TURUN
YLIOPISTO
Kauppakorkeakoulu

SATUNNAISTETTUJEN KONTROLLIKOKEIDEN ULKOISEN VALIDITEETIN ONGELMAT

Taloustieteen
pro gradu -tutkielma

Laatija:
Robert Pylkkänen

Ohjaaja:
Professori Heikki Kauppi

24. kesäkuuta 2024
Turku

Turun yliopiston laatujärjestelmän mukaisesti tämän julkaisun alkuperäisyys on tarkastettu Turnitin OriginalityCheck-järjestelmällä.

Pro gradu -tutkielma

Oppiaine: Taloustiede
Tekijä: Robert Pylkkänen
Otsikko: Satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmat
Ohjaaja: Professori Heikki Kauppi
Sivumäärä: 42 sivua
Päivämäärä: 24. kesäkuuta 2024

Tämän kirjallisuuskatsauksen tarkoituksena on arvioida satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmia. Tutkimuskysymys on rajattu koskemaan satunnaistettuja kontrollikokeita, jotka on suunnattu politiikkatoimien arviointiin. Jotta satunnaistetut kontrollikokeet ovat poliittisen päätöksenteon kannalta informatiivisia, tulee päätelmien olla sekä sisäisesti että ulkoisesti validit. Tutkimuksen ulkoiseen validiteettiin kiinnitetään monesti kuitenkin sisäistä validiteettia vähemmän huomioita. Toisinaan saatetaan myös virheellisesti ajatella, että sisäisesti validit tulokset riittäisivät sellaisinaan ulkoisen validiteetin saavuttamiseen. Kun tarkoituksena on päätelmien yleistäminen, nousee ulkoisen validiteetin ongelmien arvioiminen keskeiseen asemaan. Ongelmat voidaan luokitella kolmeen tekijään: ympäristöjen välisiin eroihin, yleisen tasapainon vaikutuksiin ja yksilöiden käyttäytymisen muutoksiin. Ulkoisen validiteetin ongemia ei usein voida suoraan ratkaista satunnaistettujen kontrollikokeiden avulla. Ongelmat voidaan kuitenkin huomioida ja niiden olemassaoloa voidaan testata. Tämä edellyttää monesti siirtymistä suuntaan, jossa satunnaistettuja kontrollikokeita hyödynnetään yhdessä muiden menetelmien ja aiemman teorian kanssa.

Avainsanat: Satunnaistetut kontrollikokeet, Ulkoinen validiteetti

Sisällys

1 Johdanto

Kokeellisen tutkimuksen suosio kasvaa taloustieteen tutkimuksessa. Kasvavasta suosiosta kertoo se, että vuoden 2019 taloustieteen Nobel-palkinto jaettiin kokeellisten tutkimusmenetelmien käytöstä. Kasvu on ollut suurta erityisesti kehittyvissä maissa, joissa kokeiden toteuttaminen on onnistunut suhteellisen pienillä budjeteilla (ks. Duflo 2006). Kokeiden voidaan odottaa lisääntyvän myös tulevaisuudessa, sillä digitalisaatio laskee jatkuvasti niiden toteuttamiseen liittyviä kustannuksia (Athey & Imbens 2017, 75).

Satunnaistetut kontrollikokeet nähdään useiden tutkijoiden keskuudessa luotettavimpana menetelmänä, kun tarkoituksena on johtaa kiinnostuksen kohteena oleva syy-seuraussuhde. Luotettavuuden taustalla on ratkaisu valintaharhaan, jonka menetelmä tarjoaa. Valintaharha viittaa tilanteeseen, jossa tutkimuksen koe- tai kontrolliryhmiin valikoituu tutkimusyksiköitä tiettyjen ominaisuuksien perusteella. Tutkittavan käsittelyn (engl. treatment) todellinen vaikutus voi tällöin sekoittua muihin taustalla vaikuttaviin tekijöihin, mikä puolestaan johtaa virheellisiin johtopäätöksiin käsittelyn vaikutuksesta. Valintaharha koskee erityisesti havaintotutkimuksia, sillä näiden aineistot kerätään usein erilaisilla kyselytutkimuksilla, kuten hallinnollisilla rekistereillä (Stock & Watson 2020, 49−50).

Validiteetin määritelmä on tärkeä, kun arvioitavana on päätelmien luotettavuus. Validiteetti kertoo, kuinka hyvin tutkimuksessa käytetty mittausmenetelmä mittaa juuri sitä tutkittavan ilmiön ominaisuutta, jota on tarkoituskin mitata (Tilastokeskus: Käsitteet). Usein validiteetti jaetaan kahteen osa-alueeseen: sisäiseen ja ulkoiseen validiteettiin. Päätelmät ovat sisäisesti validit, jos havaittu yhteisvaihtelu kahden muuttujan välillä heijastaa muuttujien välistä kausaaliyhteyttä. (Shadish ym. 2002, 53.) Päätelmät ovat ulkoisesti validit, jos tietylle perusjoukolle tehdyt kausaalipäätelmät yleistyvät muualle. Yleistyksen kohteet voivat sisältää eri perusjoukot, lopputulemat tai ympäristöt. (Athey & Imbens 2017, 79.)

Shadishin ym. (2002, 53) mukaan sisäisesti validien päätelmien saavuttamiseksi on täytettävä kolme ehtoa. Ensinnäkin tutkijan on osoitettava, että tutkittava käsittely edeltää lopputulemaa. Toisena tutkijan tulee todistaa, että käsittelyn ja

lopputuleman välillä on tilastollinen yhteys. Kolmantena tutkijan on osoitettava, että mikään muu selitys muuttujien väliselle yhteydelle ei ole todennäköinen. Satunnaistetut kontrollikokeet tarjoavat ratkaisun jokaiseen kohtaan, sillä käsittelyn jakaminen on tutkijan hallussa ja sen toteuttaminen arpomalla poistaa valintaharhan. Satunnaistettujen kontrollikokeiden sisäinen validiteetti on yleensä siis hyvä, kun ne toteutetaan asianmukaisesti.

Sisäisestä validiteetista poiketen, satunnaistettujen kontrollikokeiden ulkoinen validiteetti on usein heikko. Heikkoon ulkoiseen validiteettiin voivat vaikuttaa monet tekijät. Tyypillisesti kokeet esimerkiksi rajoittuvat yksittäisiin ympäristöihin ja ajanhetkiin, jolloin tulosten yleistettävyys myös kärsii. Taloustieteen sovelluksissa ulkoisen validiteetin ongelmat korostuvat, sillä yksilöiden preferensseissä ja rajoitteissa saattaa esiintyä merkittävää vaihtelua, kun ympäristöt ja ajanhetket vaihtelevat (Imbens 2010, 403).

Ulkoisen validiteetin arvioiminen ei kaikissa tilanteissa ole tarpeen. Ulkoista validiteettia ei esimerkiksi tarvitse edellyttää, kun tarkoituksena on tietyn teorian testaaminen tai vaikutuksen havainnollistaminen (ks. Deaton & Cartwright 2018). Ulkoisen validiteetin merkitys kuitenkin kasvaa, kun tutkimuksen tuloksia halutaan soveltaa laajemmin tai käyttää polittisen päätöksenteon tukena. Tilanne korostuu akateemisessa tutkimuksessa, jossa useimpien satunnaistettujen kontrollikokeiden tarkoituksena on politiikkasuositusten esittäminen tai tutkimustulosten yleistäminen. (Peters ym. 2018.)

Satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmiin ei toistaiseksi ole olemassa täydellistä ratkaisua. Ulkoisen validiteetin ongelmat kuitenkin tunnistetaan ja niihin on ehdotettu ratkaisuja (ks. Duflo ym. 2006; Banerjee & Duflo 2009). Ratkaisuvaihtoehtoja ovat esimerkiksi analyysin keskittäminen yleistettävissä olevien mekanismien tutkimiseen, joiden avulla voidaan selittää syitä tutkimuksen johtopäätöksille (ks. Deaton 2010; Deaton & Cartwright 2018). Toinen vaihtoehto on arvioida suoraan syitä sille, miksi kokeen käsittelyvaikutus vaihtelee eri ympäristöissä (ks. Athey & Imbens 2017).

Tämä tutkielma on kirjallisuuskatsaus, jonka tarkotuksena on perehtyä satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmiin. Tutkielmassa ar-

vioidaan ulkoisen validiteetin merkitystä, kun satunnaistettujen kontrollikokeiden avulla muodostettuja päätelmiä on tarkoitus käyttää tutkimuspohjaisen päätöksenteon tukena. Ulkoisen validiteetin ongelmat määritellään Duflon ym. (2006) tutkimuksen avulla ja ratkaisu ongelmien testaukseen esitetään Hotzin ym. (2005) tutkimuksella. Tämän tutkielman tavoitteena on kasvattaa tietoisuutta ulkoisen validiteetin ongelmiin liittyen ja esitellä lähestymistavat, joilla ongelmien olemassaoloa voidaan arvioida.

Tutkielma etenee seuraavasti. Luvussa kaksi esitetään kausaalipäättelyn kannalta keskeinen teoria ja pohjustetaan satunnaistettujen kontrollikokeiden toimintaa. Luvussa kolme arvioidaan satunnaistettujen kontrollikokeiden luotettavuutta ja tutustutaan ulkoisen validiteetin ongelmiin. Luvussa neljä tarkastellaan Hotzin ym. (2005) tutkimusta, jossa ulkoisen validiteetin ongelmien esiintymistä testataan empiirisesti. Luku viisi toimii yhteenvetona tutkielman aiheille.

2 Syy-seuraussuhteen arviointi

Taloustieteen empiiristen tutkimusten tavoitteena on usein johtaa eri muuttujien välinen syy-seuraussuhde. Kausaalikysymykset voivat liittyä esimerkiksi kuluttujien, yritysten tai valtion toimintaan, jolloin tavoitteena on löytää vastaus siihen, miten tietty ohjelma, toimenpide tai käsittely vaikuttaa kiinnostuksen kohteena olevaan lopputulemaan. Klassisia esimerkkejä taloustieteen kirjallisuudessa esiintyvistä kausaalikysymyksistä ovat luokkakoon vaikutus oppimistuloksiin (ks. Krueger 1999) sekä koulutuksen vaikutus palkkatuloihin (ks. Card 1999).

Syy-seuraussuhteen mittaaminen edellyttää kontrafaktuaalin eli vaihtoehtoisen tilan arviointia. Koulutuksen kausaalivaikutus palkkatuloihin viittaa esimerkiksi palkkatason kasvuun, jonka yksilö saa hankittuaan lisää koulutusta. Yksilön tulot on tällöin havaittava kahdessa eri tilanteessa, jotta kysymykseen saadaan vastaus. Tulot on havaittava sekä tilanteessa, jossa yksilö hankkii lisäkoulutusta, että tilanteessa, jossa yksilö ei hanki lisäkoulutusta. Ongelmana on kuitenkin se, että tiettynä ajanhetkenä on mahdollista havaita vain toinen yksilön lopputulemista. Holland (1986, 947) kutsuu tilannetta kausaalipäättelyn perimmäiseksi ongelmaksi.

2.1 Potentiaalinen lopputulema ja valintaharha

Syy-seuraussuhteen mittaamista hankaloittaa se, että emme voi suoraan havaita, mitä olisi tapahtunut ilman tiettyä toimenpidettä. Ongelmaa ja sen ratkaisua voidaan havainnollistaa *potentiaalisten lopputulemien* käsitteen (engl. potential outcome) avulla, jonka Rubin (1974) on alunperin esitellyt.

Angrist ja Pischke (2009) käsittelevät tilannetta tutkimuksessaan. Tarkastelussa on asetelma, jossa sairaalakäynnin vaikutusta arvioidaan yksilön i terveyteen Y_i. Sairaalahoitoa merkitään binäärisellä satunnaismuuttujalla $W_i = \{0, 1\}$ ja jokaisella yksilöllä on kaksi potentiaalista terveydentilaa:

$$\textit{Potentiaalinen lopputulema} = \begin{cases} Y_i(0), & \text{jos } W_i = 0 \\ Y_i(1), & \text{jos } W_i = 1. \end{cases}$$

Potentiaalinen lopputulema $Y_i(1)$ kuvaa yksilön terveyttä, kun tämä menee sairaalaan. Potentiaalinen lopputulema $Y_i(0)$ kuvaa yksilön terveyttä tilanteessa, jossa tämä ei olisi mennyt sairaalaan.

Tavoitteena on määrittää sairaalakäynnin kausaalivaikutus yksilön i terveyteen. Kausaalivaikutuksen mittaaminen edellyttää yksilön potentiaalisten terveydentilojen $Y_i(1)$ ja $Y_i(0)$ vertaamista. Kyseessä on tilanne, joka on mahdollista toteuttaa, jos yksilön molemmat terveydentilat havaitaan samanaikaisesti. Tällöin yksilön havaittu terveys voidaan kirjoittaa potentiaalisten lopputulemien avulla:

$$Y_i = Y_i(W_i) = Y_i(0) + (Y_i(1) - Y_i(0))W_i,$$

jossa erotus $Y_i(1) - Y_i(0)$ on kiinnostuksen kohteena oleva kausaalivaikutus. Vain toinen yksilön potentiaalisista lopputulemista kuitenkin realisoituu.

Useimmissa tapauksissa sairaalahoidon kausaalivaikutukselle ei saada luotettavaa estimaattia, kun samaa yksilöä verrataan eri ajanhetkinä. Tämä johtuu siitä, että muut yksilön terveyteen vaikuttavat tekijät ovat saattaneet muuttua ajan myötä. Tästä syystä johtuen sairaalahoidon vaikutusta tulee mitata siten, että sairaalahoidossa olleiden yksilöiden keskiarvoterveyttä verrataan sellaisen kontrolliryhmän keskiarvoterveyteen, jossa yksilöt eivät ole olleet sairaalahoidossa. Ehtona on, että kontrolliryhmän tulee koostua yksilöistä, joiden lopputulemat ilman sairaalahoitoa ovat samankaltaiset sairaalahoidossa olleiden kanssa. (ks. Duflo ym. 2006, 3899.)

Tavoitteena on mitata sairaalahoidon keskimääräinen käsittelyvaikutus perusjoukkossa, jossa yksilöt jakaantuvat sairaalahoidossa olleisiin ja sairaalahoidossa olemattomiin:

$$E[Y_i(1) - Y_i(0)],$$

jossa potentiaaliset terveydentilat $Y_i(1)$ ja $Y_i(0)$ ovat satunnaismuuttujia ja operaattori $E[\cdot]$ viittaa odotusarvoon.

Käsittelyn keskimääräinen vaikutus voidaan estimoida siten, että molemmista ryhmistä otetaan ensin terveydentilojen keskiarvot. Tämän jälkeen ryhmien välisiä keskiarvoterveyksiä verrataan toisinsa. Kun otos lähestyy ääretöntä, keskiar-

voterveyksien välinen erotus konvergoituu muotoon:

$$E[Y_i|W_i = 1] - E[Y_i|W_i = 0] = E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0], \quad (1)$$

jossa $E[Y_i|W_i = w]$ ja $E[Y_i(W_i)|W_i = 1]$ kuvaavat satunnaisten muuttujien ehdollisia odotusarvoja käsittelyn $W_i = w$ suhteen.

Sairaalahoidon vaikutuksesta ei välttämättä saada täysin oikeaa kuvaa vain odotettuja terveydentiloja vertaamalla. Tämä havaitaan, kun ehdollinen odotusarvo $E[Y_i|W = 1]$ summataan ja vähennetään yhtälöön (1):

$$\begin{aligned} \underbrace{E[Y_i|W_i = 1] - E[Y_i|W_i = 0]}_{\text{Keskimääräisten terveydentilojen havaittu ero}} &= \underbrace{E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 1]}_{\text{Keskimääräinen käsittelyvaikutus altistuneille}} + \\ &\quad \underbrace{E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0]}_{\text{Valintaharha}}. \end{aligned} \quad (2)$$

Odotettujen terveydentilojen välinen havaittu ero jakautuu nyt kahteen osaan: kiinnostuksen kohteena olevaan kausaalivaikutukseen ja valintaharhaan.

Yhtälön (2) ensimmäinen termi,

$$E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 1] = E[Y_i(1) - Y_i(0)|W_i = 1],$$

kuvaa sairaalahoidon keskimääräistä kausaalivaikutusta niille yksilöille, jotka joutuivat sairaalahoitoon. Operaattori $E[Y_i(1)|W_i = 1]$ kuvaa sairaalahoitoon joutuneiden odotettua terveyttä ja operaattori $E[Y_i(0)|W_i = 1]$ kuvaa sairaalahoitoon joutuneiden terveydentilaa hypoteettisessa tilanteessa, jossa yksilöt eivät olisi olleet sairaalahoidossa.

Yhtälön (2) toinen termi,

$$E[Y_i(0)|W_i = 1] - E[Y_i(0)|W_i = 0],$$

on valintaharha. Termi kuvaa terveydentilan $Y_i(0)$ keskimääräistä eroa sairaalaan joutuneiden ja sairaalaan joutumattomien välillä. Valintaharha on sairaalakäynteihin liittyvän esimerkin tapauksessa negatiivinen, sillä sairaalahoitoon hakeutuvat ensisijassa ne yksilöt, jotka ovat alunperin jo kontrolliryhmään kuuluvia yksilöitä sairaampia. Lisäksi on mahdollista, että tilanne pysyy samana, vaikka sairaalakäynneillä olisi todellisuudessa ollut yksilöiden terveydentilaa parantava vaikutus.

Valintaharhan koko voi olla niin suuri, että se peittää täysin positiivisen käsittelyvaikutuksen. Valintaharhan suuruutta on yleisesti mahdoton määrittää, sillä ehdollista odostusarvoa $E[Y_i(1)|W_i = 0]$ ei havaittu. Tämän takia on myös hankalaa muodostaa arviota siitä, mikä on valintaharhan osuus lopputulemien välisestä havaitusta erosta. (Duflo ym. 2006, 3901.) Useiden empiiristen tutkimusten tavoitteena on valintaharhan poistaminen ja siten sanoa jotain muuttujien, kuten W_i kausaalivaikutuksesta (Angrist & Pischke 2009, 15).

2.2 Satunnaistaminen ratkaisuna valintaharhaan

Valintaharha voidaan ratkaista satunnaistetun kokeen avulla. Satunnaistettussa kokeessa N yksilön kokoinen otos valitaan ensin kiinnostuksen kohteena olevasta perusjoukosta, jonka jälkeen kokeellinen otos jaetaan satunnaisesti koe- ja kontrolliryhmiin. Koeryhmä altistetaan tämän jälkeen käsittelylle ($W_i = 1$) ja kontrolliryhmä jätetään ilman käsittelyä ($W_i = 0$). Lopuksi ryhmien välisiä havaittuja lopputulemia, Y_i, verrataan keskenään. (Duflo ym. 2006, 3901.)

Angrist ja Pischke (2009) hahmottavat ratkaisua valintaharhaan. Kun käsittely, W_i, määrätään yksilöille satunnaisesti, on käsittely, W_i, riippumaton yksilöiden potentiaalisista lopputulemista. Tämän seurauksena koe- ja kontrolliryhmiin sijoitetut yksilöt eroavat toisistaan ehdollista odotusarvoa tarkasteltaessa vain yksilöiden vastaanottamissa käsittelyissä:

$$E[Y_i(0)|W_i = 1] = E[Y_i(0)|W_i = 0]. \tag{3}$$

Kun yhtälön (3) yhtäsuuruus on voimassa, voidaan yhtälön (1) odotettujen terveydentilojen välinen erotus kirjoittaa muotoon:

$$\begin{aligned} E[Y_i|W_i = 1] - E[Y_i|W_i = 0] &= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 0] \\ &= E[Y_i(1)|W_i = 1] - E[Y_i(0)|W_i = 1]. \end{aligned} \tag{4}$$

Koska yksilöt on ohjattu sairaalahoitoon arvalla, sievenee yhtälö (4) entisestään:

$$\begin{aligned} E[Y_i|W_i = 1] - E[Y_i|W_i = 0] &= E[Y_i(1) - Y_i(0)|W_i = 1] \\ &= E[Y_i(1) - Y_i(0)]. \end{aligned} \tag{5}$$

Sairaalahoidon vaikutus sairaalahoidossa oleville on nyt sama kuin sairaalahoidon vaikutus satunnaisesti valittulle potilaalle. Valintaharha onnistutaan siis poistamaan.

2.3 Satunnaistetut kokeet ja käsittelyn jakomekanismi

Satunnaistetuksi kontrollikokeeksi kutsutaan kokonaisuutta, jossa yksilöt jaetaan satunnaisesti koe- ja kontrolliryhmiin. Koeryhmään ohjatut saavat tällöin tutkittavan käsittelyn, ja kontrolliryhmän kuuluvat eivät saa käsittelyä. (Stock & Watson 2020, 48.) Kuten kohta (5) osoittaa, hyvin suunniteltu ja toteutettu satunnaistettu kontrollikoe tuottaa harhattoman estimaatin käsittelyn vaikutukselle tutkittavassa otoksessa. Estimaatit ovat tällöin sisäisesti validit. (ks. Duflo ym. 2006, 3902.)

Satunnaistettu kontrollikoe voidaan toteuttaa eri tavoin riippuen siitä, miten tutkittava käsittely arvotaan tutkimusyksiköille. Imbens ja Rubin (2015) määrittelevät, että käsittelyn jakomekanismin (engl. assignment mechanism) tulee täyttää neljä oletusta. Ensinnäkin sen funktionaalisen muodon tulee olla tiedossa ja tutkijan hallinnassa. Toisena jokaisella yksiköllä tulee olla positiivinen todennäköisyys päätyä koe- tai kontrolliryhmään. Kolmantena käsittelyn jakomekanismin tule olla riippumaton yksilöiden potentiaalisista lopputulemista. Neljäntenä yksilön todennäköisyyteen saada käsittely ei tule vaikuttaa muiden yksilöiden potentiaaliset lopputulemat.

Bernoullin kokeet (engl. Bernoulli trials) tarjoavat klassisen esimerkin käsittelyn jakomekanismista. Bernoullin koetta voidaan havainnollistaa kolikon heiton avulla, jossa kolikon puoli määrittää tutkimusyksikön, i, ... , N, vastaanottaman käsittelyn: kruunalla yksilö ohjataan koeryhmään ja klaavalla kontrolliryhmään. Käsittelyn jako on tällöin satunnainen tapahtuma, jossa käsittelyn vastaanottamisen todennäköisyys on kaikille sama:

$$Pr(W_i = 1) = 0.5.$$

Kolikonheitot ovat toisistaan riippumattomia tapahtumia, jolloin käsittelyn vastaanottamisen yhteistodennäköisyys on niiden yksittäisten todennäköisyyksien tulo:

$$Pr(W_1 = w_1, \dots, W_N = w_N) = 0.5^N$$

Vaikka Bernoullin kokeet toteuttavat neljä käsittelyn jakomekanismiin liittyvää oletusta, eivät Bernoullin kokeet ole kaikkein ihanteellisin satunnaistamisen muoto. Tämä johtuu positiivisesta todennäköisyydestä tilanteelle, jossa kaikki tutkimusyksiköt määrätään joko koe- tai kontrolliryhmään. (Imbens & Rubin 2015, 50.)

Täysin satunnaistetut kokeet (engl. completely randomized experiments) tarjoavat toisen vaihtoehdon, jolla käsittelyn arpominen voidaan toteuttaa. Täysin satunnaistetussa kokeessa kiinteä lukumäärä tutkimusyksiköitä, N_t, valitaan satunnaisesti N tutkimusyksikön kokoisesta perusjoukosta muodostamaan tutkimuksen koeryhmä. Loput perusjoukkoon kuuluvat tutkimusyksiköt, $N_c = N - N_t$, määrätään kontrolliryhmään. Käsittelyn vastaanottamisen todennäköisyys on tällöin yhtä suuri kaikille perusjoukon jäsenille:

$$Pr(W_i = 1) = \frac{N_t}{N}.$$

Eri vaihtoehtoja, joilla tutkimusyksiköt voidaan määrätä koeryhmään, on $\frac{N!}{N_t!(N-N_t)!}$ kappaletta. Kaikki toimeksiannot ovat yhtä todennäköisiä, jolloin käsittelyn vastaanottamisen yhteistodennäköisyys on:

$$Pr(W_1 = w_1, \ldots, W_N = w_N) = \begin{cases} \left(\frac{N!}{N_t!(N-N_t)!}\right)^{-1} & \text{jos } \sum_{i=1}^{N} W_i = N_t, \\ 0 & \text{muutoin.} \end{cases}$$

Täysin satunnaistetut kokeet ovat suosittuja eri käytännön sovelluksissa. Suosion taustalla ovat menetelmän yksinkertaisuus ja takuu siitä, että koe- ja kontrolliryhmään jaetaan riittävä määrä tutkimusyksiköitä. (Imbens & Rubins 2015, 50−51.)

Täysin satunnaistettujen kokeiden periaatteita voidaan soveltaa myös monimutkaisemmissa tutkimusasetelmissa. Yksi suosittu yleistys on ositus (engl. stratification), jossa tutkimusyksiköt jaetaan aluksi kerroksiin (engl. strata) potentiaalisten lopputulemien kannalta tärkeiden kovariaattien perusteella. Tämän jälkeen täysin satunnaistettu koe toteutetaan erikseen jokaisessa kerroksessa. Tällä tavalla voidaan rajata pois sellaiset toimeksiannot, jotka hyväksyttäisiin täysin satunnaistetuissa kokeissa, mutta tarjoavat suhteellisen vähän informaatiota. Tämän kaltaisia toimeksiantoja ovat esimerkiksi tapaukset, jossa kaikki miehet ovat koeryhmässä ja kaikki naiset kontrolliryhmässä. (Imbens & Rubin 2015, 51.)

3 Satunnaistettujen kontrollikokeiden luotettavuus

Satunnaistettuja kontrollikokeita on perinteisesti pidetty kaikkein luotettavimpana menetelmänä, kun tarkoituksena on kausaalisten johtopäätösten muodostaminen (Athey & Imbens 2017, 78). Imbens (2010, 401) esimerkiksi kirjoittaa, että satunnaistetut kokeet ovat luotettavuutensa suhteen ylivertaisia muihin menetelmiin verrattuna, kunhan kiinnostuksen kohteena oleva kausaalikysymys mahdollistaa satunnaistettujen kokeiden toteuttamisen. Banerjee ja Duflo (2009, 151) puolestaan mainitsevat, että satunnaistetut kokeet mahdollistavat sellaisten parametrien estimoinnin, joita ei ilman käsittelyn arpomista olisi mahdollista arvioida.

Satunnaistetut kontrollikokeet eivät sovellu kaikkiin tilanteisiin. Erityisesti eettiset ja kustannuksiin liittyvät kysymykset hankaloittavat usein satunnaistettujen kontrollikokeiden toteuttamista. Myöskään makrotalouden kausaalikysymyksiä ei aina voida ratkaista satunnaistetuilla kontrollikokeilla. (Athey & Imbens 2017, 78.) Vaikka satunnaistettujen kontrollikokeiden toteuttaminen ei aina ole mahdollista, on niillä silti tärkeä rooli kausaalipäätelyssä. Tämä johtuu siitä, että ihanteellisesti toteutettu satunnaistettu kontrollikoe tarjoaa käsitteen, joka toimii määritelmänä kausaalivaikutukselle. (Stock & Watson 2020, 48.)

3.1 Satunnaistetut kontrollikokeet ja validiteetti

Satunnaistettujen kontrollikokeiden luotettavuus perustuu kahteen keskeiseen ominaisuuteen. Ensimmäinen ominaisuus liittyy kontrolliin, joka tutkijalla on käsittelyn jakomekanismista. Kontrolli mahdollistaa koe- ja kontrolliryhmien tasapuolisen vertailun ja poistaa valintaharhana tunnetun ongelman. (Athey & Imbens 2017, 78.) Toinen ominaisuus liittyy satunnaistettujen kontrollikokeiden hyviin tilastollisiin ominaisuuksiin. Kun tietyt oletukset täyttyvät, tuottavat satunnaistetut kontrollikokeet tarkkoja ja luotettavia estimaatteja käsittelyvaikutuksen koolle (engl. treatment effect size) ja todennäköisyydelle, jolla todellinen vaikutus kuuluu tietyn luottamusvälin sisälle. (Shadish ym. 2002, 13.)

Mukaan voidaan tuoda validiteetin määritelmä, kun tutkimuksen luotettavuutta halutaan arvioidaan laajemmin. Valideetti tarkoittaa päätelmien likimääräistä

totuutta ja sen määrittäminen edellyttää arvion muodostamista kokonaisuudesta, jolla muodostettujen päätelmien todenmukaisuutta tai virheellisyyttä voidaan tukea. Arviointi tapahtuu esimerkiksi empiiristen havaintojen, aiemman tutkimuskirjallisuuden sekä aiheeseen liittyvän teorian avulla. Validiteetti on päättelyyn liittyvä ominaisuus, joka ei suoraan yleisty tiettyihin malleihin tai menetelmiin. Tämä johtuu siitä, että sama malli voi tuottaa enemmän tai vähemmän valideja johtopäätöksiä olosuhteista riippuen. (Shadish ym. 2002, 34.)

Validiteetti jaetaan usein osa-alueisiin, joihin eri menetelmät voivat vaikuttaa samanaikaisesti. Satunnaistettujen kontrollikokeiden kannalta tärkeimpiä validiteetin osa-alueita ovat sisäinen ja ulkoinen validiteetti. Menetelmän ympärillä käytävä keskustelu liittyy tyypillisesti näiden kahden väliseen arvotukseen. Keskeisenä huolenaiheena on, että satunnaistettujen kontrollikokeiden sisäinen validiteetti on usein hyvä, mutta ulkoinen validiteetti heikko. (Imbens 2010, 417.)

Sisäinen validiteetti on pitkään nähty tutkijoiden keskuudessa tärkeimpänä validiteetin osa-alueena. (ks. Athey & Imbens 2017, 79). Yksi tekijä sisäisen validiteetin ensisijaisuuden taustalla on Campbellin ja Stanleyn (1963, 5) kirjoitus, jossa tutkijat mainitsevat sisäisen validiteetin olevan kokeiden vähimmäisvaatimus, jota ilman kokeilla on vain vähän arvoa. Shadishin ym. (2002, 97) mukaan lausunto on yksi tutkimusmetodologian viitatuimmista repliikeista, joka on vahvistanut sisäisen validiteetin asemaa ensisijaisena validiteetin osa-alueena.

Sisäisen validiteetin ensisijaisuus näkyy siinä, että ulkoisen validiteetin arviointiin käytetään tutkimuksissa monesti sisäistä validiteettia vähemmän huomiota (ks. Peters ym. 2018). Ulkoinen validiteetin merkitys on tilannekohtainen. Sitä ei tarvitse edellyttää, kun tutkimuksen tavoitteena on osoittaa tietyn vaikutuksen olemassaolo tai vaikutusmekanismin toiminta. Kokeiden tuloksista ollaan usein kuitenkin kiinostuneita alkuperäisen koeasetelman ulkopuolella. Tämä luo ristiriidan yksittäisten kokeiden tuottaman kausaalitiedon ja tutkijoiden tavoittelemien yleistyvämpien kausaalipäätelmien välille. (Shadish ym. 2002, 19.)

Ulkoisen validiteetin arvostus on ollut viimeisten vuosien aikana kasvussa taloustieteen tutkimuskirjallisuudessa. Yleistyvä ajatus tutkijoiden keskuudessa on, ettei vain yhden tyyppisen validiteetin huomioiminen ole perusteltua, kun tutki-

muksen päätelmiä hyödynnetään esimerkiksi poliitikkatoimien arviointiin. Tämän kaltaisessa tilanteessa ulkoinen validiteetti nähdään yhtä tärkeänä sisäisen validiteetin kanssa. Ulkoisen validiteettin ongelmien arviointi nousee tällöin tärkeään asemaan. (ks. Imbens 2010; Manski 2013.)

3.2 Ulkoisen validiteetin ongelmat

Satunnaistettujen kontrollikokeiden heikkoon ulkoiseen validiteettiin on monia syitä. Usein kokeet esimerkiksi rajoittuvat tiettyyn ympäristöön ja ajanhetkeen, ja niissä tutkitaan vain yksittäistä käsittelyä (Shadish ym. 2002, 18). Lisäksi kokeeseen osallistuminen edellyttää tyypillisesti yksilöiden omaa suostumusta, jolloin kokeeseen osallistuvat yksilöt saattavat poikkeavat niistä yksilöistä, jotka kokeeseen eivät ole osallistuneet. (Athey & Imbens 2017, 79.) Ulkoisen validiteetin ongelmat ovat erityisen haasteellisia taloustieteen sovelluksissa. Tämä johtuu yksilöiden preferensseistä ja rajoitteista, jotka voivat vaihdella merkittävästi kontekstista riippuen. (Imbens 2010, 417.)

3.2.1 Tulosten ympäristöriippuvuus

Duflo ym. (2006) ovat jakaneet ulkoisen validiteetin ongelmat kolmeen luokkaan. Esimmäinen ongelmista liittyy ympäristöjen välisiin eroavaisuuksiin. Ongelma koskee tilannetta, jossa tietyn kokeen avulla hankitut tulokset yleistetään alkuperäisestä ympäristöstä toiseen. Ympäristöt voivat poiketa esimerkiksi maantieteellisen lokaation, ajanhetken tai perusjoukkoon kuuluvien yksilöiden suhteen. Yksilöiden erot voivat perustua havaittuihin tai havaitsemattomiin tekijöihin ja myös kokeiden käsittelyt voivat hieman poiketa toisistaan. (Athey & Imbens 2017, 80.) Koska sijaintien välillä voi esiintyä merkittäviäkin eroja, ei satunnaistettujen kontrollikokeiden tuloksia voida suoraan yleistää muualle (Banerjee & Duflo 2009, 159−160).

Tulosten ympäristöriippuvuuteen vaikuttavat tekijät voidaan jakaa kolmeen osa-alueeseen. Näistä ensimmäinen liittyy kokeen toteutustapaan: tietty ohjelma voidaan toteuttaa niin yksityiskohtaisesti, että sen myöhempi replikointi on suu-

remmassa skaalassa mahdotonta. Pilottihankkeet kuuluvat usein tähän kategoriaan, sillä ne muodostetaan tyypillisesti laadukkaissa olosuhteissa ja äärimmäisen huolellisesti. (Duflo ym. 2006, 3953.) Myös toimeenpanevan organisaation koko saattaa heikentää tutkimuksen ulkoista validiteettia, sillä estimoitu käsittelyvaikutus voi heijastaa kokeen toteuttajalle ominaisia piirteitä. Ongelmaa kutsutaan toimeenpanijan vaikutukseksi (engl. implemeter effetct) ja se korostuu, kun toimeenpanevan organisaation koko pienenee. (Banerjee & Duflo 2009, 160.)

Toinen ympäristöstä riippuva tekijä liittyy tutkittavaan perusjoukkoon: kun kokeellinen ohjelma laajennetaan uuteen ympäristöön, voi perusjoukko uudessa ympäristössä poiketa alkuperäisen ympäristön perusjoukosta. Perusjoukkoon kuuluvien yksilöiden ikäjakaumat voivat esimerkiksi olla erilaiset. Jos ikäjakaumilla on yhteys kokeen lopputuloksiin tai ohjelman vaikutukseen, voivat ohjelmien keskiarvovaikutukset poiketa ympäristöjen välillä. (Hotz ym. 2005, 243.) Ongelma ei rajoitu ainoastaan satunnaistettuihin kontrollikokeisiin, sillä kaikki empiiriset tutkimukset antavat tietoa tutkittavasta otoksesta. Tutkimusten tuloksia voidaan tällöin yleistää vain tietyin oletuksin. Satunnaistettujen kontrollikokeiden tapauksessa ongelma kuitenkin korostuu, sillä nämä totetutetaan usein logistisista syistä johtuen suhteellisen pienillä alueilla. (Duflo ym. 2006, 3953.)

Kolmas ympäristöstä riippuva tekijä liittyy ohjelman eri versioihin. Suunnitteilla olevan ohjelman tulee olla riittävän samankaltainen alkuperäisen kokeen kanssa, jotta kokeen tuloksia voidaan yleistää. Laadullisia eroja voi syntyä, jos toinen vertailtavista ohjelmista on paremmin toteutettu (Stock & Watson 2020, 481). Laadullisia eroja voi myös esiintyä, jos ohjelmat eivät ole täysin samanlaiset. Koulutusohjelma voi esimerkiksi yhdessä sijainnissa perustua lähiopetukseen, kun taas toisessa sijainnissa yksilöt voivat saada työnhakuun liittyvää avustusta. (Hotz ym. 2005, 243.)

Osa ympäristövaikutuksista voidaan pyrkiä huomioimaan jo kokeen suunnittelu- ja toteutusvaiheessa. Ohjelmien väliset erot voidaan osittain huomioida, kun kokeen toteutuksesta ja sen eri vaiheista toimitetaan riittävästi informaatiota. Ulkoinen validiteetti tietyn perusjoukon suhteen voidaan puolestaan maksimoida siten, että kokeen sijaintipaikat valitaan ensin arpomalla ja

tutkimusyksiköt sen jälkeen arpomalla koe- ja kontrolliryhmiin. Ensimmäistä vaihetta ei usein kuitenkaan toteuteta kokeiden toteuttamiseen liittyvistä käytännön rajoitteista johtuen. (Duflo ym. 2006, 3953.)

Vaikka ympäristövaikutuksiin ei usein ole suoraa ratkaisua, voidaan niiden olemassaoloa kuitenkin arvioida. Tähän tarkoitukseen suositellaan usein replikointitutkimuksia (ks Duflo ym. 2006). Replikointitutkimusten avulla saadaan arvio tulosten yleistettävyydestä, kun ylimääräisiä kokeita toteutetaan eri sijainneissa ja olosuhteissa. (Banerjee & Duflo 2009.) Onnistuneet replikoinnit eivät suoraan takaa tutkimukselle ulkoista validiteettia tai anna takeita tulevien replikointien onnistumisesta. Epäonnistuneet replikoinnit eivät kuitenkaan tee alkuperäisistä tuloksista hyödyttömiä, sillä replikointitutkimukset tarjoavat hyvän välineen arvioida syitä sille, miksi tulokset vaihtelevat eri ympäristöissä. (Deaton & Cartwright 2018, 10.)

3.2.2 Yleisen tasapainon vaikutukset

Toinen Duflon ym. (2006) luokittelema ulkoisen validiteetin ongelma liittyy tilanteeseen, jossa pieni ja tilapäinen kokeellinen ohjelma muutetaan pysyväksi ja laajalle levinneeksi ohjelmaksi. Ongelmaa kutsutaan yleisen tasapainon vaikutuksiksi (engl. general equilibrium effects). Yleisen tasapainon vaikutukset syntyvät, kun ohjelman kasvattaminen muuttaa taloudellista ympäristöä riittävästi, jotta alkuperäisen kokeen tulokset eivät ole enää yleistettävissä. (Stock & Watson 2020, 481.)

Yleisen tasapainon vaikutukset ovat seurausta yksilöiden välille muodostuvista vuorovaikutussuhteista. Tarkastellaan esimerkiksi politiikkatoimenpidettä, jossa yliopiston lukukausimaksuja alennetaan. Jos kiinnostuksen kohteena on määrittää lukukausimaksujen alentamisen vaikutus palkkoihin, voidaan satunnaistetun kontrollikokeen avulla määrittää keskiarvovaikutus tietyn yliopiston opiskelijoille. Havaittu estimaatti soveltuu koko talouden kuvaamiseen, jos yksilöiden palkkoihin ei vaikuta muiden taloudessa valmistuneiden yliopisto- tai lukio-opiskelijoiden lukumäärä. Sama huomio pätee tilanteeseen, jossa palkat liikkuvat talouden eri tiloissa täysin samansuuntaisesti. (Heckman ym. 1999, 28−29.)

Kun toimenpidettä arvioidaan kansallisella tasolla, on todennäköistä, että ai-

empaa halvempi koulutus kasvattaa yliopisto-opiskelijoiden lukumäärää. Samalla se kuitenkin pienentää opiskelijoiden suhteellista palkkatasoa. Ne opiskelijat, jotka ottavat tilanteen huomioon, eivät ilmottaudu ollenkaan yliopisto-opetukseen. Kokeellinen ohjelma ei huomioi tilannetta, sillä se ei havaitse kokeen ulkopuolisia yksilöitä. Tämän seurauksena kokeen osallistujamäärä on sen todellista tasoa suurempi, jolloin kokeen tuottama keskiarvovaikutus saattaa olla harhaanjohtava. (Heckman ym. 1999, 28−29.)

Yleisen tasapainon vaikutuksiin ei ole olemassa täydellistä ratkaisua. Siihen liittyvät ongelmat voidaan kuitenkin pyrkiä huomioimaan kokeen toteutuksessa. Crepónin ym. (2013) tutkimus tarjoaa esimerkin tilanteesta. Tutkimus sijoittui Ranskan työmarkkinoille ja siinä keskityttiin työnhakijoiden avustuspalveluihin, jotka oli suunnattu korkeakoulutetuille nuorille. Koealue sisälsi 235 julkista työttömyysvirastoa, jotka ovat hajallaan kymmenellä Ranskan hallinnollisella alueella. Koeasetelma muodostettiin siten, että ensin jokaiselle alueelle arvottiin tietty prosenttiosuus (0, 25, 50, 75, 100). Tämän jälkeen jokaisen alueen koeryhmään arvottiin prosenttiosuutta vastaava lukumäärä kelvollisia työnhakijoita. Yleisen tasapainon vaikutuksia arvioitiin muodostettujen kontrolliryhmien avulla. Tämä tapahtui vertaamalla niitä sijainteja, joissa kukaan ei päätynyt koeryhmään, niihin sijanteihin, joissa osa työntekijöistä valittiin koeryhmään. Tutkimuksen tulosten perusteella työttömät nuoret hyötyivät ohjelmaan osallistumisesta. Hyödyt olivat kuitenkin väliaikaisia ja ne esiintyivät osittain ohjelmaan osallistumattomien työnhakijoiden kustannuksella. Yleisen tasapainon vaikutusten seurauksena ohjelman nettohyödyt olivat siis erittäin pienet.

Kuten Créponin ym. (2013) tutkimus osoittaa, yleisen tasapainon vaikutukset antavat syyn suosia suurempia tutkimuksia pienempien tutkimusten sijaan. Yleisen tasapainon vaikutuksia voidaan tässä tapauksessa tarkastella ulkoisvaikutusten kaltaisena ilmiönä. Kuten ulkoisvaikutuksilla, osa yleisen tasapainon vaikutuksista voidaan havaita, jos havaintoyksikkö (engl. unit of observation) on riittävän kattava. Satunnaistetut kontrollikokeet eivät aina kuitenkaan sovellu tähän tarkoitukseen. Tämä johtuu siitä, että tasapainovaikutusten havaitseminen saattaa jossain tapauksissa edellyttää koeasetelman toteuttamista kansallisella tai jopa

kansainvälisellä tasolla, mikä itsessään on haastavaa. (Duflo ym. 2006, 3951.)

3.2.3 Hawthorne- ja John Henry -vaikutukset

Kolmas Duflon ym. (2006) luokittelema ulkoisen validiteetin ongelma liittyy tilanteeseen, jossa koehenkilöt tiedostavat olevansa osana koetta ja muuttavat tämän takia käyttäytymistään. Koeryhmään kuuluvien yksilöiden käyttäytymisen muutoksia kutsutaan Hawthorne-vaikutuksiksi ja kontrolliryhmään kuuluvien yksilöiden käyttäytymisen muutoksia John Henry -vaikutuksiksi. Koeryhmässä saatetaan esimerkiksi työskennellä normaalia kovemmin, koska koeryhmään kuuluminen nähdään positiivisena asiana ja tarkkailun alaisuudessa oleminen tiedostetaan. Kontrolliryhmässä tekemisen laatu voi puolestaan heikentyä, jos yksilöt olisivat mielummin olleet osana koeryhmää ja kontrolliryhmään kuuluminen ei herätä vastaavaa kiinnostusta. (Duflo ym. 2006, 3951.)

Hawthorne- ja John Henry -vaikutukset ovat ongelmallisia, sillä ne muokkaavat arvioitavana olevaa ohjelmaa. Kun ohjelma muuttuu, voidaan satunnaistetun kontrollikokeen avulla edelleen tuottaa sisäisesti validit tulokset, mutta vain muuttuneelle ohjelmalle. Tutkimuksen tulokset eivät toisin sanoen yleisty enää muuttuneen ohjelman ulkopuolelle. (Heckman & Vytlacil 2007, 5066.)

Hawthorne- tai John Henry -vaikutukset eivät ole vain satunnaistettuja kontrollikokeita koskeva ongelma, sillä myös havaintotutkimuksissa yksilöiden käyttäytyminen voi muuttua. Esimerkiksi kouluille tarjottavat tuotantopanokset voivat tilapäisesti kohottaa opettajien ja opiskelijoiden motivaatiota, mikä puolestaan voi johtaa parempiin suorituksiin lyhyellä aikavälillä. Havaintotutkimukset, jotka tutkivat tuotantopanosten lisäämistä, saattavat tällöin yhtä lailla altistua yksilöiden käyttäytymisen muutoksille. Satunnaistetuissa kontrollikokeissa ongelma on kuitenkin erityisesti läsnä, sillä yksilöt saattaavat tiedostaa olevansa arvioitavana. (Duflo ym. 2006, 3951−3952.)

Hawthorne- ja John Henry -vaikutukset voidaan minimioida muodostamalla koeasetelma siten, että tutkija ja koehenkilöt eivät tiedä, kuka kuuluu koeryhmään ja kuka kontrolliryhmään. Tämän kaltaista asetelmaa ei usein kuitenkaan ole mahdollista toteuttaa taloustieteen käytännön sovelluksissa. Tämän takia on-

gelmien arvioinnin merkitys kasvaa. (Stock & Watson 2020, 480.)

Hawthorne- ja John Henry -vaikutuksia voidaan arvioida eri tavoin. Duflon ym. (2012) tutkimuksessa tämä tapahtui siten, että kokeen vaikutusten arviointia jatkettiin varsinaisen kokeen päättymisen jälkeen. Ashrafin ym. (2006) tutkimuksessa ongelma taas lähestyttiin muodostamalla kaksi koeryhmää, jotka saivat toisistaan hieman poikkeavan käsittelyn. Tutkimusten toimenpiteet eivät poista ongelmia, mutta tarjoavat keinon niiden huomiointiin. Tällöin voidaan paremmin arvioida yksilöiden käyttäytymisen muutosten osuutta kokeessa havaittuun vaikutukseen.

3.3 Satunnaistettujen kontrollikokeiden käyttökohteet

Satunnaistettujen kontrollikokeiden tuottamia tuloksia kutsutaan ajoittain kultaiseksi standardiksi, kun vertailukohteena ovat muut ekonometriset menetelmät (Deaton 2010, 438). Satunnaistettuihin kontrollikokeisiin kohdistuu paikoitellen myös kritiikkiä. Iso osa kritiikistä liittyy satunnaistettujen kontrollikokeiden heikkoon ulkoiseen validiteettiin. Osansa kritiikistä saavat myös keskiarvovaikutuksen mittaaminen[1] ja varsinaisten kokeiden kohtaamat käytännön ongelmat, jotka saattavat heikentää päätelmien sisäistä validiteettia (ks. Deaton 2010; Bédécarrats ym. 2020).

Banerjeen ja Duflon (2009, 152) mukaan satunnaistettujen kontrollikokeiden kohtaama kritiikki on suurimmaksi osaksi hyödyllistä. He huomauttavat kritiikin usein kuitenkin jättävän huomioimatta pääsyyt sille, miksi kokeellinen tutkimus on herättänyt niin valtavasti kiinnostusta taloustieteellisessä tutkimuksessa. Monet ongelmista ovat yleisiä kaikille mikrotutkimuksille (engl. microevaluation), mutta satunnaistettujen kontrollikokeiden yhteydessä ne nostetaan useammin esille. Todennäköinen selitys ilmiölle on se, että suurin osa muista yleisistä ongelmista ratkaistaan, kun käsittely jaetaan tutkimusyksiköille satunnaisesti. (Banerjee & Duflo 2009, 159.)

[1] Jossain tapauksissa ideaalia voisi olla muiden havaintosuureiden, kuten mediaanin, prosenttipisteen tai varianssin hyödyntäminen. Satunnaistetut kontrollikokeet eivät kuitenkaan sovellu tähän ilman merkittäviä lisäoletuksia. (Deaton 2010, 439.)

Satunnaistettujen kontrollikokeiden ulkoista validiteettia koskevat huolenaiheet liittyvät usein siihen, ettei tutkimuksissa huomioida ulkoisen validiteetin ongelmia riittävästi. Ulkoisen validiteetin arvioimiseen käytetään tyypillisesti huomattavasti vähemmän resursseja kuin sisäisen validiteetin arvioimiseen eikä perusteluja tulosten yleistettävyydelle välttämättä esitetä. Toisinaan tutkijoiden keskuudessa saatetaan myös virheellisesti ajatella, että sisäisesti validin kokeen tulokset pätevät automaattisesti tai sellaisenaan myös muualla. (Deaton & Cartwright 2018.)

Petersin ym. (2018) tutkimus havainnollistaa tilanteen laajutta. Tutkimuksessa tutkittiin, miten ulkoisen validiteetin ongelmat on huomioitu taloustieteen pääjournaaleissa[2]. Tutkijat analysoivat kaikki tutkimukset vuosilta 2009−2014, joissa käytettiin satunnaistettuja kontrollikokeita. Tulosten perusteella ulkoisen validiteetin ongelmia ei huomioitu suurimmassa osassa tutkimuksista. Monet tutkimuksista eivät myöskään tarjonneet riittävää informaatiota, jotta ongelmien olemassaoloa olisi voitu arvioida. Usean tutkimuksen tarkoituksena oli tästä huolimatta tulosten yleistäminen.

Tulosten yksinkertainen yleistäminen ei satunnaistettujen kontrollikokeiden tapauksessa ole useinkaan perusteltua, jos ulkoisen validiteetin ongelmia ei ole ensin huomioitu (ks. Deaton & Cartwright 2018). Deaton (2010) korostaa, että tuloksille tulee löytää jokin selitys, jotta ne yleistyät lokaalin ympäristönsä ulkopuolelle. Satunnaistettuja kontrollikokeita ei tyypillisesti ole kuitenkaan suunnattu tähän käyttötarkoitukseen. Tästä syystä Deaton (2010) painottaa analyysin suuntaamista yleistettävissä olevien mekanismien tutkimiseen, jotka selittävät, miksi ja missä yhteyksissa tietyn hankkeen voidaan odottaa toimivan.

Gautierin ym. (2018) tutkimus toimii esimerkkinä tilanteesta. Tutkimuksessa yhdistyivät käsittelyvaikutuksen empiirinen estimointi sekä makrotalouden etsintämallit (engl. search model). Tutkimuksessa analysoitiin työttömille työnhakijoille

[2]Aineistoon kuuluvat tieteelliset aikakausjulkaisut ovat *American Economic Review*, *Quarterly Journal of Economics*, *Econometrica*, *Economic Journal*, *Review of Economic Studies*, *Review of Economics and Statistics*, *Journal of Political Economy* ja *American Economic Journal: Applied Economics*

suunnattua aktivointiohjelmaa sekä mahdollisten yleisen tasapainon vaikutusten olemassaoloa ja suuruutta. Tutkijat hyödynsivät kahden Tanskan eri maakunnissa toteutetun satunnaistetun kokeen aineistoa ja havaintoaineistoa, joka sisälsi työttömät työntekijät muista Tanskan maakunnista. Havaintoaineiston tarkoituksena oli toimia vertailujoukkona kokeissa olleille kontrolliryhmille, mikä mahdollisti yleisen tasapainon vaikutusten testauksen. Empiirisiä havaintoja käytettiin tämän jälkeen yleisen tasapainon etsintämallin (engl. equilibrium search model) estimoimiseen. Etsintämallin perusteella aktivointiohjelmalla oli todellisuudessa yksilöiden hyvinvointia laskeva vaikutus. Mallissa hyvinvointia laskivat kasvanut työttömyysaste, aktivointiohjelmaan liittyvien julkisten menojen nousu ja työttömien työntekijöiden lisääntyneet työnhakukustannukset.

Ulkoisen validiteetin ongelmien huomioinnilla on tärkeää merkitys, kun tutkimuksen tuloksia ollaan kiinnostuneita laajentamaan alkuperäisen koeasetelman ulkopuolelle. Tulosten yksinkertaisella yleistämisellä satunnaistettujen kontrollikokeiden käyttökohteet helposti yliarvioidaan. Samalla mahdolliset käyttökohteet myös aliarvioidaan, sillä satunnaistettujen kontrollikokeiden tuloksia voidaan käyttää laajasti alkuperäisen koeasetelman ulkopuolella, esimerkiksi hypoteesien muodostamiseen, ymmärtämiseen ja testaamiseen. Monesti tämä edellyttää siirtymistä suuntaan, jossa satunnaistettujen kontrollikokeita käytetään yhdessä muiden menetelmien ja aiemman teorian kanssa. (Deaton & Cartwright 2018.)

4 Ympäristövaikutusten testaaminen

Ulkoinen validiteetti on satunnaistettujen kontrollikokeiden keskeinen huolenaihe. Ulkoisen validiteetin ongelmien merkitys korostuu, kun satunnaistettuja kontrollikokeita käytetään osana tutkimuspohjaista päätöksentekoa. Päätös poliittisen ohjelman käyttöönotosta perustuu usein arvioon sen todennäköisestä vaikutuksesta. Monesti tämänkaltainen arvio tehdään hyödyntämällä aineistoa samankaltaisesta ohjelmasta, joka on toteutettu aikaisemmin tai eri sijainnissa. Arvioiden luotettavuuden kannalta on tällöin tärkeää, että mahdolliset ulkoisen validiteetin ongelmat huomioidaan. (Hotz ym. 2005, 242−243.)

Tutkimuksen ulkoiseen validiteettiin voivat vaikuttaa useat tekijät. Ympäristöjen väliset erot ovat yleinen selitys tilanteelle, jossa kokeellisen ohjelman tulokset vaihtelevat eri sijainneissa. Ratkaisuksi ympäristövaikutuksiin on kehitetty erilaisia lähestymistapoja, joista yhden tarjoaa Hotzin ym. (2005) tutkimus. Tutkimuksen tavoitteena on ennustaa uuden koulutusohjelman keskiarvovaikutus käyttämällä aineistoja aiemmin toteutetuista kokeellisista ohjelmista. Seuraavissa alaluvuissa Hotzin ym. (2005) tutkimus käydään yksityiskohtaisesti läpi.

4.1 Keskiarvovaikutuksen ennustaminen

Hotzin ym. (2005) tutkimuksen tavoitteena on ennustaa uuden koulutusohjelman keskiarvovaikutus. Ennusteet muodostetaan kokeellisten aineistojen avulla, jotka ovat peräisin aiemmin toteutetuista koulutusohjelmista. Tutkijat huomioivat kaksi mahdollista tekijää, jotka voivat hankaloittaa ennusteiden muodostamista. Molemmat ongelmista liittyvät ympäristövaikutuksiin. Ensinnäkin yksilöiden ominaisuuksien jakaumissa voi esiintyä poikkeavuuksia eri perusjoukkojen välillä. Toisena ohjelmat voivat erota toteutustapojensa suhteen, vaikka ne nimellisesti olisivatkin samankaltaisia.[3] Tutkijat arvioivat ongelmien empiiristä merkitystä analysoimalla neljän satunnaistetun kokeen aineistoa, jotka toteutettiin Yhdysvaltojen eri osissa 1980-luvulla.

[3]Tutkijat huomioivat myös yleisen tasapainon vaikutukset, mutta olettavat, ettei yksilöiden välillä ole vuorovaikutussuhteita.

Tutkimuksen teoriakehys muodostuu yksilön potentiaalisten lopputulemien avulla. Tutkimuksen aineisto on suuresta perusjoukosta poimittu N tutkimusyksikön kokoinen satunnaisotos. Jokainen tutkimusyksikkö, i = 1, 2, ... , N, on peräisin jommastakummasta sijainnista, jota merkitään indikaattorilla $D_i \in \{0, 1\}$. Tässä $D_i = 0$ kuvaa alkuperäistä sijaintia ja $D_i = 1$ uutta sijaintia. Jokaisella tutkimusyksiköllä on kaksi potentiaalista lopputulemaa: $Y_i(1)$ kuvaa lopputulemaa tilanteessa, jossa tutkimusyksikkö i vastaanottaa harjoittelun ja $Y_i(0)$ kuvaa lopputulemaa, kun tutkimusyksikkö i ei vastaanota harjoittelua. Tutkimusyksiköiden välillä ei ole vuorovaikutussuhteita. Lisäksi tutkimusyksiköiden vastaanottamat käsittelyt ovat homogeeniset. Tutkimusyksikön vastaanottamaa käsittelyä merkitään indikaattorilla $W_i \in \{0, 1\}$, jossa $W_i = 1$ vastaa harjoittelua ja $W_i = 0$ kontrolliryhmään kuulumista. Vektorilla X_i kuvataan kovariaattien tai ennen käsittelyä havaittavien muuttujien joukkoa. Tutkimusyksikölle i havaittava lopputulema on:

$$Y_i \equiv Y_i(W_i) = W_i \cdot Y_i(1) + (1 - W_i) \cdot Y_i(0).$$

Tutkimuksen tavoitteena on estimoida harjoittelun keskiarvovaikutus perusjoukossa, jossa $D_i = 1$:

$$\tau_1 = E[Y_i(1) - Y_i(0)|D_i = 1]. \tag{6}$$

Parametri (6) estimoidaan N havainnon avulla. Sijaintiin $D_i = 0$ kuuluvien tutkimusyksiköiden osalta havaitaan tällöin kovariaatit X_i, sijainti D_i, käsittely W_i ja varsinainen lopputulema Y_i. Sijaintiin $D_i = 1$ kuuluville tutkimusyksiköille havaitaan vain kovariaatit X_i ja sijainti D_i.

Tutkimuksen teoriakehyksessä muodostetaan kolme oletusta. Ensimmäisen oletuksen mukaan tutkimusyksiköt, jotka kuuluvat sijaintiin $D_i = 0$, jaettiin satunnaisteti koe- ja kontrolliryhmiin:

$$W_i \perp (Y_i(0), Y_i(1))|D_i = 0, \tag{7}$$

jossa (7) on käsittelyn W_i ja potentiaalisten lopputulemien $Y_i(1)$ ja $Y_i(0)$ välinen ehdollinen riippumattomuus, kun annettuna on $D_i = 0$. Alkuperäisessä sijainnissa toteutetun harjoittelun keskimääräinen vaikutus saadaan tällöin estimoitua, kun verrataan koe- ja kontrolliryhmien odotettuja tuloksia:

$$E[Y_i|W_i = 1, D_i = 0] - E[Y_i|W_i = 0, D_i = 0] = E[Y_i(1) - Y_i(0)|D_i = 0],$$

jossa $E[Y_i|W_i = w, D_i = d]$ on lopputuleman Y_i ehdollinen odotusarvo käsittelyn $W_i = w$ ja sijainnin $D_i = d$ suhteen ja $E[Y_i(1) - Y_i(0)|D_i = 0]$ on harjoittelun keskiarvovaikutus perusjoukossa, jossa $D_i = 0$.

Harjoittelun keskimääräinen vaikutus saadaan yksinkertaisimmillaan estimoida sijaintiin $D_i = 1$ kuuluville tutkimusyksiköille, kun sijainnilla ole vaikutusta lopputuloksiin. Sijainnin ja lopputulemien välinen riippumattomuus on mahdollista taata tutkimusasetelmassa vain, kun perusjoukkoon kuuluvat tutkimusyksiköt arvotaan eri sijainteihin. Tämä on kuitenkin oletuksena epärealistinen, sillä yksilöt todennäköisemmin valitsevat oman asuinpaikkansa.

Sijainnin satunnaisuuteen liittyvää oletusta voidaan lieventää ennen käsittelyä havaittavilla muuttujilla. Tähän liittyy tutkimuksen teoriakehikon toinen oletus:

$$D_i \perp (Y_i(0), Y_i(1))|X_i.$$

Oletuksen mukaan sijainti D_i on riippumaton tutkimusyksiköiden potentiaalisista lopputulemista, kun kovariaatit X_i on kiinnitetty. Tutkijat nimittävät oletusta ”ei makrovaiktuksia”-oletukseksi. Kun oletus on voimassa, harjoittelun ehdollinen keskiarvovaikutus kovariaattien suhteen ei enää riipu sijainnista:

$$E[Y_i(1) - Y_i(0)|X_i = x, D_i = 1] = E[Y_i(1) - Y_i(0)|X_i = x, D_i = 0]. \quad (8)$$

Kolmas oletus tutkimuksen teoriakehikossa liittyy kovariaattien jakaumiin. Oletuksen mukaan kovariaattijakaumat ovat uudessa ja vanhassa sijainnissa täysin päällekäiset (engl. complete overlap). Kaikilla kovariaattivektorin X_i arvoilla pätee tällöin ehto:

$$\delta < Pr(D_i = 1|X_i = x) < 1 - \delta,$$

jollekin $\delta > 0$ ja kaikille kovariaattivektorin X_i mahdollisille arvoille. Ehdon mukaan tapahtuman $D_i = 1$ todennäköisyys on kiinnitetyillä kovariaattivektorin arvoilla positiivinen. Oletus takaa sen, että kaikilla kovariaattivektorin arvoilla voidaan löytää tutkimusyksiköitä alkuperäisen sijainnin osapopulaatiosta (engl. sub-population).

Kun kaikki kolme oletusta ovat yhtä aikaa voimassa, yhtälön (8) toisesta termistä voidaan ottaa odotusarvo uuden sijainnin kovariaattien jakaumalla:

$$E[Y_i(1) - Y_i(0)|D_i = 1] = E[E[Y_i(1) - Y_i(0)|X_i = x, D_i = 0]|D_i = 1],$$

jossa yhtäsuuruus seuraa iteroitujen odotusarvojen laista (engl. Law of Iterated Expectation). Koska käsittely on alkuperäisessä sijainnissa jaettu tutkimusyksiköille arpomalla, voidaan yhtälö (6) kirjoittaa muotoon:

$$\tau_1 = E[E[Y_i|W_i = 1, D_i = 0, X_i] - E[Y_i|W_i = 0, D_i = 0, X_i]|D_i = 1]. \quad (9)$$

Uuden sijainnin keskiarvovaikutus saadaan nyt estimoitua vanhalle sijainnille estimoidun keskiarvovaikutuksen avulla.

Estimaattorin (9) validiteettiin liittyy kaksi mahdollista uhkaa, jotka voivat heikentää ennusteiden luotettavuutta. Ensimmäinen ongelmista on seurausta makrovaikutuksista. Tutkijat viittaavat termillä sellaisten muuttujien vaikutuksiin, joiden arvot ovat vakioita tietyn sijainnin tai alasijaintien (engl. sub-location) sisällä. Koska makrovaikutukset poikkeavat sijaintien välillä, ne estävät kovariaattijakaumien päällekkäisyyteen liittyvän oletuksen toteutumisen. Tämä taas johtaa tilanteeseen, jossa yhtälön (9) yhtäsuuruus ei ole enää voimassa.

Tutkijat ehdottavat kolmea strategiaa ratkaisuna makrovaikutuksiin. Ensimmäisenä strategiana on käyttää alkuperäisinä sijainteina sellaisten sijaintien toteutuksia, jotka ovat ominaisuuksiltaan ja ajanhetkeltään mahdollisimman samankaltaisia kiinnostuksen kohteena olevan sijainnin kanssa. Toisena strategiana on kerätä mahdollisimman paljon tietoa kovariaateista, jotka voivat toimia korvikkeena (engl. proxy) sijaintien ja ajanhetkien välisille eroille. Tutkijoiden mukaan kiinnostuksen kohteena olevien tulosten aiemmat arvot soveltuvat erityisen hyvin tähän tarkoitukseen. Kolmantena strategiana on hyödyntää estimoinnissa useampaa sijaintia kahden sijaan. Tämä mahdollistaa sen, että sijaintitason ominaisuudet voidaan kontrolloida hyödyntämällä mallipohjaista muokkausta (engl. model-based adjustment).

Toinen estimaattorin (9) validiteettiin liittyvistä ongelmista koskee tilannetta, jossa yksilöiden vastaanottamissa käsittelyissä esiintyy sijaintikohtaisia eroja. Tutkijoiden mukaan on harvinaista, jopa koulutusohjelmien satunnaistetuissa arvioinneissa, että kaikki yksilöt vastaanottaisivat tietyssä ohjelmassa täysin identtisen käsittelyn. Tyypillisesti koeryhmään kuuluville määrätään joukko palveluita erilaisten lisäseulontojen ja haastatteluiden perusteella. Toimenpiteet ja näiden

pohjalta muodostettavat toimeksiannot saattavat erota eri sijainneissa, mikä puolestaan johtaa heterogeenisuuteen yksilöiden vastaanottamissa käsittelyissä.

Käsittelyiden heterogeenisuuteen liittyvää ongelmaa voidaan havainnollistaa teoreettisesti. Oletuksena on, että kontrolliryhmään kuuluvat yksiköt vastaanottavat kaikissa sijainneissa saman käsittelyn. Koeryhmässä tilanne on toinen, sillä kolutusohjelma voi koostua K + 1 koulutusosiosta. Jokaiselle koulutusosiolle $t \in T = \{0, 1, ..., K\}$ ja jokaiselle tutkimusyksikölle $i \in 1, ..., N$ on olemassa potentiaalinen lopputulema $Y_i(t)$. Tutkimusyksikön vastaanottamaa koulutusosiota merkitään indikaattorilla $\tilde{T} \in T$. Kun $W_i = 1$, tutkija havaitsee alkuperäisen toimeksiannon, mutta ei tutkimusyksikön vastaanottamaa koulutusosiota. Kun $W_i = 0$, tutkimusyksiköt jaetaan satunnaisesti kontrolliryhmään, jossa nämä eivät saa koulutuspalveluja.

Alkuperäisessä sijainnissa tutkimusyksikön vastaanottama käsittely W_i on riippumaton potentiaalisista lopputulemista, sillä käsittelyn jako suoritetaan arpomalla:

$$W_i \perp \{Y_i(0), ... Y_i(K)\} | D_i = 0.$$

Myös koeryhmä muodostetaan alkuperäisessä sijainnissa arpomalla, mutta toimenpidettä ei välttämättä jaeta satunnaisesti koeryhmän sisällä. Tutkimusyksikölle määrätty koulutusosio voi tällöin olla riippuvainen tutkimusyksikön potentiaalisista lopputulemista:

$$\tilde{T}_i \not\perp \{Y_i(0), ... Y_i(K)\} | D_i = 0. \tag{10}$$

Riippuvuus on seurausta menetelmästä, jolla toimenpiteet kohdennetaan koeryhmään kuuluville.

Kun teoriakehikon kolme oletusta ovat voimassa, pätee seuraava yhtäsuuruus:

$$E[Y_i | W_i = 0, D_i = 1] = E[E[Y_i | W_i = 0, D_i = 0, X_i] | D_i = 1].$$

Lopputuleman Y_i ehdollinen odotusarvo, kun annettuna on $W_i = 0$ ja $D_i = 1$, voidaan edelleen estimoida alkuperäisen sijainnin aineistolla ilman lisäoletuksia. Sama huomio ei kuitenkaan päde koeryhmään kuuluville, sillä:

$$E[Y_i | W_i = 1, D_i = 1] \neq E[E[Y_i | W_i = 1, D_i = 0, X_i] | D_i = 1]. \tag{11}$$

Kohdan (10) riippuvuus estää tässä tapauksessa yhtälön (11) voimassaolon. Tämä tarkoittaa sitä, että vain kontrolliryhmiä voidaan validisti vertailla uuden ja vanhan sijainnin välillä. Kiinnostuksen kohteena olevalle kausaalivaikutukselle ei toisin sanoen voida muodostaa tarkkoja ennusteita ilman lisäoletuksia.

4.2 Empiirinen strategia

Hotzin ym. (2005) tutkimuksen empiirisenä strategiana on testata oletuksia, joiden mukaan sijaintien välillä ei esiinny makrovaikutuksia tai käsittelyn heterogeenisuutta. Testauksessa hyödynnetään teoriakehikon ominaisuutta, jonka mukaan eri sijaintien kontrolliryhmiä voidaan validisti verrata toisiinsa käsittelyn heterogeenisuudesta huolimatta. Empiirinen strategia on kaksivaiheinen ja sitä voidaan havainnollistaa, kun annettuna on kaksi satunnaistettua koetta. Molemmat kokeet arvioivat tässä tapauksessa samaa koulutusohjelmaa, mutta eri sijainneissa.

Tutkijat arvioivat ensimmäisellä testillä makrovaikutusten olemassaoloa. Testi toteutetaan hyödyntämällä molempien koulutusohjelmien kontrolliryhmiä. Ideana on, että kontrolliryhmien väliset lopputulemat ovat vertailukelpoiset, jos oletus sijaintien välisestä ehdollisesta riippumattomuudesta on voimassa. Kun kiinnostuksen kohteena on estimoida kontrollien keskiarvotulema jälkimmäisessä sijainnissa, voidaan estimointi toteuttaa kahdella tapaa. Ensin kontrollien keskiarvotulema voidaan suoraan estimoida jälkimmäisen kokeen aineistolla:

$$E[Y_i(0)|D_i = 1] = E[Y_i|W_i = 0, D_i = 1]. \tag{12}$$

Toisena kontrollien keskiarvotulema voidaan estimoida hyödyntämällä ensimmäisen sijainnin aineistoa:

$$\begin{aligned} E[Y_i(0)|D_i = 1] &= E[E[Y_i(0)|D_i = 1, X_i]|D_i = 1] \\ &= E[E[Y_i|W_i = 0, D_i = 0, X_i]|D_i = 1] \end{aligned} \tag{13}$$

Estimaattorit (12) ja (13) ovat toisistaan funktionaalisesti riippumattomia, sillä näiden käyttämät lopputulemat ovat erit. Tutkijoiden mukaan molempien estimaattoreiden voidaan odottaa tuottavan toisiaan lähellä olevia arvoja, jos sijaintien välillä ei ole eroja. Estimaattoreiden tuottamia arvoja voidaan tällöin verrata tilastollisilla

testeillä ja jos testit hylätään, voidaan tämä tulkita todisteena makrovaikutusten olemassaololle.

Toisella testillä tutkijat arvioivat käsittelyn homogeenisuusoletusta. Testi on toteutukseltaan samanlainen kuin ensimmäinen testi, mutta se suoritetaan tällä kertaa koeryhmille. Ensimmäisen testin tapaan koeryhmän keskiarvotulema voidaan suoraan estimoida jälkimmäisen sijainnin aineistolla

$$E[Y_i(1)|D_i = 1] = E[Y_i|W_i = 1, D_i = 1], \tag{14}$$

tai ensimmäisen kokeen aineistolla

$$E[Y_i|W_i = 1, D_i = 1] = E[E[Y_i|W_i = 1, D_i = 0, X_i]|D_i = 1]. \tag{15}$$

Tutkijoiden mukaan estimaattoreiden (14) ja (15) tuottamien arvojen voidaan odottaa olevan lähellä toisiaan, jos ensimmäisen vaiheen testia ei hylätty ja käsittelyt ovat homogeeniset. Jos testi hylätään, kun ensimmäisen vaiheen testi on ensin hyväksytty, voidaan tämä tulkita käsittelyn homogeenisuusoletuksen hylkäyksenä.

4.3 Empiirinen analyysi

Hotz ym. (2005) hyödyntävät tutkimuksessaan neljän satunnaistetun kokeen aineistoa, jotka ovat peräisin Yhdysvalloissa 1980-luvulla toteutetuista Work INcentive-ohjelmista. Kokeet on alunperin toteutettu Arkansasissa, Viriginiassa, San Diegossa ja Baltimoressa ja ne keskittyivät koulutusohjelmiin, jotka tarjosivat työllistymistä tukevia palveluita. Ohjelmat poikkesivat toisistaan ajanhetken, sijainnin, väestön, rahoituksen ja ohjelman toimintojen suhteen. Myös ohjelmille havaitut keskiarvovaikutukset vaihtelivat sijainneittain.

Tutkijoiden tavoitteena on ennustaa kunkin ohjelman keskiarvovaikutus hyödyntämällä aineistoa muista ohjelmista. Tutkimuksessa arviodaan neljää lopputulemaa: työllisyyden ja kokonaistulojen indikaattoreita, joiden arvot määritetään vuosi ja kaksi vuotta satunnaistamisen jälkeen. Mukana on myös kaksitoista kovariaattia. Nämä koostuvat neljästä dummy-muuttujasta, jotka kuvaavat yksilöiden

väestöllisiä ominaisuuksia, neljän satunnaistamista edeltävän neljänneksen tulotasosta sekä indikaattoreista, jotka osoittavat, ovatko kyseisten neljännesten tulot olleet positiivisia. Lopputulemia analysoidaan jokaisessa sijainnissa erikseen koulutusohjelmaan ja kontrolliryhmään kuuluneilla. Lisäksi tuloksia tarkastellaan ottamalla huomioon yksilöiden aiempi työtausta.

Tutkijat muodostavat ennusteet hyödyntämällä kaltaistusta (engl. matching). Tutkijat tarkastelevat esimerkin vuoksi San Diegoon kuuluvia kontrolleja ja keskittyvät tilanteeseen, jossa yksilöiden keskiarvoansiot ennustetaan kolmen muun sijainnin avulla, kun satunnaistamisesta on kulunut vuosi ja yksilöillä on jotain aiempaa työkokemusta:

$$\mu_{SD,0,1} = E[Y_i | D_i = SD, W_i = 0, E_i = 1], \tag{16}$$

jossa $E_i \in \{0, 1\}$ kuvaa yksilön aiempa työhistoriaa (ei työkokemusta tai jotain työkokemusta) ja $D_i \in \{SD, AK, VA, MD\}$ kuvaa sijaintia.

Tutkijoiden tarkoituksena on estimoida parametri (16) siten, että jokainen San Diegon kontrolliryhmään kuuluva havainto kaltaistetaan lähimpään kolmessa muussa sijainnissa olevaan kontrolliryhmän havaintoon. Jos X on San Diegoon kuuluvan havainnon kovariaattivektori, on tavoitteena tällöin löytää kolmesta muusta sijainnista se kontrolliryhmän havainto, jonka kovarittivektori Z minimoi lausekkeen $(X - Z)'(X - Z)$. Ennen kaltaistusta kovariaatit normalisoidaan siten, että niiden keskiarvo on 0 ja varianssi on 1. Kaltaistus toteutetaan takaisinpanolla, jolloin sen järjestyksellä ei ole merkitystä. San Diegon tapauksessa kaltaistus johti aineistoon, joka sisälsi yhteensä 519 paritettua havaintoa.

Tutkijat muodostavat ennusteet siten, että kaltaistetut havainnot regressoidaan käytössä olleilla 12 kovariaatilla ja kahdella ylimääräisellä aggregaatti-muuttujalla (työllisyys/väestö ja reaalitulot/työntelijä). Kovariaatit sisällytetään regressioon, sillä kaltaistuksen avulla muodostetut parit eivät todellisuudessa ole täysin tarkkoja. Kun annettuna on regression kertoimet, tutkijat muokkaavat (engl. adjust) kaltaistetun aineiston lopputulemat kaltaistettujen kovariaattiarvojen välisellä erotuksella:

$$Y + \hat{\beta}(X - Z),$$

jossa Y kuvaa kaltaistettua lopputulemaa, $\hat{\beta}$ kuvaa regression kertoimia, X on San Diegoon kuuluvan havainnon kovariaattivektori ja Z on sille poimitun parin kovariaattivektori. Kun X ja Z ovat lähellä toisiaan, ei muokkauksen tulisi olla suuri. Tutkijoiden mukaan toimenpide saattaa kuitenkin merkittävästi laskea keskiarvovaikutukseen liittyvää harhaisuutta.

Tutkijat vertaavat tämän jälkeen kaltaistetulle aineistolle laskettua estimaattia kontrollien keskiarvotuloihin, joka saadaan laskettua San Diegon aineistolla:

$$\hat{\mu}_{SD,0,1} = \frac{1}{N_{SD,0,1}} \sum_{D_i=SD, W_i=0, E_i=1} Y_i,$$

jossa $N_{SD,0,1}$ on havaintojen määrä, kun yksilöt kuuluvat San Diegon kontrolliryhmään ja näillä on aiempaa työkokemusta. Tutkijat raportoivat kahden estimaatin välisen eron ja t-testisuureen nollahypoteesista, jonka mukaan estimaatit ovat yhtäsuuret. Sama toimenpide toteutetaan kaikissa sijainnissa siten, että huomioon otetaan eri lopputulemat, yksilöiden aiempi työkokemus sekä koulutusohjelmaan osallistuminen.

4.4 Tutkimuksen tulokset

Hotzin ym. (2005) tutkimuksen tavoitteena oli ennustaa kunkin koulutusohjelman keskiarvovaikutus käyttämällä aineistoja kolmesta muusta ohjelmasta. Arvioitavana oli neljä lopputulemaa, jotka koostuivat työllisyyttä ja kokonaistuloja kuvaavista indikaattoreista. Tutkimuksen empiirisenä strategiana oli muodostaa lopputulemista ennusteet erikseen koe- ja kontrolliryhmille, mikä mahdollisti ulkoisen validiteetin ongelmien arvioinnin. Ennusteita verrattiin tämän jälkeen lopputulemien keskiarvoihin, jotka saatiin laskettua kohdesijainnin aineiston avulla.

Tutkijat toteuttivat analyysin monivaiheisesti. Ensin kaikista aineistosta poistettiin yksilöt, joiden ominaisuudet eivät olleet tasaisesti edustettuina eri sijainneissa. Tähän kategoriaan kuuluivat kaikki miehet sekä naiset, joilla ei ollut vähintään kuusivuotiaita lapsia. Toimenpiteen validiteettia testattiin nollahypoteesilla, jonka mukaan aineistosta poistettujen yksilöiden keskimääräinen ansiotaso oli kussakin sijainnissa yhtäsuuri kuin aineistoon jätettyjen yksilöiden keskimääräinen ansio-

taso. Useimmissa sijainneissa nollahypoteesia ei hylätty, kun aineistosta pudotettuja yksilöitä arvioitiin yhdessä. Tilanne oli päinvastainen, kun aineistoista pois jätetyt yksilöt jaettiin miehiin, lapsettomiin naisiin sekä naisiin, joiden lapset olivat alle kuusivuotiaita.

Tutkijat jatkoivat seuraavaksi analyysiä ottamalla huomioon yksilöiden aiemman työkokemuksen. Tilastollinen yhteenveto muodostettiin siten, että yksilöt jaettiin jokaisessa sijainnissa kahteen ryhmään koulutusohjelmaa edeltäneen neljän neljänneksen työllisyystilanteen perusteella (ei työkokemusta tai jotain työkokemusta). Yhteenvedon perusteella molempien ryhmien väestölliset ominaisuudet olivat samanlaiset. Ainoa selkeä ero liittyi yksilöiden kouluttautumiseen: yksilöt, joilla oli aiempaa työkokemusta, olivat todennäköisemmin suorittaneet lukiotutkinnon.

Koulutusohjelmien keskiarvovaikutukset poikkesivat kuitenkin merkittävästi muodostettujen ryhmien välillä, vaikka muodostetut ryhmät olivat väestöllisiltä ominaisuuksiltaan samankaltaisia. Koulutusohjelman havaittiin kaikissa sijainneissa kasvattavan työllisyysastetta ja yksilöiden ansiotasoa niillä, joilla ei ollut aiempaa työkokemusta. Aiemman työkokemuksen omaavilla yksilöillä keskiarvovaikutukset vaihtelivat sijainneittain; osassa sijainneista vaikutukset olivat positiivisia ja osassa negatiivisia. Tämän vuoksi tutkijat muodostivat ennusteet erikseen kummallekkin ryhmälle.

Myös ennustustarkkuuksissa havaittiin vaihtelua, kun yksilöiden aiempi työkokemus otettiin huomioon. Työllisyyden ja tulotasojen ennusteet olivat yleensä lähempänä niiden oikeita keskiarvotuloksia ja tilastolliselta merkitsevyydeltään heikompia, kun arvioitavana olivat kontrolliryhmään kuuluvat yksilöt, joilla oli aiempaa työkokemusta. Esimerkiksi San Diegossa ennuste poikkesi keskiarvosta 181 dollarilla (t-testisuure on 0.6), kun ennuste koski yksilöiden tuloja vuosi satunnaistamisen jälkeen ja se muodostettiin kolmen muun sijainnin avulla. Vastaava ero oli 285 dollaria (t-testisuure 2.8), kun ennuste muodostettiin kontrolliryhmään kuuluneille, joilla ei ollut aiempaa työkokemusta. Suhteellinen ero oli merkittävä, sillä ensimmäisellä ryhmällä ero keskiarvoansioihin oli viisi prosenttia, kun taas jälkimmäisellä ryhmällä ero oli 28 prosenttia.

Tutkijat muodostivat vastaavat ennusteet myös koulutusohjelmaan osallistuneille. Toimenpiteen tarkoituksena oli arvioida käsittelyn heterogeenisuuden mahdollista vaikutusta ennustustarkkuuksiin. Käsittelyn heterogeenisuudella ei kuitenkaan havaittu olevan merkittävää vaikutusta, sillä tulokset olivat pienin poikkeuksin yhdenmukaisia kontrolliryhmälle muodostettujen ennusteiden kanssa.

Ennustustarkkuuksien havaittiin heikkenevän, kun ennusteet muodostettiin yksittäisillä aineistoilla. San Diegon tapauksessa ennuste ensimmäisen vuoden ansioille poikkesi sen keskiarvotuloksesta 1278 dollarilla, kun yksilöillä oli aikaisempaa työkokemusta ja ennusteet tehtiin Arkansasin aineiston avulla. Ero oli huomattava, sillä vastaava ero oli vain 181 dollaria, kun kaikkia kolmea aineistoa käytettiin yhdessä. Sama trendi toistui myös muissa sijainneissa. Ainoa selkeä poikkeus esiintyi Arkanssissa, jossa yksittäisten aineistojen avulla muodostetut ennusteet olivat yhdistetyn aineiston ennusteita tarkempia tilanteessa, jossa yksilöillä ei ollut aiempaa työkokemusta.

Myös yksittäisten aineistojen avulla muodostetut ennusteet olivat tarkempia, kun yksilöillä oli aiempaa työkokemusta. Ennustustarkkuuksissa havaittiin kuitenkin vaihtelua sijaintien välillä, joten yhdistetyn aineiston hyödyntäminen vähensi ennusteisiin liittyvää epävarmuutta molemmissa ryhmissä. Ennusteet olivat kaikissa sijainneissa yhdenmukaisia kontrolliryhmien ennusteiden kanssa, kun ne muodostettiin harjoitusohjelmaan osallistuneille. Yksittäisten sijaintien väliset suhteelliset erot vaihtelivat tässä tapauksessa kuitenkin hieman.

Lopuksi tutkijat analysoivat vaikutusta, joka eri kovariaattijoukoilla oli ennustustarkkuuksiin. Arvioitavana oli neljä eri tapausta. Ensin ennusteet muodostettiin ilman, että regressiota muokattiin kovariaateilla. Toisena ennusteet muodostettiin muokkaamalla regressiota vain yksilöiden ominaisuuksia kuvaavilla kovariaatteilla. Kolmantena mukaan lisättiin yksilöiden aiempia tuloja kuvaavat kovariaatit. Neljäntenä mukaan lisättiin aggregaattimuuttujat. Tulosten perusteella ennusteet eivät reagoineet merkittävästi ylimääräisiin kovariaatteihin, kun yksilöt jaettiin ensin ryhmiin aiemman työkokemuksen perusteella. Esimerkiksi San Diegossa ennustusvirhe laski 209 dollarista 178 dollariin, kun ensimmäisen vuoden ansioista muodostettiin ennuste yksilöille, joilla ei ollut aiempaa työkokemusta, ja yksilöiden

ominaisuuksia kuvaavat kovariaati lisättiin mukaan regressioon. Ennustuksen virhe kasvoi -136 dollarista -227 dollariin, kun vastaava ennuste muodostettiin yksilöille, joilla oli aiempaa työkokemusta. Sama trendi toistui riippumatta sijainnista, lisätyistä kovariaateista tai siitä, oliko yksilö osallistunut koulutusohjelmaan.

Hotzin ym. (2005) tutkimuksessa ennusteiden muodostaminen onnistui melko hyvin, kun yksilöillä oli aiempaa työkokemusta. Sijaintien välinen valintaharha onnistuttiin tässä tapauksessa poistamaan ja ennusteet paranivat, kun käytössä oli useampi aineisto. Ennusteet olivat kuitenkin epätarkkoja, kun yksilöillä ei ollut aiempaa työkokemusta. Havaitsematon heterogeenisuus vaikeutti tässä tapauksessa ennusteiden muodostamista. Tutkimuksessa ei esiintynyt viitteitä sille, että ohjelmien välisellä heterogeenisuudella olisi ollut ennustustarkkuuksia heikentävää vaikutusta. Hotzin ym. (2005) tutkimuksen tulokset osoittavat, että ympäristöjen väliset erot on mahdollista tasata ennen koetta havaittavien muuttujien avulla. Tutkimuksen menetelmät tarjoavat tällöin keinon tutkia sitä, minne ja mille joukolle alkuperäisen kokeen tulokset ovat yleistettävissä.

5 Yhteenveto

Tämän tutkielman tarkoituksena oli arvioida satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmia. Tutkimusmysymys rajattiin satunnaistettuihin kontrollikokeisiin, joiden avulla pyritään informoimaan poliittista päätöksentekoa. Jotta tutkimuksissa johdetut päätelmät soveltuvat tähän tarkoitukseen, tulee niiden olla sekä sisäisesti että ulkoisesti validit. Ulkoisen validiteetin huomioimisessa on edelleen kuitenkin puutteita, vaikka sen arvostus kasvaa taloustieteen tutkimuskirjallisuudessa.

Satunnaistettujen kontrollikokeiden ulkoiseen validiteettiin voivat vaikuttaa useat tekijät. Mahdolliset ongelmat voidaan luokitella kolmeen kategoriaan: ympäristöjen välisiin eroihin, koeasetelman skaalamisesta syntyviin vaikutuksiin sekä koehenkilöiden käyttäytymisen muutoksiin. Ongelmat eivät koske vain satunnaistettuja kontrollikokeita, mutta usein korostuvat niissä. Tämä johtuu siitä, että kokeet rajoittuvat tyypillisesti yksittäisiin ympäristöihin ja ajanhetkiin. Satunnaistettujen kontrollikokeiden ulkoisen validiteetin ongelmiin ei ole yhtä täydellistä ratkaisua. Ongelmat voidaan useimmissa tapauksissa kuitenkin huomioida, jolloin niiden vaikutuksia voidaan myös arvioida.

Tässä tutkimuksessa on esitetty lähestymistavat, joiden avulla ulkoisen validiteetin ongelmia voidaan arvioida. Tutkimuksessa on tämän lisäksi esitetty ratkaisu, jolla ympäristöjen väliset erot voidaan ottaa huomioon kahden tai usemaan kokeellisen ohjelman välillä. Esitetty menetelmä ei tee alkuperäisen kokeen tuloksista ulkoisesti valideja, mutta antaa keinon tutkia sitä, minne tulokset yleistyvät ja minne eivät. Ratkaisuehdotus soveltuu suunnitteilla olevien polittisten hankkeiden arviointiin, sillä päätös hankkeiden toteuttamisesta perustuu usein arvioon niiden todennäköisestä vaikutuksesta.

Mahdollinen jatkotutkimuskohde on laajentaa Hotzin ym. (2005) tutkimuksessa esitetyt menetelmät johonkin toiseen aineistokokonaisuuteen, jossa yksilöiden ominaisuudet jakautuvat eri tavalla. Analyysi voitaisiin esimerkiksi toteuttaa aineistoilla, jotka ovat peräisin pohjoismaissa toteutetuista kokeellisista koulutusohjelmista. Tällöin voitaisiin tutkia sitä, miten maantieteelliset erot vaikuttavat tut-

kimustulosten yleistettävyyteen. Toinen mahdollinen jatkotutkimuskohde on yhdistää Hotzin ym. (2005) tutkimuksessa esitetyt menetelmät osaksi yleisen tasapainon testausta, sillä tutkimuksessa oletettiin, ettei yksilöiden välillä ole vuorovaikutussuhteita.

Lähteet

Angrist, J. D − Pischke, J−S. (2009) *Mostly harmless econometrics*. Princeton University Press, Princeton.

Ashraf, N. − Karlan, D. − Yin, W. (2006) Tying Odysseus to the Mast: Evidence From a Commitment Savings Product in the Philippines. *The Quarterly Journal of Economics*. vol 121 (2), 635−672.

Athey, S. − Imbens, G. (2017) The Econometrics of Randomized Experiments. Teoksessa *Handbook of Field Experiments*. Toim. A. B. Banerjee – E. Duflo, 73−140. Elsevier, Amsterdam.

Banerjee, A.V. − Duflo, E. (2009) The Experimental Approach to Development Economics. Annual Review of Economics, Vol.1 (1), 151−178.

Bédécarrats, F. − Guérin, I. − Roubaud, F. (2020) *Randomized control trials in the field of development: a critical perspective* 1. p. Oxford University Press, Oxford.

Card, D.E (1999) The Causal Effect of Education on Earnings. Teoksessa: *Handbook of Labor Economics*, toim. O. Ashenfelter − D. E. Card, 1801−1863. North Holland, Amsterdam.

Campbell, D. T − Stanley, J. C (1963). *Experimental and quasi-experimental designs for research*. Chicago, Rand McNally.

Crépon, B. − Duflo, E. − Gurgand, M. − Rathelot, R. − Zamora, P. (2013) Do labor market policies have displacement effects? evidence from a clustered randomized experiment. *The Quarterly journal of economics*. Vol. 128 (2), 531−580.

Deaton, A. (2010) Instruments, Randomization, and Learning about Development. *Journal of economic literature*. Vol. 48 (2), 424−455.

Deaton, A. − Cartwright, N. (2018) Understanding and misunderstanding randomized controlled trials. *Social science* & *medicine*. Vol. 210, 2−21.

Duflo, E. − Glennerster, R. − Kremer, M. (2006) Using Randomization in Development Economics Research: A Toolkit. Teoksessa *Handbook of Development Economics. Volume 4*, toim. T. P. Schultz − J.A. Strauss, 3895−3962. North-

Holland, Amsterdam.

Duflo, E. – Hanna, R. – Ryan, S (2012) Incentives work: Getting teachers to come to school. *The American Economic Review*. Vol. 102 (4), 1241–1278.

Gautier, P. – Muller, P. – van der Klaauw, B. – Rosholm, M. – Svarer, M. (2018) Estimating Equilibrium Effects of Job Search Assistance. *Journal of labor economics*. Vol. 36 (4), 1073–1125.

Heckman, J. J. – Lochner, L. – Taber, C (1999) Human Capital Formation and General Equilibrium Treatment Effects: A Study of Tax and Tuition Policy. *Fiscal studies*. Vol. 20 (1), 25–40.

Heckman, J. J. – Vytlacil, E.J. (2007) Econometric Evaluation of Social Programs, Part II: Using the Marginal Treatment Effect to Organize Alternative Econometric Estimators to Evaluate Social Programs, and to Forecast Their Effects in New Environments. Teoksessa: *Handbook of Econometrics.*, toim. J.J. Heckman – E.E. Leamer, Vol.6B, 4875–5143. North-Holland, Amsterdam.

Holland, P. (1986) Statistics and causal inference. *Journal of the American Statistical Association*. Vol. 81 (396), 945–960.

Hotz, V. J. – Imbens, G. W. – Mortimer, J.H. (2005) Predicting the efficacy of future training programs using past experiences at other locations. *Journal of econometrics.* Vol. 125 (1), 241–270.

Imbens, G. W. (2010) Better LATE Than Nothing: Some Comments on Deaton (2009) and Heckman and Urzua (2009). *Journal of economic literature*. Vol. 48 (2), 399–423.

Imbens, G. W. – Rubin, D. B. (2015) *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press, Cambridge.

Krueger, A. B. (1999) Experimental Estimates of Education Production Functions. *The Quarterly journal of economics* Vol. 114(2) (1999), 497–532.

Manski, C. F. (2013) *Public Policy in an Uncertain World Analysis and Decisions*. Cambridge, Harvard University Press.

Peters, J. – Langbein, J. – Roberts, G. (2018) Generalization in the Tropics – Development Policy, Randomized Controlled Trials, and External Validity. *The*

World Bank research observer. Vol. 33 (1), 34−64.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational psychology*. Vol. 66 (5), 688−701.

Shadish, W. R. − Cook, T. D. − Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference.* Houghton Mifflin, Boston.

Stock, J. H. − Watson, M. W. (2020) *Introduction to Econometrics*. 4.p., Pearson education limited, Harlow.

Tilastokeskus: Käsitteet. <https://stat.fi/meta/kas/validiteetti.html>, haettu 30.5.2024.