Väärä tulos

Ari Hyytinen
Professori
Jyväskylän yliopiston kauppakorkeakoulu

Tutkimustulosten toistettavuus on riippumattoman vertaisarvioinnin lisäksi yksi tärkeimmistä keinoista, joiden avulla tiedeyhteisö pyrkii varmistamaan tuottamansa tiedon oikeellisuuden ja luotettavuuden. Empiirinen tutkimus on sekä itseään korjaavaa, jos virheelliset eli ei-toistettavissa olevat tutkimustulokset tulevat kumotuksi, että itseään vahvistavaa, jos oikea tulos toistetaan onnistuneesti myöhemmissä tutkimuksissa.

Tilastollisessa testauksessa lähtökohtaväitteen eli nollahypoteesin hylkäämiseen tai hylkäämättä jättämiseen voi liittyä kahdenlaisia virheitä. Tyypin I virhe tapahtuu, kun nollahypoteesi on tosi, mutta tilastollinen testi hylkää sen. Toisaalta nollahypoteesi voi olla virheellinen, mutta tilastollinen testi ei hylkää sitä. Tällöin on kyse tyypin II virheestä.

Viime aikoina tutkimustulosten toistettavuus on kyseenalaistettu monilla tieteenaloilla. Esimerkiksi lääketieteessä on käyty keskustelua siitä, ovatko useimmat alan tieteellisissä aikakauskirjoissa julkaistuista, tilastollisesti merkitsevistä – ja siis nollahypoteesin hylkäävistä – tuloksista vääriä (ks. esim. Ioannidis 2005). Tähän liittyen on pohdittu, mikä merkitys toistettavuudella on näiden ”väärien positiivisten” tulosten vähentämisessä (ks. esim. Moonesinghe, Khoury ja Janssens 2007).

Myös psykologiassa on käyty laajaa keskustelua tutkimustulosten luotettavuudesta: Huolta ovat aiheuttaneet paitsi suoranaiset tieteelliset petokset (ks. Stroebe, Postmes ja Spears 2012) ja tutkijoiden haluttomuus antaa aineistojaan muille tutkijoille uudelleentarkastelua varten myös tutkijoiden moninaiset vapausasteet tutkimusasetelman suunnittelussa ja tutkimuksen toteutuksessa (Simmons, Nelson ja Simonsohn 2011; John, Loewenstein ja Prelec 2012). Riskiksi tässä keskustelussa on koettu mm. se, että lähes mikä tahansa psykologinen väite tai mekanismi saattaa päätyä saamaan (näennäistä) empiiristä tukea. Keskustelun vilkkaudesta kertoo sekin, että *Perspectives on Psychological Science* -aikakauskirja julkaisi vuonna 2012 ao. ongelmia eri näkökulmista tarkastelevan erikoisnumeron.

Taloustieteessä toistettavuuden vaatimus ei vaikuta olevan yhtä vahva kuin muilla tieteenaloil-

la (ks. esim. Hamermesh 2007), vaikka viime keväänä keskusteltiinkin näkyvästi Carmen M. Reinhartin ja Kenneth M. Rogoffin (2010, 2012) tutkimuksien toistettavuusongelmista liittyen julkisen velan ja talouskasvun väliseen yhteyteen. Välttämättä aina ei edes ole selvää, mitä empiirisen taloustieteellisen tutkimuksen toistettavuudella tarkalleen ottaen tarkoitetaan.

Tutkimuksen replikointi voi tarkoittaa teknistä toistettavuutta, tilastollista toistettavuutta tai tieteellistä toistettavuutta (Hunter 2001 ja Hamermesh 2007). Näistä ensimmäinen viittaa tutkimuksen täsmälliseen toistamiseen alkuperäisellä aineistolla ja lähestymistavalla ja toinen tutkimuksen toistamiseen uudella otoksella, mutta käyttäen samaa menetelmää ja kohdistuen samaan populaatioon kuin alkuperäistutkimuksessa. Tieteellinen toistettavuus viittaa tässä jaottelussa puolestaan tutkimuksen uusimiseen toisenlaiseen populaatioon kohdistuvalla aineistolla, kenties eri menetelmiä hyödyntäen.

Ei varmaankaan ole väärin todeta, että useimmille empiiristä tutkimusta tekeville taloustieteilijöille toistettavuus tarkoittaa käytännössä edellä listatuista viimeisintä eli tietyn aiemman tutkimuksen ”osittaista toistamista” uudella, eri populaatiota koskevalla aineistolla ja/tai eri menetelmiä hyödyntäen. Tilastollista toistettavuutta peräänkuulutetaan taloustieteessä selvästi harvemmin.

Osalle taloustieteilijöistä ja ainakin taloustieteellisen tutkijayhteisön ulkopuolisille jonkinlainen yllätys oli se, että Reinhartin ja Rogoffin tutkimuksien toistettavuusongelmat vaikuttavat olleen pitkälti teknisluonteisia. Ne liittyvät ohjelmointivirheisiin, aineiston harkitsemattomaan valikointiin ja havaintojen epätavalliseen painottamiseen.

Teknisen toistettavuuden ongelmat eivät ole uusi asia taloustieteessä, sillä jo 1980-luvulta lähtien on ollut tiedossa ns. *Journal of Money, Credit and Banking* -aineistohankkeen tulokset. Tämän hankkeen ansiosta on voitu selvittää, miten hyvin ao. aikakauskirjassa julkaistut empiiriset tutkimustulokset ovat teknisesti toistettavissa. Dewald, Thursby ja Anderson (1986) raportoivat hankkeen ensimmäiset tulokset. He totesivat, että vaikka monien tarkasteluun päätyneiden tutkimuksien keskeiset johtopäätökset eivät välttämättä muuttuneet merkittävästi kun ne teknisesti toistettiin, ”*… inadvertent errors in published empirical articles are a commonplace rather than a rare occurrence*.” Hamermeshin (2007) mukaan Dewaldin, Thursbyn ja Andersonin raportoimat löydökset olivat osaltaan syynä siihen, että American Economic Review alkoi kiinnittää huomiota siinä julkaistujen empiiristen tutkimusten aineistojen saatavuuteen. McCullough, McGeary ja Harrison (2005) ovat sittemmin vahvistaneet samaa *Journal of Money, Credit and Banking* -aineistohanketta hyödyntäen, että empiiristen taloustieteellisten tutkimusten tulokset eivät useimmiten ole suoraan teknisesti toistettavissa (ks. myös Anderson ym. 2008).

Edellä todettu kertoo paitsi tutkimusaineistojen ja niihin liittyvien ohjelmistokoodien saatavuuteen ja uudelleenkäyttöön liittyvistä ongelmista, myös siitä, että erilaisten teknisluontoisten virheiden ja epäselvien mallinnusvalintojen mahdollisuutta on vaikea sulkea pois, kun moniulotteisia aineistoja muokataan ja käsitellään tilastollisten ohjelmien avulla ja kun tutkimuksessa hyödynnetään monimutkaisiakin ekonometrisia ja tilastollisia menetelmiä.

Kuten muutkin tieteenalat, myös taloustiede kärsii julkaisuharhasta: aikakauskirjojen toimit-

tajilla ja vertaisarviointiprosesseilla on taipumus valikoida julkaistavaksi empiirisiä tutkimuksia, joissa saadaan tilastollisesti merkitseviä tuloksia (ks. DeLong ja Lang 1992, Card ja Krueger 1995 ja Stanley 2005). Yksi konkreettinen julkaisuharhan merkki on se, että pienempiin otoksiin perustuvissa julkaistuissa tutkimuksissa raportoidaan keskimäärin suurempia vaikutuksia (kertoimia) tietylle ilmiölle tai vaikutusmekanismille. Tästä on näyttöä myös taloustieteessä. On mielenkiintoista, että julkaisuharhan aste vaikuttaa vaihtelevan tutkimusalueittain (Doucouliagos ja Stanley 2013). Julkaisuharha on järjestelmätason ongelma. Se tarkoittaa, että kenties suurikin osa julkaistuista, tilastollisesti merkittävistä tuloksista voi olla vääriä positiivisia löydöksiä (eli kärsii tyypin I virheestä).

Aikakauskirjat julkaisevat toisinaan tutkimuksia, jotka *eivät* tuota näyttöä tietystä vaikutuksesta ja jotka eivät siis hylkää nollahypoteesia. Taloustieteessä tämä liittyy De Longin ja Langin (1992) mukaan usein tilanteeseen, jossa aikaisempi empiirinen kirjallisuus on jo tuottanut näyttöä ao. vaikutuksesta tai mekanismista. Julkaisukannustin syntyy siitä, että vain tässä tilanteessa nollatuloksella vaikuttaisi olevan uutuusarvoa tutkijayhteisölle. De Long ja Lang osoittavat, että tällöin on kuitenkin hyvin todennäköistä, että ao. tutkimuksien tulokset ovat virheellisiä. Heidän tyly arvionsa on, että lähes kaikki keskeisissä taloustieteellisissä aikakauskirjoissa julkaistut ei-merkitsevät tulokset ovat vääriä. Ne eivät siis onnistu hylkäämään nollahypoteesia, joka on epätosi.

Toistettavuusvaatimus tieteellisenä kriteerinä on empiirisessä taloustieteessä ollut tähän asti epämääräinen ja monilta osin lähinnä halpaa puhetta. Tämä koskee ymmärtääkseni myös kokeellista taloustiedettä. Ainakaan toistettavuus ei ole saanut ansaitsemaansa huomiota, vaikka aina ei olekaan selvää, missä raja menee toistettavuuteen keskittyvän replikointitarkastelun ja aiemman tutkimuksen varaan rakentavan, mutta sinällään itsenäisen tutkimuksen välillä.

Painopiste empiirisen taloustieteellisen tutkimuksen luotettavuutta koskevassa viimeaikaisessa keskustelussa ei ole ollut toistettavuusongelmissa, vaan pitkälti muissa kysymyksissä, kuten mm. luonnollisia koeasetelmia ja instrumenttimuuttujia hyödyntävän soveltavan mikroekonometrian ja rakenteellisen ekonometrian eroissa ja suhteellisissa vahvuuksissa.[1]

Yllä sanotun valossa paljon julkisuutta saaneet Reinhartin ja Rogoffin tutkimuksien ongelmat asettuvat oikeaan kontekstiin: Ensinnäkin, eriasteiset toistettavuusongelmat ovat yleisempiä kuin usein luullaan. Toiseksi, kuten muillakin tieteenaloilla, empiirisessä taloustieteellisessä tutkimuksessa on paljon moniulotteisempia ja periaatteellisempia ongelmia kuin jonkin tietyn yksittäisen tutkimuksen toistettavuus. Aikakauskirjojen julkaisuprosesseihin, tutkijoiden kannustimiin, tutkimusmenettelyihin ja muihin koko tiedeyhteisöä koskeviin ongelmiin ei valitettavasti ole yksinkertaisia ratkaisuja. □

[1] *Katso esimerkiksi Journal of Economics Perspectives -aikakauskirjassa vuonna 2010 julkaistu Con out of economics -teemanumero, jossa julkaistiin mm. varsin paljon huomiota saanut Angristin ja Pischken (2010) -artikkeli, tai Journal of Economic Literature -lehden samana vuonna julkaisema Forum on the estimation of treatment effects- artikkelikokoelma, sekä Keane (2010)*

Kirjallisuus

Anderson, R. Greene, W. H., McCullough, B. D. and Vinod, H. D. (2008), "The role of data/code archives in the future of economic research", *Journal of Economic Methodology* 15: 99-119.

Angrist, J. ja Pischke, J-S. (2010), "The credibility revolution in empirical economics: How better research design is taking the con out of econometrics", *Journal of Economic Perspectives* 24: 3–30.

Card, D. ja Krueger, A. B. (1995), "Time-series minimum wage studies: a meta-analysis", *American Economic Review* 85, s. 238-243.

De Long, B. J. ja Lang, K. (1992), "Are all economic hypotheses false?", *Journal of Political Economy* 100: 1257-1272.

Dewald, W., Thursby, J. ja Anderson, R. (1986), "Replication in empirical economics: the Journal of Money, Credit, and Banking project", *American Economic Review* 76: 587-603.

Doucouliagos, C. ja Stanley T. D. (2013), "Are all economic facts greatly exaggerated? Theory competition and selectivity", *Journal of Economic Surveys* 27: 316-339.

Hamermesh. D. (2007), "Replication in economics", *Canadian Journal of Economics* 40: 715-733.

Hunter, J. (2001), "The desperate need for replications", *Journal of Consumer Research* 28: 31-43.

Ioannidis (2005), "Why most published research findings are false", *PLoS Medicine* 2: 696-701.

John L. K., Loewenstein G. ja Prelec D, (2012) "Measuring the prevalence of questionable research practices with incentives for truthtelling", *Psychological Science* 23: 524-532.

Keane, M. P. (2010), "Structural vs. atheoretic approaches to econometrics", *Journal of Econometrics* 156: 3-20.

McCullough, B. D., McGeary, K. A. and Harrison T. D. (2005), "Lessons from the JMCB Archive", *Journal of Money, Credit, and Banking* 38: 1093-1107.

Moonesinghe, R., Khoury M. J. ja Janssens, A. C. J. W. (2007), "Most published research findings are false – but a little replication goes a long way", *PLoS Medicine* 4: 218-221.

Reinhart C. M. ja Rogoff K. S. (2010), "Growth in a time of debt", *American Economic Review: Papers & Proceedings* 100: 573–578.

Reinhart C. M. ja Rogoff K. S. (2012), "Public debt overhangs: Advanced-economy episodes since 1800", *Journal of Economic Perspectives* 26: 69-86.

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

Stanley, T. D. (2005), "Beyond publication bias", *Journal of Economic Surveys* 19: 309-345.

Stroebe W., Postmes T. ja Spears R. (2012), "Scientific misconduct and the myth of self-correction in science", *Perspectives on Psychological Science* 7: 670-688.