jälkeen riippumatta tulosten tilastollisesta merkitsevyydestä. Tämän voisi ajatella poistavan sekä julkaisuharhan että eksploratiivisen ja konfirmatorisen tutkimuksen sekoittamisen (sekä sekoittamisen piilottamisen) eli *p*-hakkeroinnin. Näiden kahden seikan poistaminen johtaa *p*-arvojen ja efektikokojen tulkinnan kannalta luotettavampaan kirjallisuuteen. Datan ja analyysikoodien avoimesta julkaisemisesta sekä analyysisuunnitelmien esirekisteröinnistä artikkeliin saatavat merkit (ks. esim. https://cos.io/our-services/open-science-badges/) näyttävät suosituilta.

KS: Oma kokemukseni on, että tiedeyhteisö on replikaatiokriisin suhteen hyvin jakautunut. On olemassa koko ajan laajeneva reformiliikkeen kupla, jossa replikaatiokriisi otetaan hyvin vakavasti. Kuplan ulkopuolinen keskustelu pääosin jatkuu kuin kriisiä ei olisi. Reformiliikkeen kuplassa keskustelu on vilkasta, kansainvälistä ja moninaista. Mielestäni on hyvä, että keskustelussa ei keskitytä vain yhteen asiaan, vaan pureudutaan ongelmaan monesta eri suunnasta. Esimerkiksi mittaamiseen ja teoriaan liittyvät ongelmat ovat vielä jääneet paitsioon, mutta tilanne on muuttumassa nopeasti (ks. esim. Flake & Fried, 2019).

MS-K: Minusta keskeisistä asioista on puhuttu melko paljon.

4. Jotkut sanovat, että replikaatiokriisiin liittyviä ongelmia on liioiteltu. Mitä mieltä sinä olet?

MH: Mielestäni niitä on lähinnä vähätelty, mikä on ymmärrettävää, jos oma ura ja asema jossain määrin riippuu aikaisemman tutkimuksen luotettavuudesta.

V-JI: Olen eri mieltä. Luulen, että ongelmia vähätellään, kun huomioidaan ala koko laajuudessaan. Havainto liioittelusta voi johtua siitä, että osa kriitikoista on todella äänekkäitä ja räikeitä sävyltään, mutta mielestäni monet heistä puhuvat oikeista asioista.

SL: Olen täysin eri mieltä. Teen itse perustutkimusta, ja tutkimuksen suunnittelussa olisi voitava luottaa aiempaan tutkimukseen ja rakentaa sen päälle. Tällä hetkellä suuri osa oman alani perustutkimuksen tuloksista on menettänyt uskottavuutensa, vaikka vain osa on suoranaisesti osoitettu epäluotettaviksi. Samoin jos tällä hetkellä pitäisi pitää sosiaalipsykologian peruskurssi, kurssin sisällön suunnittelu tuntuisi hyvin vaikealta.

Jotkut sanovat, että kyse on normaalista tieteen itsekorjaavuudesta – jotkut tulokset vain osoittautuvat luotettavammiksi kuin toiset, ja vähitellen ”väärät” tulokset putoavat pois tieteen kelkasta. Tämä näkökulma ei kuitenkaan ota huomioon laajalle levinneitä kyseenalaisia tutkimuskäytäntöjä, jotka ovat se varsinainen ongelma. Kriisin ydin ei ole replikaatioprojekteissa kyseenalaisiksi joutuneet tutkimuslinjat, vaan se, että metodologiset heikkoudet ja tulosten kalastelu ovat osoittautuneet niin yleisiksi alalla. Tämä heittää varjon kaiken alan tutkimuksen ylle.

JL: On ja ei ole. Lähtökohtaisesti kriisin aikana paljastuneet *p*-hacking ja kyseenalaiset tieteelliset käytännöt on hyvä kitkeä pois, puhumattakaan paljastuneista datan väärentämisistä ja muista vilpeistä.

Toisaalta replikaatiokriisikeskustelussa on esiintynyt myös valitettavia ylilyöntejä. Tuloksen replikoitumattomuuteen liittyy tietynlaista häpeää, vaikka sen pitäisi olla aivan normaali osa tieteen tekemistä. Blogi- ja sosiaalisen median keskusteluissa esiintyy myös paljon syyllistämistä niitä tutkijoita kohtaan, joiden tulokset eivät ole replikoituneet. Tämä on erittäin valitettavaa ja uskon, että suuressa osassa tapauksia tarpeetonta. Ainakin itse uskon, että oikeastaan yksikään tutkimus, jota olen ollut itse tekemässä, ei lopulta ole täysin totta, ja oikeastaan toivon, että suuri osa niistä saataisiin lopulta kumottua. Kuitenkin haluan uskoa, että jokainen on vienyt suuressa mittakaavassa tiedettä edes hiukan eteenpäin (ja jopa harha-askeleet ovat osa tätä prosessia) ja luo pohjaa tulevalle tutkimukselle.

J-EL: Päinvastoin, ongelmaa vähätellään ja/tai sitä ei ymmärretä.

MO: Tietyiltä osin ongelmat ovat olleet vakavia, muun muassa positiivisen julkaisemisen harha. Tämä tarkoittaa sitä, että julkaistavaksi hyväksytään vain tutkimuksia, joissa löydetään tukea testatulle efektille. Tästä seuraa vääristymiä julkaistuun tutkimustietoon. Alkuperäisen efektin löytäjällä saattaa olla pöytälaatikossa 19 tutkimusta, joissa ei löytynyt merkitseviä efektejä, mutta näitä ei olisi saanut julkaistua, koska niissä on nollatulos. Myös muut tutkijat ovat saattaneet yrittää löytää saman efektin, mutta eivät ole julkaisseet nollatuloksia. Tästä kaikesta seuraa se, että efektin voimakkuus yliarvioidaan, ja kun sitä myöhemmin yritetään replikoida, se ei välttämättä onnistu, koska alkuperäinen efekti onkin paljon heikompi kuin miltä se on alkuperäisen raportin mukaan näyttänyt.

EP: Joitain ongelmia on ehkä liioiteltu, mutta tavalla, joka saattaa paljastaa vielä syvempiä ongelmia koulutuksessa ja tutkimuksessa. Tutkimusten replikoitumattomuus ei sinänsä minusta ole ongelma. Sen ongelmallisena pitäminen on tavallaan liioittelua. Jokin yksittäinen tutkimus voi tuottaa virheellisen johtopäätöksen, vaikka kaikki tehtäisiin oikein ja saataisiin pieni *p*-arvo. Ongelma on ollut ehkä ennemminkin siinä, että on uskottu ja luotettu, että replikaatiotutkimuksia ei tarvitsisi tehdä eikä saa julkaistua, vaan että kun yksi merkitsevä tulos on saatu, pitää tehdä jotain muuta uutta ja innovatiivista. Replikaatioidenkin olisi oltava erilaisia kuin alkuperäinen tutkimus. Fyysikko Richard Feynman havaitsi psykologien tekevän tässä suhteessa pseudotiedettä jo vuon-

na 1935 ja varoitti fyysikkoja sortumasta samaan (http://calteches.library.caltech.edu/51/2/CargoCult.htm). Kun on havaittu, että kokeiden suorat replikaatiot eivät tuota samassa määrin pieniä *p*-arvoja kuin alkuperäiset tutkimukset, se toivottavasti herättää epäilemään yksittäisten tutkimusten todistusarvoa ja tekemään useita replikaatioita. Yksittäinen tutkimus ei vielä kerro paljon, mutta ei kerro yksittäinen replikaatiokaan. Replikaatioita on syytä tehdä paljon, kuten Many Labs -projekteissa, jos halutaan varmistua jonkin ilmiön toistettavuudesta. Yksi ”epäonnistunut” replikaatio voi johtua esimerkiksi liian pienestä tilastollisesta voimasta replikaatiossa.

Replikaatiokriisiin liittyvät julkaisuharha- ja *p*-hakkerointiongelmat ovat nähdäkseni vakavia.

TAR: En sanoisi, että ongelmia on liioiteltu, mutta kyseessä on niin monimutkainen vyyhti, että keskustelussa menevät helposti puurot ja vellit sekaisin. On esimerkiksi muistettava, että heikkoon toistettavuuteen on hyvin monenlaisia syitä, kuten vaikkapa sosiaalisen vuorovaikutuksen tilanne- ja kontekstisidonnaisuus, mutta kaikki ei silti mene pelkän kontekstisidonnaisuuden piikkiin. Toisaalta on myös huomattava, että replikaatioprojekteihin on valikoitunut vain osa (kokeellisista) tutkimuksista, eli niissä löydetyt replikaatioprosentit eivät kuvaa koko tieteenalan kaikkien tutkimusten toistettavuutta ja uskottavuutta.

KS: Replikaatiokriisiin liittyviä ongelmia on varmasti joissain yhteyksissä liioiteltu, mutta niitä vähätellään paljon enemmän. Osa tutkijoista myös kyseenalaistaa toistettavuuden tärkeyden ja vetoaa tutkittujen ilmiöiden monimutkaisuuteen, vaikka lähes kaikessa psykologian tutkimuksessa on ainakin näennäisesti luonnontieteitä matkiva viitekehys. Tutkimustulokset päätyvät esimerkiksi oppikirjoihin ja hoitosuosituksiin yleistyksinä eivätkä kuvauksina tutkijan kohtaamasta ainutkertaisesta tilanteesta. Jos vain esitämme tekevämme tutkimusta samalla tarkkuudella kuin luonnontieteet, kyseessä on fyysikko Richard Feynmanin kuvaama lastikulttitiede (*cargo cult science*; Feynman, 1974). Jos emme aio pyrkiä toistettaviin tuloksiin, tulisi tutkimukset toteuttaa avoimesti laadullisina tai esimerkiksi teoreettisina pohdintoina.

MS-K: Siinä mielessä kyllä, että joidenkin merkittävien tutkimusten replikoitumattomuus ei minusta tarkoita, että ne eivät sinänsä olisi päteviä. Varsinkin sosiaalipsykologiassa saattaa olla paljon sellaisia ilmiöitä, jotka riippuvat ajasta, paikasta ja tutkittavasta populaatiosta. Minusta täytyy pohtia tarkkaan, missä määrin ilmiöitä voi yleistää.

5. Mitkä ovat sellaisia tiedeyhteisön viime vuosina käytäntöihinsä (mm. journalien editoriaalikäytännöt) tekemiä muutoksia, joihin olet tyytyväinen?

MH: Rekisteröidyt raportit ovat ainakin uusi artikkeliformaatti, joka auttaa moniin ongelmiin.

V-JI: Kaikki uudet käytännöt liittyen läpinäkyvyyteen ja avoimuuteen ovat tervetulleita. Muun muassa vaatimukset datan ja tulosten tuottamisen avoimuudesta, voimalaskelmat, esirekisteröinnit sekä etenkin rekisteröidyt raportit ovat sellaisia, joihin on vaikea olla tyytymätön.

SL: 1) Registered reports -prosessi, jossa lehti arvioi tarkan tutkimussuunnitelman ja hyväksyessään sen sitoutuu julkaisemaan tutkimuksen tuloksesta riippumatta. Tämä vähentää motivaatiota kalastella merkitsevää *p*-arvoa ja ehkäisee pöytälaatikkoefektiä. 2) Vaatimus aineiston, kaikkien mitattujen muuttujien ja kaikkien aineistolle tehtyjen analyysien julkisuudesta. Tämäkin ehkäisee tulosten kalastelua esimerkiksi muuttujien uudelleenkoodauksen tai tarkoitushakuisen analyyseihin valikoimisen avulla. Ongelmana on se, miten varmistetaan, että tutkijat todella julkaisevat rehellisesti kaikki mitatut muuttujat, havainnot ja analyysit. 3) Tutkimuksen esirekisteröinnistä palkitseminen joissakin lehdissä; samat hyödyt kuin edellä. Sitäkin voidaan käyttää väärin, esimerkiksi yrittämällä esirekisteröidä tutkimus retrospektiivisesti.

JL: Voimalaskelmien ja efektin kokojen raportoinnin vaatiminen on ollut hyvää kehitystä, ja ne ovat myös tilastollisesti perusteltuja.

Toisaalta tämän varjopuolena on ollut havaittavissa myös erikoisia vaatimuksia, kuten esimerkiksi voimalaskelmat niin sanotulle post-hoc-voimalle, joka on käsitteellisesti haastava,

vaatimus käyttää bayesilaista päättelyä, jonka tunnusluvut kuitenkin ovat suoraan verrannollisia muun muassa *p*-arvoihin, mikäli käytetään ei-informatiivisia priori-jakaumia ja niin edelleen. Tällöin ei tunnuta tuntevan kovinkaan hyvin tilastollisten menetelmien perustaa ja sitä lähtökohtaa, että tilastolliset menetelmät ovat työkalupakki, jossa on välineitä paljon muuhunkin kuin hypoteesintestaukseen. Meillä on muun muassa frekventistinen, bayesilainen, informaatioteoreettinen ja suurimman uskottavuuden lähestymistapa. Tavoitteena voi olla maksimaalinen ennuste tai tilastollinen päättely, joihin molempiin voi liittyä a priori -uskomuksia ilmiön käyttäytymisestä tai sitten ei. Mahdollisuuksia on monia, ja lopulta puhumme kuitenkin aina todennäköisyyksistä, joissa ei välttämättä ole absoluuttista dikotomista totuutta.

MO: Monet lehdet ovat alkaneet hyväksyä ”rekisteröityjä raportteja” (*registered reports*), joissa tutkijat kirjoittavat artikkelin johdannon ja metodit, ja artikkeli arvioidaan ja mahdollisesti hyväksytään julkaistavaksi tämän perusteella (mm. *eNeuro*). Tämä toivottavasti pienentää painetta saada positiivisia tuloksia, millä pitäisi olla pitkällä aikavälillä myönteinen vaikutus tieteelliseen julkaisemiseen.

EP: Olen tyytyväinen erityisesti edellä mainitsemiini rekisteröityihin raportteihin sekä mahdollisimman suuresta läpinäkyvyydestä palkitsemiseen. Oli ilo huomata, että *Nature Human Behaviour* -lehti ymmärtää, mistä Neyman–Pearson-tyyppisessä frekventisessä tilastollisessa päättelyssä on kyse ja vaatii rekisteröityjen raporttien suunnitelluilta frekventistisiltä tutkimusasetelmilta 95 %:n voimaa pienimmälle sisällöllisesti merkittävälle efektikoolle. Nähdäkseni fisheriläistä, eksploratiivisempaa tutkimusta ilman voimalaskelmia tulisi kuitenkin myös voida tehdä, jos se raportoidaan läpinäkyvästi.

TAR: Esimerkiksi tutkimuskysymysten ja hypoteesien esirekisteröinti, tutkimusaineistojen jakaminen sekä metodologisen vaatimustason kasvu ovat hyviä muutoksia.

KS: Tiedeyhteisössä kokonaisuutena on toistaiseksi tapahtunut hyvin vähän konkreettista muutosta. Poikkeuksena tästä on avoin julkaiseminen (*open access*), jonka eteen Suomessakin on tehty paljon. Reformiliikkeen kuplassa taas tapahtuu valtavaa ja nopeatahtista edistystä. Kriisin ratkaisemiseksi on perustettu kokonaisia organisaatioita, kuten Center for Open Science (COS) ja Society for the Improvement of Psychological Science (SIPS). COS on muun muassa julkaissut lehdille ja rahoittajille TOP Guidelines -ohjeistuksen (Nosek ym., 2015). Sen avulla on helppo ottaa käyttöön eri tasoisia vaatimuksia avoimuudelle esimerkiksi ennakkorekisteröintiin ja aineiston jakamiseen liittyen. Nyt aivan viime päivinä on julkaistu myös uusi TOP Factor, joka toimii vaihtoehtona yleensä käytetylle vaikuttavuuskertoimelle (JIF).

MS-K: Datan avoimuus on minusta hyvä ja tärkeä periaate. Käsikirjoitusta arvioivien henkilöiden on tarvittaessa voitava tarkistaa, että aineistossa ei ole mitään epäilyttävää ja analyysien laskemisessa ei ole virheitä. Hypoteesien esirekisteröinti on hyvä käytäntö myös.

6. Mitä vielä pitäisi tehdä, jotta ongelmilta jatkossa vältyttäisiin?

MH: Mielestäni ensimmäinen saavutettava askel on radikaali läpinäkyvyys raportoinnissa. Tämä ei estä tuhlaamasta resursseja sellaisen tutkimuksen tekemiseen ja arviointiin, joka ei olisi voinut alun alkaenkaan tuottaa informatiivisia tuloksia – mutta ainakin sen rajoitukset ovat edes periaatteellisessa mielessä selvitettävissä.

Toki yksi asia, mistä usein puhutaan, on ajattelun lisääminen, mitä hankaloittaa se, että tieteenfilosofien – samoin kuin tilastotieteilijöiden – ulosanti ei aina sytytä tulta soveltavan tutkijan sydämeen. Mutta peruskysymyksiä olisi tärkeää pohtia: Mitä oikeastaan tehdään, kun testataan hypoteesia, ja miksi? Voiko jonkin väitteen todistaa todeksi vai vain epätodeksi? Onko suurimmassa osassa satunnaistamatonta tutkimusta (esim. miesten ja naisten eroavuudet asiassa x), oikeasti kiinnostavaa tietää, ettei ryhmien välinen ero ole tismalleen 0.00000? Tämä jälkimmäinen on kysymys, jota tunnun toistavan ai-

nakin kerran viikossa, kun kuulen jonkun onnistuneesti hylänneen nollahypoteesinsa. Itselleni näitä asioita avasivat huomattavasti Daniël Lakensin kaksi ilmaista verkkokurssia (https://www.coursera.org/learn/statistical-inferences ja https://www.coursera.org/learn/improving-statistical-questions).

V-JI: Suoraan edelliseen kysymykseen liittyen näitä käytäntöjä ei ole otettu tarpeeksi laajasti käyttöön. Artikkeliväitöskirjojen laadullisia ja määrällisiä vaatimuksia voisi myös muuttaa suoraan tukemaan tällaisten käytäntöjen oppimista ja edistämistä. Yleisemmin tutkijakoulutuksessa pitäisi päästä eroon analyysirituaalien opettelusta ja painottaa tilastotieteen ja tieteenfilosofian ymmärrystä. Vertaisarviointikäytännöt ovat osa-alue, jonka toivoisi vielä muuttuvan siihen suuntaan, jossa varmistetaan, että näitä uusia vaatimuksia läpinäkyvyydestä osataan myös hyödyntää arvioinnissa. Vertaisarviointi voisi olla jatkuvampi prosessi, joka tapahtuu myös ennen käsikirjoituksen lähettämistä ja julkaisemisen jälkeen. Tämähän voi johtaa siihen, että julkaisusarjoja, ainakaan sellaisina kuin nyt, ei enää tulevaisuudessa samalla tavalla olisi.

SL: Syvin ongelma on tutkimuskulttuuri, jossa jahdataan merkitsevää tulosta tai teorian/hypoteesin todistamista ”todeksi”. Tämä pitäisi saada muuttumaan. Tähän suuntaan ollaan menossa, ja mainitut lehtien muuttuneet käytännöt tukevat asennemuutosta. Samoin *PLoS*-tyyppiset lehdet, jotka julkaisevat kaikki metodologisesti pätevät artikkelit. Kuitenkin niin kauan kuin teorian tai hypoteesin osoittaminen todeksi on arvostettavin tutkimuksen lopputulos, ongelmia tulee olemaan. Kaikkia hyviä käytäntöjä voi valitettavasti kiertää.

Eräs taustaongelma sosiaalipsykologiassa on mielestäni muun muassa Rozinin (2001) esiin nostama perushavaintojen puute. Toisin kuin useimmissa luonnontieteissä, sosiaalipsykologiassa ei historiallisesti lähdetty siitä, että oltaisiin kerätty tutkimuskohteesta eli ihmisen käyttäytymisestä runsaasti havaintoja luonnollisissa tilanteissa. Sen sijaan hypättiin nopeasti laboratorio-olosuhteisiin ja monimutkaisten teorioiden testaamiseen (toki poikkeuksia oli). Tämä perushavaintojen puute näkyy edelleen alallamme esimerkiksi siinä, ettemme tiedä, miten tilanteita tulisi käsitteellistää ja luokitella psykologisesti relevantilla tavalla, sekä siinä, että luotettavasti tunnistettavia säännönmukaisuuksia ihmisen käyttäytymisessä tunnetaan vain vähän. Lisäksi teorian testaamisesta on tullut arvostetuin tutkimuksen päämäärä. Näin ei tulisi olla, vaan esimerkiksi pitäisi olla aivan ok ja itse asiassa hyvinkin arvostettavaa kirjoittaa aikovansa tutkia ateoreettisesti ja laajalla muuttujien kirjolla, miten ihmiset käyttäytyvät erilaisissa arkipäiväisissä tilanteissa.

Konkreettisemmalla tasolla eräs tärkeä muutos ajattelussa olisi se, että jos tulosta täytyy kalastella, tulos ei todennäköisesti ole kovin luotettava tai sitten efekti on niin äärimmäisen pieni, että se ei ole kovin kiinnostava. Eli pitäisi siirtyä pois ajattelusta ”tässä on varmasti jotain, minun täytyy vain saada efekti esiin”. Koko tiedeyhteisön tulisi tukea nollatulosten hyväksymistä sekä laadukasta eksploratiivista tutkimusta.

JL: Replikaatiokriisin ja siihen liittyvään hypoteesin testaukseen liittyvien ongelmien osalta riittää, että tilastollisia menetelmiä ei nähtäisi pelkästään tilastollisina ajoina: temppuina, jotka tehdään, jotta saadaan merkitsevä tulos $p < .05$. Aineiston analysointi on kokonaisuus, joka lähtee liikkeelle tutkimusasetelman suunnittelusta ja parhaiten tutkimusongelmaan liittyvän analyysimenetelmän valinnasta. Nämä ovat asioita, joita ei tehdä vain siinä sivussa, ja ne vaativat koulutusta sekä tutkittavan ilmiön ymmärtämiseen että analyysimenetelmien ymmärrykseen.

J-EL: Ehkä yhden tutkimuksen replikointi voisi jopa olla pakollinen osajulkaisu väitöskirjaa tehdessä?

MO: Useampien lehtien pitäisi hyväksyä rekisteröityjä raportteja, tutkimusten toistoja ja nollatuloksia. Ymmärrän, etteivät kaikki lehdet näin voi tehdä (esim. erittäin valikoivat yleistiedelehdet kuten *Nature* ja *Science*), mutta tästä pitäisi tulla paljon yleisempää.

EP: Rahoittajat voisivat perehtyä asiaan ja kannustaa julkaisemaan rekisteröityjä raportteja. Si-

ten ehkä nekin tutkijat, jotka nyt tietävät, etteivät tee kuten pitäisi, tekisivät kuten pitäisi. En ole vielä lukenut sellaista oppikirjaa, jossa olisi käsitelty selkeästi Fisherin ja Neyman–Pearsonin lähestymistapoja ja niistä seuraavia erilaisia tulkintoja *p*-arvoille. Tutkijat ja opettajat voisivat kuitenkin tutustua kyseisiä *p*-arvotulkintoja käsitteleviin artikkeleihin ja alkuperäislähteisiin (Perezgonzales, 2015). *P*-arvojen tulkintojen opettamisessa voisivat toimia konkreettiset esimerkit (Colquhoun, 2014) ja Daniël Lakensinkin käyttämät simulaatiot, joista selviää esimerkiksi, että *p*-arvot ovat tasajakautuneita eli kaikki *p*-arvot ovat yhtä todennäköisiä, kun nollahypoteesi on tosi, tai että .05:n lähellä oleva *p*-arvo voi olla parempaa todistusaineistoa nollahypoteesin puolesta kuin sitä vastaan. Simulaatiot löytyvät edellä mainitsemaltani Courseran kurssilta (luento: https://www.youtube.com/watch?v=RVxHlsIw_Do; R-koodi: https://www.coursera.org/learn/statistical-inferences/supplement/jvAYS/assignment-1-which-p-values-can-you-expect). Hieman muokattu simulaatiokoodi on myös pitämäni kurssin sivulla (https://users.aalto.fi/~palosae2/).

TAR: Kaikenlainen avoimuus tieteen tekemisessä ei koskaan ole pahitteeksi.

KS: Ongelmilta ei tulla välttymään jatkossa, ja osa ratkaisuyrityksistä tulee todennäköisesti epäonnistumaan. Joitain vaikutuksia ei osata ennustaa oikein, tai uusi järjestelmä on yllättävällä tavalla pelattavissa. Virheiden tekeminen voi kuitenkin saada aikaan suurtakin edistystä, jos niistä pystytään oppimaan. Tällä hetkellä samoja virheitä toistetaan vuodesta toiseen.

MS-K: Arvioinnit voisivat aina olla anonyymejä sekä arvioijan että arvioitavan suuntaan. Joskus ”isot nimet” saavat julkaistua tutkimuksia, joiden laatu ei riittäisi tuntemattomalle tutkijalle.

7. Millaisia väärinkäsityksiä replikaatiokriisistä puhuttaessa olet huomannut esiintyvän tutkijoiden (tai maallikoiden) joukossa?

MH: Yksi ärsyttävimpiä on se, miten usein ihmiset ajattelevat esirekisteröinnin tarkoittavan sitä, ettei dataa voisi enää eksploroida tai analyyseissa olla luova – ainoa, mitä esirekisteröinti rajoittaa, on eksploratiivisen tutkimuksen esittäminen konfirmatorisena.

Yksi iso ongelma esimerkiksi lääketieteessä, jossa esirekisteröinti on ollut jo pitkään muodollisesti pakollista, on se, ettei rekisteröintien vastaavuutta artikkelissa käytettyihin menetelmiin tarkista kukaan (Goldacre, Drysdale, Dale ym., 2019; Goldacre, Drysdale, Marston ym., 2019). Tämä ongelma – eli tutkijat esirekisteröivät jotain ja sittemmin tekevät mitä haluavat ja esittävät sen konfirmatorisena – on jo tullut esiin psykologiassa (Claesen, Gomes, Tuerlinckx & Vanpaemel, 2019) ja tulee varmasti nousemaan näkyvyydessään.

V-JI: Ainakin kaksi väärinkäsitystä liittyy esirekisteröintiin. Ensimmäinen on, että esirekisteröinti olisi mahdollista vain, jos aineistoa ei ole vielä kerätty. Esirekisteröinti ennen aineiston keräämistä onkin mielestäni aivan ehdoton käytäntö, jos jostain syystä ei halua ryhtyä suoraan rekisteröityyn raporttiin, jossa tutkimuksen aihiot vertaisarvioidaan ennen aineiston keruuta. Esirekisteröintiä voidaan kuitenkin hyödyntää myös aineistoihin, jotka on jo kerätty ja joista on julkaistu jo muita tutkimuksia (Van den Akker ym., 2019). Esimerkiksi monien pitkittäisaineistojen keruu on aloitettu kauan ennen replikaatiokriisiä, eikä esirekisteröintiä tietenkään ole tehty ennen keruuta. Tällöin esirekisteröity osa ei voi sisältää kaikkia esirekisteröinnin palasia, kuten etukäteen määriteltyä tarvittavaa otoskokoa, mutta hypoteesit ja analyysisuunnitelman voi silti esirekisteröidä. Olennaista on, että esirekisteröinnissä selvästi mainitaan, mitä analyysejä on jo tehty ja mitä aineistosta tiedetään. Tällä tavalla on mahdollista määrittää etukäteen yksiselitteisempi suunnitelma testattavalle hypoteesille ja se, miten siitä tullaan tekemään tulkintoja.

Toinen väärinkäsitys on suoraa jatkoa tälle: voidaanko aineistosta sitten analysoida mitään muuta kuin esirekisteröity hypoteesi määritetyllä tavalla. Mielestäni on selkeää, että voidaan, mutta silloin nämä ei-rekisteröityjen analyysien tulokset täytyy esittää eksploratorisina, ei konfirmatorisina. Eksploratorisissa analyyseissä on myös tarpeen hyödyntää konservatiivisia korjausmenetelmiä *p*-arvoille, jotta tyypin I virhettä saadaan hillittyä.

Kolmas väärinkäsitys, johon kohtalaisen usein törmää, on, että meta-analyysit kertoisivat selkeästi tulosten toistettavuudesta. Koska tuloksilla, jotka eivät ole tilastollisesti merkitseviä, on huomattavasti pienempi todennäköisyys päätyä osaksi meta-analyysiä, on niissä huomattavasti julkaisuharhaa, ja tällöin myös meta-analyysin tuottama arvio ilmiöstä on huomattavan vääristynyt.

SL: Kollegoiden kesken meillä on suunnilleen samat ajatukset. Joissain keskusteluissa on noussut esiin, miten vaikea on uskoa, että jokin tietty ilmiö ei olekaan olemassa (tämä koskee myös itseäni). Tässä huomaa, miten vaikeaa on ajatella ihmisten käyttäytymistä ja tunteita koskevia säännönmukaisuuksia kylmän analyyttisesti, kun itse elää niitä.

JL: Katso vastaus kysymykseen 8.

J-EL: Monet vanhemman polven tutkijat vaikuttavat pitävän tätä kriisiä aiempien kaltaisena, ikään kuin sen voisi kuitata toteamalla, että olemme jo kauan tienneet, että tulokset usein ovat kulttuurisidonnaisia. Kun ongelma pikemmin on, että mitään tuloksia ei alun perin ollutkaan, vaan ne syntyivät tieteellisten käytäntöjemme seurauksena.

MO: Jos tieteellistä prosessia ei ymmärrä kunnolla, asian saattaa nähdä liian mustavalkoisesti. Esimerkiksi jos joku tutkimustulos ei replikoidu, jotkut saattavat ajatella, että kyseessä on tarkoituksellinen huijaus, vaikka kysymys on usein vain siitä, että tietyllä todennäköisyydellä tilastollisissa testeissä saadaan merkitseviä efektejä, vaikka efektiä ei oikeasti ole (ns. ykköstyypin virhe). Voi olla hankala ymmärtää ilman tilastollista koulutusta, että yleisesti hyväksytyt merkitsevyyden kriteerit ovat melko mielivaltaisia ja että vääriä positiivisia tulee aina jossain vaiheessa. Tämän takia tutkimuksia pitää monta kertaa replikoida, jotta päästään lähemmäs totuutta.

EP: Olen kuullut maallikoiden tekevän johtopäätöksiä, että mihinkään psykologian alan tutkimuksiin ei voi luottaa. Sanoisin, että nyt on vihdoin sellaisia tutkimuksia, joihin voi luottaa, kuten Many Labs -projektien toistuvat tulokset.

TAR: Opiskelijoiden keskuudessa melko usein kohtaamiani väärinkäsityksiä ovat muun muassa käsitykset tutkijoiden yleisestä vilpillisyydestä ja toiminnasta julkaisujen määrä ja rahoitus, ei tieteellinen laatu edellä. Keskustelu replikaatiokriisistä tuntuu siis vaikuttaneen ihmisten käsityksiin tieteen etiikasta laajemminkin. Kokeelliseen tutkimukseen liittyvät ongelmat tuntuvat heijastuvan melko usein myös muilla teoreettis-metodologisilla lähestymistavoilla tehtyjen tutkimusten kritiikkiin. Tämä ilmenee toiveina, että replikaatiokriisi ”iskisi” seuraavaksi esimerkiksi sosiaalis-konstruktionistisesta paradigmasta käsin tehtyihin tutkimuksiin, vaikka ne edustavat täysin toisenlaisia tieteenfilosofisia taustaoletuksia.

KS: Hyvin yleinen väärinkäsitys on, että kyseessä ovat uudet vaatimukset tai uudet tilastolliset menetelmät. Tietokoneiden kehittynyt laskentateho sekä internet tietysti mahdollistavat paljon sellaista, mikä aiemmin olisi ollut mahdotonta. Aina ei ole voinut esimerkiksi jakaa aineistoja internetin välityksellä tai testata tilastollisia menetelmiä simulaatiotutkimuksella. Olennaisimmat vaatimukset ja menetelmät kuitenkin perustuvat tietoon, joka on ollut olemassa ja saatavilla aivan alusta saakka ja jonka huomiotta jättämisestä on jatkuvasti varoiteltu (ks. esim. Meehl, 1990). Ei ole koskaan ollut hyvän tieteellisen käytännön mukaista käyttää tilastollisia menetelmiä miten sattuu tai jättää kertomatta seikkoja, jotka vaikuttavat merkittävästi tulosten tulkintaan. Toistotutkimukset ovat olleet hyvin olennainen osa tieteellistä metodia jo kauan ennen kuin nykyisiä tilastollisia menetelmiä alettiin käyttää. Nyt ollaan vain palaamassa niihin vaatimuksiin ja metodiseen tarkkuuteen, jotka ovat mahdollistaneet aiemmat tieteelliset läpimurrot.

MS-K: Se, että sosiaalipsykologia olisi tieteenalana jotenkin muita heikkotasoisempi. Meillä on siivottu omaa pesää paljon tarmokkaammin kuin monella muulla alalla, ja esimerkiksi lääketieteessä on myös ollut tapauksia, joissa

tutkimusaineistot ovat paljastuneet keksityiksi tai väärennetyiksi.

8. Millaisia tieteenfilosofisia kysymyksiä tai implisiittisiä taustaoletuksia replikaatiokriisikeskustelussa olet huomannut ponnahtavan esiin?

MH: Monilla tutkijoilla on ajatus siitä, että psykologiset ilmiöt ovat kompleksisia, eikä siinä mielessä tulosten edes voi odottaa replikoituvan. Silti tehdään tutkimusta menetelmillä, jotka olettavat, että tutkittavan kokonaisuuden voi pilkkoa palasiksi, tutkia erillään ja kasata palaset myöhemmin yhteen. Tällainen reduktionismi ei ikävä kyllä toimi kovin hyvin lineaarisen fysiikan ulkopuolella ja erityisesti psykologiassa (Hasselman & Bosman, 2020) – viime vuosina onkin tullut paljon uutta, niin sanottuun kompleksisuusnäkökulmaan pohjaavaa tutkimusta (Carello & Moreno, 2005; Kelty-Stephen & Wallot, 2017; Pincus, Kiefer & Beyer, 2018; Richardson, Dale & Marsh, 2014). Tämä on tärkeää, sillä yksi replikoitumattomuuden perussyistä voi olla käytettyjen mallien implisiittiset taustaoletukset (Wallot & Kelty-Stephen, 2017).

V-JI: Ei välttämättä tieteenfilosofiaa tai taustaoletuksia, mutta ei-toistuvien tulosten laajan esiintyvyyden spekuloiminen on luonnollisesti noussut esiin. Ainakin vääränlaiset kannustinjärjestelmät (*publish or perish*), tulosten erittäin voimakas kontekstiriippuvaisuus, tutkijoiden tilastollinen osaamattomuus tai halu uskoa omiin tuloksiin, koska ne kertovat miellyttäviä tai muutettavia asioita ihmisistä, on nostettu esiin selitettäessä, miksi niin monet tulokset eivät toistu.

JL: Yhteinen vastaus oikeastaan kysymyksiin 7 ja 8. Suurin taustalla oleva ongelma on pyrkimys löytää keinoja, joilla voitaisiin varmistaa, että yksittäisen julkaistun artikkelin tulos on ”totta”. Tämä unelma/tavoite on täysin utopistinen. Jopa kokeneiden tutkijoiden tuntuu olevan vaikea ymmärtää satunnaisvaihtelun voimaa eli sitä, miten paljon vaihtelua aineistoihin tulee puhtaasti otantavirheestä ja myös mittausvirheestä, puhumattakaan tilanteesta, jossa otos on harhainen ja mittarit vähän sinne päin.

Tutkimuksessa on useita virhelähteitä. Ilmiössä voi olla otantavirhettä, mittausvirhettä, se ei välttämättä ole pysyvä ajan suhteen tai siihen vaikuttaa moderoivia tekijöitä, eli tulos ei ole välttämättä samanlainen kaikissa populaatioissa. Kaikki nämä vaikuttavat tuloksen replikoitumiseen, ja lopullinen arviointi vaatii varsin pitkäjänteistä tutkimusta. Julkaistu tutkimustulos olisikin nähtävä pikemminkin lupaavana alkuna kuin päätepisteenä.

J-EL: Muut tieteenalan ongelmat ovat ehkä jääneet taka-alalle. Olettamus vaikuttaa olevan, että kunhan tästä kriisistä selvitään, niin kaikki on kunnossa ja voimme taas kertoa Totuuksia. Käsitteiden epäselvyyteen, mittarien validiteettiin, ilmiöiden mitattavuuteen ja tulosten tulkintaan liittyvät ongelmat on sivuutettu, ikään kuin olisi selvä, että tieteenala pystyy tulevaisuudessa perustelemaan olemassaolonsa, kunhan vain saamme replikaatioon liittyvät ongelmat korjattua.

EP: Vastaan on tullut kysymyksiä induktiivisesta päättelystä ja popperilaisesta falsifikationismista (Popper, 1959). Olin aikaisemmin ihmetellyt falsifikationismin ja tilastollisen päättelyn välistä ristiriitaa eli sitä, miksi psykologiassa ei pyritä tilastollisissa testeissä kumoamaan falsifikationismin mukaisesti varsinaisia mallien ennustuksia vaan niille vastakkaisia hypoteeseja (Holtz & Monnerjahn, 2017; Meehl, 1967, 1978). Koska perinteinen nollahypoteesimerkitsevyystestaus (NHST) on monitulkintainen silloin, kun tulos ei ole tilastollisesti merkitsevä, ei varsinaisia malleja tai ennustuksia voida silloin ehkä koskaan falsifioida, vaikka tulosten ei-merkitsevyys replikoitaisiin joka kerta. Mitä ei-merkitsevä tulos tarkoittaa nykyisessä käytännössä? Todistusaineistoa sille, että malli tai hypoteesi on väärä? Vai ei yhtään mitään? Kirjallisuudessa on molempia tulkintoja, eikä NHST:ssä ole yksikäsitteistä tulkintaa, koska NHST on sekoitus sisäisesti johdonmukaisista mutta keskenään osin ristiriitaisista Fisherin ja Neyman–Pearsonin lähestymistavoista (Perezgonzales, 2015). Sivujuonteena minulle on selvinnyt, että tieteellisesti kunnianhimoisempaan varsinaisten hypoteesien asettamiseen falsifikaatiouhan alle on olemassa frekventistinen menetelmä nimeltä ekvivalenssitestaus (Lakens,

Scheel & Isager, 2018). Myös Neyman–Pearson-tyyppisessä päättelyssä vaikuttaa olevan aineksia falsifioituvuuteen kykenevän tieteen tekemiseen (Perezgonzales, 2015).

TAR: Liittyen implisiittisiin taustaoletuksiin olen törmännyt ajatukseen, että kokeellinen tutkimus ja luonnontieteiden ihanteisiin tähtääminen ovat (sosiaali)psykologisen tutkimuksen tärkeimpiä ja kauneimpia muotoja.

KS: Replikaatiokriisi on lähtökohtaisesti tieteenfilosofinen. Kriisin ytimessä ovat kysymykset tietämisen rajoitteista ja siitä, miten tiede erotetaan näennäistieteestä, politiikasta ja uskonnosta. Tutkijat tekevät aina tieteenfilosofisia oletuksia, halusivat tai eivät (ks. esim. Dienes, 2008). Jos näitä oletuksia ei tutkita kriittisesti, seuraa ongelmia.

MS-K: Monet tuntuvat ajattelevan, että psykologiassa voisi havaita luonnonlakien tapaisia pysyviä ja universaaleja periaatteita. Itse ajattelen, että ihmisyhteisöt muotoutuvat hyvinkin erilaisiksi erilaisissa konteksteissa, vaikka joitakin universaaleja ilmiöitä on.

9. Replikaatio-ongelmia on havaittu useilla muilla tieteenaloilla, vaikka vuosikymmenen alkuvaiheilla ongelma paikantui julkisessa keskustelussa psykologiaan. Verrattuna muihin tieteenaloihin, miten psykologiatiede on nähdäksesi pärjännyt käytäntöjen parantamisessa tähän mennessä?

MH: En osaa sanoa muista tieteenaloista.

V-JI: Luulisin, että melko hyvin, mutta en toisaalta ole kovin tarkkaan seurannut muita aloja. Psykologialla on mielestäni myös velvoite tehdä jotain ollakseen vakavasti otettava tieteenala, ja sillä on periaatteessa mahdollisuus olla edelläkävijä näissä asioissa.

SL: Psykologia on ollut etulinjassa puuttumassa oman alansa toimintaan ja ratkaisemassa kriisiä. Mielestäni alamme tutkijat ovat tehneet todella paljon tilanteen parantamiseksi. Suuret replikaatioyritykset (Reproducibility Project, Many Labs) ovat suunnattoman arvostettavia. Lehtien muuttuneet käytännöt konkreettisesti tukevat tutkimuskulttuurin muutosta. Monilla muilla aloilla – uskoisin, että kaikilla kvantitatiiviseen tutkimukseen luottavilla aloilla – tämä sama on vielä edessä, ja toivon, että ne ottavat mallia psykologian ponnisteluista ja kehittävät uusia tapoja, joista me taas voimme oppia.

Haluan tässä nostaa erityisesti esiin yhden varhaisen replikaatioyrityksen. Olen kiitollinen kaikille tutkijoille, jotka ovat osallistuneet replikaatioprojekteihin, mutta on vaikea kuvailla, kuinka suuressa kiitollisuudenvelassa olen Stephane Doyenille ja hänen kollegoilleen, jotka tekivät ensimmäisen replikaatioyrityksen sosiaalisesta/emotionaalisesta primingista (Doyen ym., 2012). Omien tutkimusaiheideni vuoksi olin hyvin kiinnostunut sosiaalisesta primingista, ja ilman replikaatioyrityksiä olisin todennäköisesti uhrannut aikaa ja vaivaa tutkimuksiin ja suunnitelmiin, jotka perustuvat tähän ilmiöön. On mielestäni merkillepantavaa, että kun Doyen ja kollegat päättivät tutkia sosiaalisen primingin luotettavuutta, heidän suunnitelmansa ei luvannut heille mainetta, kunniaa tai julkaisua jossain arvostetuimmista lehdistä, päinvastoin – he pyrkivät suurella vaivalla tekemään replikaation toisen tutkijan jo aikaa sitten tekemästä, erittäin arvostetusta tutkimuksesta: jos he replikoisivat tutkimuksen, heidän tuloksensa olisi tiedeyhteisön silmissä vain tylsä; jos eivät, he joutuisivat vahvan vastareaktion kohteeksi, kuten kävikin. Tämä on mielestäni ihailtavaa.

JL: Tieteenä psykologia pärjää mielestäni varsin hyvin, vaikka edelleen toivoisin parantamista otantaan ja mittaamiseen liittyvissä asioissa, kuten ehkä aikaisemmissa vastauksissa on näkynyt. Samoin julkaisukäytännöt eivät varsinaisesti ole muuttuneet toivomaani suuntaan, jossa jokainen järkevä tutkimus saataisiin julkaistua jossakin tuloksesta riippumatta ja julkaisukelpoisuuden arviointi ei keskittyisi tulokseen vaan asetelman arviointiin.

Lopulta on kuitenkin sanottava, että psykologiaa vaivaa kuitenkin jonkinlainen huono itsetunto, joka on ehkä nuoren alan ongelma. Lähtökohtaisesti sitä, että vanhoja tuloksia saadaan kumottua, pitäisi pikemminkin juhlia kuin kauhistella. Se on juuri sitä tieteen itsekorjautuvuutta. Samalla on hyvä, että yhä kasvava osa tutkijoista on ollut

viime vuosina kiinnostuneempi tutkimustensa analysointimenetelmistä, ja muutamista keskusteluun liittyvistä ylilyönneistä huolimatta on hyvä, että keskustelua käydään.

J-EL: Psykologia sai muutaman vuoden etumatkaa, mutta ainakin taloustiede ja lääketiede taitavat olla kovaa vauhtia seuraamassa. Nyt olisi tilaisuus vielä hyödyntää tätä edelläkävijän asemaa ja kehittää toimintatapoja, joita muut voisivat matkia.

MO: Täytyy huomata, että psykologiassa on osa-alueita, joissa replikaatiokriisi on ollut huomattavasti pienempi ongelma, esimerkkinä havaintopsykologia. Tämä johtuu nähdäkseni siitä, että havaintopsykologiassa efektien koot ovat keskimäärin isompia kuin monilla muilla psykologian osa-alueilla ja on kenties vähemmän painetta saada huomiota herättäviä tuloksia (toki tähän on poikkeuksia). Kognitiivisessa neurotieteessä oli aikaisemmin jonkin verran ongelmia varsinkin aivokuvantamisdatan tilastollisessa testaamisessa, mutta näistä teini-iän ongelmista on nähdäkseni päästy aika hyvin eroon.

EP: Toistaiseksi psykologiassa on parannettu joissain lehdissä läpinäkyvyyttä sekä julkaisuharhaa ja *p*-hakkerointa ehkäiseviä käytäntöjä paremmin ja nopeammin kuin muualla. Taloustiede näyttää kulkevan rinnalla, koska huippulehdissä näyttää nykyään olevan sääntönä julkaista kaikki aineistot ja analyysikoodit, ja ainakin *Journal of Development Economics* julkaisee myös rekisteröityjä raportteja.

KS: Psykologian tutkimuksen reformiliikkeen kuplassa tapahtuu päivittäistä valtavaa edistystä. Kuplan ulkopuolella kaikki näyttää kuitenkin pysyvän lähes samana. Esimerkiksi lääketieteessä on otettu suurempia askelia yleisten käytäntöjen parantamiseksi. Ennakkorekisteröinti on tehty useilla tutkimusalueilla lähes pakolliseksi. Toivon, että eri tieteenalat voivat oppia toisiltaan. Monet käytetyt tilastolliset menetelmät, julkaisujärjestelmän toimintamalli, monet muut rakenteet ja tieteenfilosofiset ongelmat ovat samoja.

MS-K: Psykologia on pärjännyt oikein hyvin (katso vastaus 7).

10. Mitä tulevaisuuden vinkkejä antaisit suomalaisille tutkijoille tähän aiheeseen liittyen?

MH: Muiden tekemistä virheistä lukemisessa on se ilo, että siitä voi oppia, ja tässä mielessä kriisikeskustelu on erittäin hyvää opetusmateriaalia. Kun tekee työnsä niin hyvin kuin voi ja raportoi sen äärimmäisen läpinäkyvästi, ei tarvitse pelätä virheitä; niistäkin muut oppivat. Epävarmuuden ja turbulenssin silottelu johtaa sitä suurempaan romahdukseen, mitä pidempään sitä tehdään – tämä on itse asiassa kiinnostava yleisperiaate, joka näkyy monissa järjestelmissä metsäpalojen torjunnasta talouden vakauttamispyrkimyksiin (Taleb & Blyth, 2011).

V-JI: Ohje voi kuulostaa erikoiselta, mutta kannattaa seurata Twitterin replikaatiokeskustelua. Tai ei välttämättä niinkään keskustelua siellä, mutta sinne suoraan linkittyvää blogeissa, konferensseissa ja artikkeleissa käytävää keskustelua. Toinen käytännön vinkki on Daniël Lakensin vapaahintainen Improve your statistical inferences -verkkokurssi (https://www.coursera.org/learn/statistical-inferences), jonka hyödyllisyys ei rajoitu pelkkiin opiskelijoihin (sen voi käydä myös anonyymisti).

SL: Kannattaa olla mukana tutkimuskulttuurin muutoksessa. Se on tehty suhteellisen helpoksi esimerkiksi OSF-palvelun myötä. Aineiston avoimuus on iso mahdollisuus: aineisto ei välttämättä mene hukkaan, ellei itse löydä siitä tuloksia. On myös tärkeää hyväksyä, että todennäköisesti tiedämme vähemmän kuin olemme tähän saakka luulleet.

JL: Tulosten replikoitumista tai replikoimattomuutta ei kannata pelätä, kunhan voi raportoida tulokset omatunto puhtaana. Vanha sanonta ”jos dataa tarpeeksi kauan kiduttaa, niin se alkaa jotain tunnustamaan” pitää valitettavan hyvin paikkansa.

J-EL: Muutos on täällä, halusimme tai emme. Ei kannata jäädä taistelemaan tuulimyllyjä vastaan, vaan loikata rohkeasti muutoksen aallolle. Kuten muutosjohtaja tai konsultti sanoisi.

MO: Omia käytäntöjä mietittäessä vankka tilastollinen osaaminen auttaa, joten jos tilastolliset taidot ovat päässeet ruostumaan, kannattaa päivittää tietoja. Kannattaa ottaa selvää, mitkä lehdet hyväksyvät rekisteröityjä raportteja, ja harkita tätä ainakin joidenkin tutkimusten julkaisukanavaksi. Jos on erityisen kiinnostunut jostain efektistä ja sen toistettavuus mietityttää, on ehkä hyvä idea kokeilla toistaa efekti, koska on nykyään mahdollista julkaista tietyissä arvostetuissa tiedejulkaisuissa (esim. *eNeuro*).

EP: Olen kuullut, että työhaastatteluissa Suomessa on jo kysytty hakijoiden näkemyksiä replikaatiokriisistä. Paitsi että muut tutkijat tietävät julkaisuharhasta ja *p*-hakkeroinnista, myös maallikot alkavat olla tietoisia. Termit ovat päätyneet populaarikulttuuriin kuten populaarilehtiin, talk show'ihin ja korttipeleihin (https://www.wired.com/story/were-all-p-hacking-now/). Tutkimustuloksiamme ja -käytäntöjämme kyseenalaistavat siis muut tutkijat ja koko yhteiskunta. Tämänkin vuoksi kannattaa tutustua aihepiiriin ja keinoihin tehdä omasta tutkimuksestaan uskottavampaa. Läpinäkyvyyttä lisäävät käytännöt tulevat lehdissä lisääntymään, ja tutkimuksiin luotetaan nykyisiä julkaisutapoja enemmän, jos ne on julkaistu aineistoineen ja analyysikoodeineen rekisteröityinä raportteina ja jos niistä on useita suoria replikaatioita.

KS: Tilanne on hyvin haastava, koska tutkijoilta vaaditaan paljon jo valmiiksi. Pitäisi tuottaa nopeasti julkaisuja, hakea ja saada rahoitusta, pysyä kartalla oman alueen kehityksestä, opettaa ja tuottaa opetusmateriaalia, esiintyä hyvin, setviä byrokraattisia sotkuja, tehdä kansainvälistä yhteistyötä ja samalla säilyttää oma mielenterveytensä. Nyt pitäisi lisäksi vielä opiskella tilastotiedettä, tieteenfilosofiaa ja avoimen tieteen käytäntöjä sekä lähteä aktiivisesti vaikuttamaan tiedepolitiikkaan. Ongelmaa ei kuitenkaan voida lakaista maton alle. Tällä hetkellä yliopistojen järjestelmä yrittää pakottaa kilpailemaan julkaisuista ja rahoituksesta keinolla millä hyvänsä, mutta voimme kieltäytyä siitä ja alkaa tukea sekä auttaa toisiamme. Yhteistyöllä voimme myös muuttaa niitä ehtoja, joilla tutkimusta tehdään. Yliopiston tulisi olla turvapaikka vapaalle ja kriittiselle ajattelulle, ei julkaisujen tai innovaatioiden tehotuotantoa. Koko ajan suurempi osa tutkijoista on valmis haastamaan paitsi aiemmat tutkimuskäytännöt, myös vallitsevat rakenteet ja asenteet. Yhteistyötä tehdään koko ajan enemmän, ja se on mahdollistanut esimerkiksi replikaatiokriisin paljastaneet laajat toistotutkimusprojektit. Tässä hengessä olemme perustaneet myös Suomeen Läpinäkyvää tiedettä -verkoston, jotta kriisin ratkaisemisesta kiinnostuneet tutkijat voisivat löytää toisensa (nettisivu tulossa: läpinäkyvätiede.fi).

MS-K: Pohdi tarkasti, onko mitään syytä uskoa, että jonkin ilmiön pitäisi replikoitua. Älä tee johtopäätöksiä yksittäisten tutkimusten perusteella. Lue paljon ja monipuolisesti!

Lähteet

Arguello, P. A. (2019). Market forces influence editorial decisions. *Cortex, 113*, 363–364. doi:10.1016/j.cortex.2018.11.033.

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*(sup1), 262–270. doi:10.1080/00031305.2018.1543137.

Aschwanden, C. (2018). Psychology's replication crisis has made the field better. *FiveThirtyEight*, December 6. Haettu osoitteesta https://fivethirtyeight.com/features/psychologys-replication-crisis-has-made-the-field-better.

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature, 533*(7604), 452–454.

Bakker, M., van Dijk, A. & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science, 7*(6), 543–554. doi:10.1177/1745691612459060.

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. doi:10.1037/0022-3514.71.2.230.

Begley, C. G. & Ellis, L. M. (2012). Raise standards for preclinical cancer research. *Nature, 483*(7391), 531–533. doi:10.1038/483531a.

Bem, D. J. (2011). Feeling the future: Experimental evidence for anomalous retroactive influences on cognition and affect. *Journal of Personality and Social Psychology, 100*, 1–19.

Benjamin, D. J., Berger, J. O., Johannesson, M., Nosek, B. A., Wagenmakers, E.-J., Berk, R., … & Johnson, V. E. (2018). Redefine statistical significance. *Nature Human Behaviour, 2*(1), 6–10. doi:10.1038/s41562-017-0189-z.

Bones, A. K. (2012). We knew the future all along: Scientific hypothesizing is much more accurate than other forms of precognition – A satire in one part. *Perspectives on Psychological Science, 7*(3), 307–309. doi:10.1177/1745691612441216.

Bower, B. (2012). The hot and cold of priming: Psychologists are divided on whether unnoticed cues can influence behavior. *Science News, 181*(10), 26–29. doi:10.1002/scin.5591811025.

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. doi:10.1038/nrn3475.

Carello, C. & Moreno, M. (2005). Why nonlinear methods. Teoksessa M. A. Riley & G. C. Van Orden (toim.), *Tutorials in contemporary nonlinear methods for the behavioral sciences* (s. 1–25). Haettu osoitteesta https://nsf.gov/pubs/2005/nsf05057/nmbs/chap1.pdf.

Carp, J. (2012). The secret lives of experiments: Methods reporting in the fMRI literature. *NeuroImage, 63*(1), 289–300. doi:10.1016/j.neuroimage.2012.07.004.

Cassidy, S. A., Dimova, R., Giguère, B., Spence, J. R. & Stanley, D. J. (2019). Failing grade: 89% of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science.* doi:10.1177/2515245919858072.

Chambers, C. D. (2019). The battle for reproducibility over storytelling. *Cortex, 113*, A1–A2. doi:10.1016/j.cortex.2019.03.009.

Claesen, A., Gomes, S. L. B. T., Tuerlinckx, F. & Vanpaemel, W. (2019). *Preregistration: Comparing dream to reality* [ennakkojulkaisu]. *PsyArXiv.* doi:10.31234/osf.io/d8wex.

Cohen, J. (1962). The statistical power of abnormal-social psychological-research – A review. *Journal of Abnormal Psychology, 65*(3), 145–153. doi:10.1037/H0045186.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45*(12), 1304–1312.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science, 1*(3), 140216. doi:10.1098/rsos.140216.

Dienes, Z. (2008). *Understanding psychology as a science: An introduction to scientific and statistical inference.* Macmillan International Higher Education.

Dominus, S. (2017). When the revolution came for Amy Cuddy. *The New York Times Magazine*, October 18. Haettu osoitteesta https://www.nytimes.com/2017/10/18/magazine/when-the-revolution-came-for-amy-cuddy.html.

Doyen, S., Klein, O., Pichon, C.-L. & Cleeremans, A. (2012). Behavioral priming: It's all in the mind, but whose mind? *PLoS ONE, 7*, e29081.

Engber, D. (2017). Daryl Bem proved ESP is real. Which means science is broken. *Slate*, May 17. Haettu osoitteesta https://slate.com/health-and-science/2017/06/daryl-bem-proved-esp-is-real-showed-science-is-broken.html.

Fanelli, D. (2018). Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proceedings of the National Academy of Sciences, 115*(11), 2628–2631. doi:10.1073/pnas.1708272114.

Feilden, T. (2017). Most scientists 'can't replicate studies by their peers'. *BBC News*, February 22. Haettu osoitteesta https://www.bbc.com/news/science-environment-39054778.

Feynman, R. P. (1974). Cargo cult science. *Engineering and Science, 37*(7), 10–13. Haettu osoitteesta http://calteches.library.caltech.edu/51/2/CargoCult.pdf.

Fidler, F. & Wilcox, J. (2018). Reproducibility of scientific results. Teoksessa E. N. Zalta (toim.), The Stanford encyclopedia of philosophy (Winter 2018 Edition). Haettu osoitteesta https://plato.stanford.edu/archives/win2018/entries/scientific-reproducibility.

Flake, J. K. & Fried, E. I. (2019). Measurement schmeasurement: Questionable measurement practices and how to avoid them. *PsyArXiv.* doi:10.31234/osf.io/hs7wm.

French, C. (2012). Precognition studies and the curse of the failed replications | Professor Chris French. *The Guardian*, March 15. Haettu osoitteesta https://www.theguardian.com/science/2012/mar/15/precognition-studies-curse-failed-replications.

Galak, J., LeBoeuf, R. A., Nelson, L. D. & Simmons, J. P. (2012). Correcting the past: Failures to replicate psi. *Journal of Personality and Social Psychology, 103*(6), 933–948. doi:10.1037/a0029709.

Gelman, A. (2011). Some thoughts on academic cheating, inspired by Frey, Wegman, Fischer, Hauser, Stapel. September 12. Haettu osoitteesta https://statmodeling.stat.columbia.edu/2011/09/12/some-thoughts-on-academic-cheating-inspired-by-frey-wegman-fischer-hauser-stapel/.

Gelman, A. (2016a). What has happened down here is the winds have changed. September 21. Haettu osoitteesta https://statmodeling.stat.columbia.edu/2016/09/21/what-has-happened-down-here-is-the-winds-have-changed/.

Gelman, A. (2016b). Why does the replication crisis seem worse in psychology? *Slate*, October 3. Haettu osoitteesta https://slate.com/technology/2016/10/why-the-replication-crisis-seems-worse-in-psychology.html.

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. Department of Statistics, Columbia University.

Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics, 33*(5), 587–606. doi:10.1016/j.socec.2004.09.033.

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.

Giner-Sorolla, R. (2012). Science or art? How aesthetic standards grease the way through the publication bottleneck but undermine science. *Perspectives on Psychological Science, 7*(6), 562–571. doi:10.1177/1745691612457576.

Glasziou, P., Altman, D. G., Bossuyt, P., Boutron, I., Clarke, M., Julious, S., … & Wager, E. (2014). Reducing waste from incomplete or unusable reports of biomedical research. *The Lancet, 383*(9913), 267–276. doi:10.1016/S0140-6736(13)62228-X.

Goldacre, B., Drysdale, H., Dale, A., Milosevic, I., Slade, E., Hartley, P., … & Mahtani, K. R. (2019). COMPare: A prospective cohort study correcting and monitoring 58 misreported trials in real time. *Trials, 20*(1), 118. doi:10.1186/s13063-019-3173-2.

Goldacre, B., Drysdale, H., Marston, C., Mahtani, K. R., Dale, A., Milosevic, I., … & Heneghan, C. (2019). COMPare: Qualitative analysis of researchers' responses to critical correspondence on a cohort of 58 misreported trials. *Trials, 20*(1), 124. doi:10.1186/s13063-019-3172-3.

Halfmann, E., Bredehöft, J. & Häusser, J. A. (2020). Replicating roaches: A preregistered direct replication of Zajonc, Heingartner, and Herman's (1969) social-facilitation study. *Psychological Science, 31*(3), 332–337. doi:10.1177/0956797620902101.

Hardwicke, T. E. & Ioannidis, J. P. A. (2018). Mapping the universe of registered reports. *Nature Human Behaviour, 2*(11), 793–796. doi:10.1038/s41562-018-0444-y.

Hasselman, F. & Bosman, A. M. T. (2020). Studying complex adaptive systems with internal states: A recurrence network approach to the analysis of multivariate time series data representing self-reports of human experience. *Frontiers in Applied Mathematics and Statistics, 6*. doi:10.3389/fams.2020.00009.

Holtz, P. & Monnerjahn, P. (2017). Falsificationism is not just 'potential' falsifiability, but requires 'actual' falsification: Social psychology, critical rationalism, and progress in science. *Journal for the Theory of Social Behaviour, 47*(3), 348–362. doi:10.1111/jtsb.12134.

Huber, D. E., Potter, K. W. & Huszar, L. D. (2019). Less "story" and more "reliability" in cognitive neuroscience. *Cortex, 113*, 347–349. doi:10.1016/j.cortex.2018.10.030.

Inzlicht, M. (2016). Reckoning with the past. February 29. Haettu osoitteesta https://michaelinzlicht.com/getting-better/2016/2/29/reckoning-with-the-past.

Inzlicht, M. (2019). Transcending humanness or: Doing the right thing for science. *Cortex, 113*, 360–362. doi:10.1016/j.cortex.2018.11.032.

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), e124. doi:10.1371/journal.pmed.0020124.

Ioannidis, J. P. A. (2012). Why science is not necessarily self-correcting. *Perspectives on Psychological Science, 7*(6), 645–654. doi:10.1177/1745691612464056.

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. doi:10.1177/0956797611430953.

Kelty-Stephen, D. G. & Wallot, S. (2017). Multifractality versus (mono-) fractality as evidence of nonlinear interactions across timescales: Disentangling the belief in nonlinearity from the diagnosis of nonlinearity in empirical data. *Ecological Psychology, 29*(4), 259–299. doi:10.1080/10407413.2017.1368355.

Kerr, N. L. (1998). HARKing: Hypothesizing After the Results are Known. *Personality and Social Psychology Review, 2*(3), 196–217. doi:10.1207/s15327957pspr0203_4.

Lakens, D., Scheel, A. M. & Isager, P. M. (2018). Equivalence testing for psychological research: A tutorial. *Advances in Methods and Practices in Psychological Science, 1*(2), 259–269. doi:10.1177/2515245918770963.

Letzter, R. (2016). We talked to the scientist at the center of a brutal firestorm in the field of psychology. *Business Insider*, September 27. Haettu osoitteesta https://www.businessinsider.com/susan-fiske-methodological-terrorism-qa-2016-9?r=US&IR=T.

Lindsay, S. (2016). A commitment to replicability: An interview with the editor of Psychological Science. *APS Observer*, January. Haettu osoitteesta https://www.psychologicalscience.org/observer/a-commitment-to-improving-replicability-an-interview-with-psychological-science-editor-d-stephen-lindsay.

Makel, M. C., Plucker, J. A. & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science, 7*(6), 537–542. doi:10.1177/1745691612460688.

Meehl, P. E. (1967). Theory-testing in psychology and physics: A methodological paradox. *Philosophy of Science, 34*(2), 103–115. doi:10.1086/288135.

Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology, 46*(4), 806–834. doi:10.1037/0022-006X.46.4.806.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports, 66*, 195–244.

Munafò, M. R., Nosek, B. A., Bishop, D. V. M., Button, K. S., Chambers, C. D., Percie du Sert, N., … & Ioannidis, J. P. A. (2017). A manifesto for reproducible science. *Nature Human Behaviour, 1*, 0021. doi:10.1038/s41562-016-0021.

Nelson, L. D., Simmons, J. & Simonsohn, U. (2018). Psychology's renaissance. *Annual Review of Psychology, 69*(1), 511–534. doi:10.1146/annurev-psych-122216-011836.

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., … & Yarkoni, T. (2015). Promoting an open research culture. *Science, 348*, 1422–1425.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … & Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences, 23*(10), 815–818. doi:10.1016/j.tics.2019.07.009.

Nosek, B. A. & Lakens, D. (2014). Registered reports: A method to increase the credibility of published results. *Social Psychology, 45*(3), 137–141. doi:10.1027/1864-9335/a000192.

Nosek, B. A., Spies, J. R. & Motyl, M. (2012). Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspectives on Psychological Science, 7*(6), 615–631. doi:10.1177/1745691612459058.

Open Science Collaboration (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science, 7*(6), 657–660. doi:10.1177/1745691612462588.

Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251). doi:10.1126/science.aac4716.

Pashler, H. & de Ruiter, J. P. (2017). Taking responsibility for our field's reputation. *APS Observer*, August 31. Haettu osoitteesta https://www.psychologicalscience.org/observer/taking-responsibility-for-our-fields-reputation.

Pashler, H. & Wagenmakers, E. (2012). Editors' introduction to the special section on replicability in Psychological Science: A crisis of confidence? *Perspectives on Psychological Science, 7*(6), 528–530. doi:10.1177/1745691612465253.

Perezgonzalez, J. D. (2015). Fisher, Neyman-Pearson or NHST? A tutorial for teaching data testing. *Frontiers in Psychology, 6*. doi:10.3389/fpsyg.2015.00223.

Pincus, D., Kiefer, A. W. & Beyer, J. I. (2018). Nonlinear dynamical systems and humanistic psychology. *Journal of Humanistic Psychology, 58*(3), 343–366. doi:10.1177/0022167817741784.

Popper, K. (1959). *The logic of scientific discovery.* New York, NY: Basic Books.

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), 712–712. doi:10.1038/nrd3439-c1.

Richardson, M., Dale, R. & Marsh, K. (2014). Complex dynamical systems in social and personality psychology: Theory, modeling and analysis. Teoksessa H. T. Reis & C. M. Judd (toim.), *Handbook of research methods in social and personality psychology* (s. 251–280). Cambridge University Press.

Rodgers, J. L. & Shrout, P. E. (2017). Psychology's replication crisis as scientific opportunity: A précis for policymakers. *Policy Insights from the Behavioral and Brain Sciences.* doi:10.1177/2372732217749254.

Rosenthal, R. (1979). The file drawer problem and tolerance for null results. *Psychological Bulletin, 86*, 638–641.

Rosnow, R. L. & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44*, 1276–1284.

Rozin, P. (2001). Social psychology and science: Some lessons from Solomon Asch. *Personality and Social Psychology Review, 5*, 2–14.

Scheel, A. M., Schijen, M. & Lakens, D. (2020). An excess of positive results: Comparing the standard psychology

literature with registered reports. *PsyArXiv*. doi:10.31234/osf.io/p6e9c.

Shea, C. (2011). Fraud scandal fuels debate over practices of social psychology. *Chronicle of Higher Education*, November 13. Haettu osoitteesta https://www.chronicle.com/article/As-Dutch-Research-Scandal/129746.

Shrout, P. E. & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology, 69*(1), 487–510. doi:10.1146/annurev-psych-122216-011845.

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. doi:10.1177/0956797611417632.

Simonsohn, U. (2013). Just post it: The lesson from two cases of fabricated data detected by statistics alone. *Psychological Science, 24*(10), 1875–1888. doi:10.1177/0956797613480366.

Spellman, B. A. (2012). Introduction to the special section on research practices. *Perspectives on Psychological Science.* doi:10.1177/1745691612465075.

Spellman, B. A. (2015). A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science.* doi:10.1177/1745691615609918.

Stanford, K. (2017). Underdetermination of scientific theory. Teoksessa E. N. Zalta (toim.), The Stanford encyclopedia of philosophy (Winter 2017 edition). Haettu osoitteesta https://plato.stanford.edu/archives/win2017/entries/scientific-underdetermination/.

Stroebe, W. & Strack, F. (2014). The alleged crisis and the illusion of exact replication. *Perspectives on Psychological Science.* doi:10.1177/1745691613514450.

Sullivan, P. F. (2007). Spurious genetic associations. *Biological Psychiatry, 61*(10), 1121–1126. doi:10.1016/j.biopsych.2006.11.010.

Szucs, D. & Ioannidis, J. P. (2017). Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biology, 15*(3), e2000797. doi:10.1371/journal.pbio.2000797.

Taleb, N. N. & Blyth, M. (2011). The black swan of Cairo. *Foreign Affairs*, June. Haettu osoitteesta https://fooled-byrandomness.com/ForeignAffairs.pdf.

Valtonen, J. (2019). Tiedeyhteisön ankarat kasvukivut. *Tiedetoimittaja*, 2/2019. Haettu osoitteesta http://www.tiedetoimittajat.fi/tiedetoimittaja/tiedeyhteison-ankarat-kasvukivut/.

Van den Akker, O., Weston, S. J., Campbell, L., Chopik, W. J., Damian, R. I., Davis-Kean, P., … & Bakker, M. (2019). Preregistration of secondary data analysis: A template and tutorial [ennakkojulkaisu]. *PsyArXiv.* doi:10.31234/osf.io/hvfmr.

Vazire, S. (2018). Implications of the credibility revolution for productivity, creativity, and progress. *Perspectives on Psychological Science.* doi:10.1177/1745691617.

Vul, E., Harris, C., Winkielman, P. & Pashler, H. (2009). Puzzlingly high correlations in fMRI studies of emotion, personality, and social cognition. *Perspectives on Psychological Science, 4*(3), 274–290. doi:10.1111/j.1745-6924.2009.01125.x.

Wallot, S. & Kelty-Stephen, D. G. (2017). Interaction-dominant causation in mind and brain, and its implication for questions of generalization and replication. *Minds and Machines*, 1–22. doi:10.1007/s11023-017-9455-0.

Yarkoni, T. (2011). The psychology of parapsychology, or why good researchers publishing good articles in good journals can still get it totally wrong. Haettu osoitteesta https://www.talyarkoni.org/blog/2011/01/10/the-psychology-of-parapsychology-or-why-good-researchers-publishing-good-articles-in-good-journals-can-still-get-it-totally-wrong/.

Yong, E. (2012). Nobel laureate challenges psychologists to clean up their act. *Nature News*, October 3. doi:10.1038/nature.2012.11535.

Zajonc, R. B. (1965). Social facilitation. *Science, 149*(3681), 269–274.

Zajonc, R. B., Heingartner, A. & Herman, E. M. (1969). Social enhancement and impairment of performance in the cockroach. *Journal of Personality and Social Psychology, 13*(2), 83–92. doi:10.1037/h0028063.