Toimenpiteiden vaikutusten arviointi empiirisessä mikrotaloustieteessä

Tuomas Pekkarinen
Uppsalan yliopisto

1. Johdanto

Toimenpiteiden arvioinnilla (engl. program evaluation) tarkoitetaan menetelmiä, joilla pyritään mittaamaan erilaisten politiikkainterventioiden vaikutuksia yksilöiden taloudelliseen käyttäytymiseen tai asemaan. Havainnollistava esimerkki toimenpiteen arvioinnista taloustieteessä on valtion rahoittaman koulutusohjelman vaikutusten tutkimus. Arvioinnin tavoitteena on selvittää, nostiko kyseinen koulutusohjelma vaikkapa osallistujien kognitiivisia kykyjä tai tuloja verrattuna tilaan, jossa koulutusohjelmaa ei olisi lainkaan toteutettu. Arviointi tuottaa tietoa toimenpiteen kausaalivaikutuksista ja tämä tieto voi ohjata toimenpiteiden kohdentamista tai päätöksiä niiden laajentamisesta tai lopettamisesta.

Toimenpiteiden vaikutusten arviointi on haastavaa, koska yksittäiset ihmiset voidaan havaita samanaikaisesti vain joko toimenpiteeseen osallistuvina tai sen ulkopuolella. Tätä ongelmaa kutsutaan kirjallisuudessa arviointiongelmaksi. Arviointimenetelmillä pyritään ratkaisemaan tämä ongelma tuottamalla estimaatteja toimenpiteen ulkopuolelle jäämisestä aiheutuvista vaihtoehtoisista tulemista. Tämän kirjoituksen tavoitteena on luoda yleistajuinen katsaus arviointimenetelmiä käsittelevään kirjallisuuteen sekä niitä käyttäviin sovelluksiin.[1]

Toimenpiteiden arvioinnista tuli jo varhain vakiintunut käytäntö monilla tieteenaloilla. Esimerkiksi lääketieteessä uusien lääkkeiden on jo pitkään pitänyt läpäistä tarkka vaikutusten arviointi päästäkseen markkinoille. Talouspoliittisessa päätöksenteossa toimenpiteiden arvioinnin tarpeellisuus tuli ajankohtaiseksi aktiivisten työvoimapoliittisten toimenpiteiden tehokkuudesta käydyn keskustelun yhteydessä. Näiden toimenpiteiden kustannukset tunne-

[1] *Erinomaisia englanninkielisiä katsauksia tähän kirjallisuuteen ovat Heckman, LaLonde ja Smith (1999), Blundell, Dearden ja Sianesi (2005) sekä Blundell ja Costas Dias (2005). Hämäläinen (2005) on suomenkielinen katsaus työvoimapoliittisten toimenpiteiden vaikutusten arvinointiin.*

taan hyvin, mutta niiden todellinen vaikutus osallistujien tuloihin tai työllistymistodennäköisyyteen on usein epäselvä. Juuri näiden vaikutusten arvioimiseksi taloustieteilijät turvautuivat arviointimenetelmiin.

Yhdysvalloissa työministeriö (Department of Labor) perusti jo 1970-luvulla arviointiin erikoistuneen osaston (Office of Evaluation), jonka tehtävänä oli käyttää tieteellisiä arviointimenetelmiä ministeriön rahoittamien toimenpiteiden vaikutusten tutkimiseen. Sittemmin toimenpiteiden arviointi on yleistynyt myös eurooppalaisessa talouspoliittisessa päätöksenteossa. Brittiläinen Institute for Fiscal Studies ja ruotsalainen Institutet för Arbetsmarknadspolitisk Utvärdering ovat esimerkkejä eurooppalaisista tutkimuslaitoksista, joiden tehtävänä on arvioida säännöllisesti talouspoliittisten toimenpiteiden vaikutuksia. Toimenpiteiden arviointi on edelleen yleisintä työvoimapoliittisten ohjelmien vaikutusten tutkimuksessa, mutta viimeaikoina arviointimenetelmiä on käytetty yhä enemmän myös muilla taloustieteen osa-alueilla, kuten julkisessa taloudessa, toimialan taloustieteessä sekä erityisesti kehitysmaataloustieteessä. Vuonna 1999 James Heckmanille myönnettiin taloustieteen Nobelpalkinto osittain hänen arviointimenetelmiä kehittäneestä työstään.

Tämä katsaus keskittyy viimeaikaisiin taloustieteellisessä kirjallisuudessa käytettyihin menetelmiin. Näille menetelmille on yhteistä se, että ne ottavat lähtökohdaksi toimenpiteiden vaikutusten heterogeenisyyden ja pyrkivät estimoimaan niiden kausaalivaikutukset mahdollisimman heikoin ja läpinäkyvin oletuksin. Niinpä tässä kirjallisuudessa nojaudutaan huomattavan vähän ekonometriassa tavallisesti käytettyihin funktiomuoto- ja jakaumaoletuksiin. Arviointimenetelmien tavoitteena on tuottaa valideja estimaatteja yksittäisten toimenpiteiden vaikutuksista. Näiden tulosten yleistäminen ei yleensä ole arviointikirjallisuuden kiinnostuksen kohteena. Lisäksi arviointikirjallisuudessa keskitytään miltei yksinomaan osittaistasapainovaikutuksiin ja oletetaan, että toimenpiteillä ei ole ulkoisvaikutuksia. Tässä mielessä arviointimenetelmät poikkeavat perinteisistä rakenteellisista valikoitumismalleista.[2]

Esittelen tässä katsauksessa taloustieteellisissä sovelluksissa yleisimmin käytettyjä menetelmiä. Aloitan katsauksen esittelemällä ensin arviointiongelman yleisellä tasolla ja määrittelemällä ne kysymykset, joihin arvioinnilla pyritään vastaamaan. Arviointimenetelmät jaetaan yleensä kokeellisiin ja ei-kokeellisiin menetelmiin sen mukaan, määräytyykö toimenpiteeseen osallistuminen satunnaisesti vai ei. Kokeelliset menetelmät ovat yleisimpiä luonnontieteissä, joissa arviointi tehdään yleensä satunnaiskokeilla. Katsauksen kolmannessa jaksossa käsittelen satunnaiskokeita taloustieteessä ja tuon esiin niiden mahdollisuudet ja rajoitukset. Koska tällaisten kokeiden järjestäminen on yhteiskuntatieteellisissä sovelluksissa vaikeaa, ovat ei-kokeelliset menetelmät edelleen hyvin suosittuja taloustieteellisessä tutkimuksessa. Ei-kokeelliset menetelmät jaetaan usein kahteen ryhmään riippuen siitä, oletetaanko toimenpiteeseen valikoitumisen tapahtuvan tutkijalle havaittavien vai havaitsemattomien tekijöiden perusteella. Tässä katsauksessa käsitellään viimeaikoina suosituksi tullut vertaistamismenetelmä esimerkkinä ensimmäisen ryhmän ei-kokeellisista menetelmistä ja instrumentti-

[2] *Rakenteelliset valikoitumismallit, jotka ovat eräässä mielessä yksi vaihtoehtoinen arviointimenetelmä, jäävät tämän katsauksen ulkopuolelle. Vella (1998) on erinomainen katsaus valikoitumismalleja käsittelevään kirjallisuuteen.*

muuttujat ja erotukset erotuksissa -menetelmän esimerkkeinä toisen ryhmän menetelmistä. Katsauksen lopussa käyn läpi kuuluisan empiirisen esimerkin, joka havainnollistaa eri menetelmien kykyä estimoida aitoja toimenpiteiden kausaalivaikutuksia.

2. Arviointiongelma

Arviointiongelma on helppo havainnollistaa esimerkillä, jossa toimenpiteeseen osallistumista kuvataan dummy-muuttujalla D_i, joka saa arvon 1, jos yksilö i osallistuu toimenpiteeseen ja arvon 0, jos hän ei osallistu. Kiinnostuksen kohteena on mitata toimenpiteen vaikutus jatkuvaan muuttujaan Y_i.

Jokaisen yksilön i kohdalla havaitaan tulema

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

joka saa arvon $Y_i(1)$ tai $Y_i(0)$, riippuen siitä, osallistuuko i toimenpiteeseen ($D_i = 1$) vai ei ($D_i = 0$). Toimenpiteen vaikutus henkilön i tulemaan on

$$\Delta_i = Y_i(1) - Y_i(0).$$

Arviointiongelma on seurausta siitä, että Δ_i:ta ei voida suoraan laskea i:lle, koska emme havaitse i:n vaihtoehtoista tulemaa. Esimerkiksi kirjoituksen alussa mainitun koulutusohjelman tapauksessa emme voi tietää ohjelmaan osallistuneiden henkilöiden testituloksia tai palkkoja, jos he olisivat jääneet toimenpiteen ulkopuolelle. Nykyaikaisessa arviointikirjallisuudessa oletetaan lähtökohtaisesti, että toimenpiteen vaikutukset ovat heterogeenisiä. Niinpä yksilön i vaihtoehtoista tulemaa ei voida suoraan johtaa muiden yksilöiden vaihtoehtoisista tulemista. Arviointimenetelmät pyrkivät ratkaisemaan tämän ongelman konstruoimalla eri tavoin vaihtoehtoisen tuleman toimenpiteeseen osallistuville yksilöille.

Arvioinnin perimmäisenä tavoitteena on tuottaa jonkinlainen tunnusluku toimenpiteen vaikutuksista. Mikä tämä tunnusluku on, riippuu kiinnostuksen kohteena olevasta kysymyksestä. Jos arvioinnin tavoitteena on selvittää, mikä toimenpiteen vaikutus on populaatiosta satunnaisesti poimittuun yksilöön, ollaan kiinnostuneita toimenpiteen keskimääräisestä vaikutuksesta (engl. average treatment effect, ATE):

$$\text{ATE: } E[\Delta_i] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 0].$$

Tämä parametri on kiinnostava, jos tarkoituksena on laajentaa toimenpide koskemaan koko väestöä. Esimerkiksi pohjoismaisissa hyvinvointivaltioissa erilaiset tuet ovat usein ”universaaleja” siinä mielessä, että kaikilla kansalaisilla on oikeus niihin. Tällöin niiden vaikutusten arvioinnin kannalta relevantti parametri on ATE.

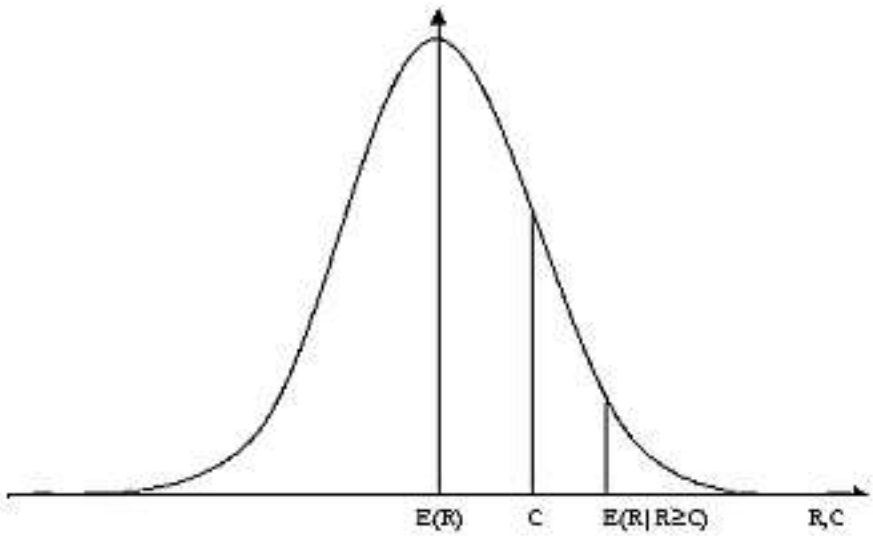
Usein taloustieteellisissä sovelluksissa ollaan kiinnostuneita jo olemassa olevan toimenpiteen vaikutuksista. Tällöin halutaan estimoida toimenpiteen vaikutus siihen osallistuville (engl. average effect of treatment on the treated, ATT):

$$\text{ATT: } E[\Delta_i \mid D_i = 1] = E[Y_i(1) \mid D_i = 1] - E[Y_i(0) \mid D_i = 1].$$

ATT on relevantti parametri esimerkiksi tilanteissa, joissa halutaan selvittää, kannattaako jo käynnissä olevaa toimenpidettä jatkaa.

On tärkeää huomata, että ATE ja ATT ovat harvoin samansuuruisia taloustieteellisissä sovelluksissa, joissa toimenpiteeseen valikoitumisen oletetaan määräytyvän hyödyn maksimoinnin peusteella. Carneiro, Heckman ja Vytlacil (2005) ovat havainnollistaneet ATE:n ja ATT:n

Kuvio 1. Koulutusohjelman tuottojen jakauma



eroa jo edellä esitetyn koulutusohjelman esimerkin avulla. Oletetaan, että koulutusohjelman yksityinen tuotto, *R*, vaihtelee populaatiossa kuviossa 1 esitetyn jakauman mukaisesti ja koulutuksen kustannukset, *C*, ovat kaikille samat. Yksilöt osallistuvat ohjelmaan vain jos *R* on suurempi kuin *C*. Tällöin ohjelman vaikutus siihen osallistuviin, $E(R \mid R = C)$, on suurempi kuin ohjelman keskimääräinen vaikutus $E(R)$. Koulutusohjelman vaikutusten arviointi on hyvä esimerkki taloustieteellisestä arviointitilanteesta, jossa osallistujien ja ei-osallistujien keskimääräiset vaihtoehtoiset tulemat ovat eri suuret.

3. Satunnaiskokeet

Monessa mielessä ihanteellinen ratkaisu arviointiongelmaan ovat satunnaiskokeet, joissa yksilöt ohjataan toimenpiteeseen osallistuvaan koeryhmään ja sen ulkopuolelle jäävään kontrolliryhmään satunnaisesti. Koska toimenpiteeseen osallistuminen on tässä populaatiossa satunnaista, ovat yksilöiden havaittavien ja ei-havaittavien ominaisuuksien jakaumat samat koe- ja kontrolliryhmissä. Tällöin havaittavat erot tulemissa näiden ryhmien välillä heijastavat toimenpiteen aitoa kausaalivaikutusta.

Jos toimenpiteeseen osallistuminen on aidosti satunnaistettua, toimenpiteen keskimääräinen vaikutus populaatiossa (ATE) voidaan laskea yksinkertaisesti vertaamalla keskiarvotulemia toimenpiteen piirissä ja sen ulkopuolella:

$$\hat{\Delta} = \bar{Y}_i(1) - \bar{Y}_i(0).$$

Satunnaiskokeita voidaan käyttää myös järkevästi suunniteltuina osallistujien keskimääräisen vaikutuksen (ATT) estimoimiseksi. Esimerkiksi, jos populaatio rajataan tietyn toimenpiteen kohteena olevaan ryhmään ja satunnaisesti suljetaan osa tästä populaatiosta toimenpiteen ulkopuolelle, koe estimoi keskimääräisen vaikutuksen osallistujille hieman vahvempiin oletuksiin nojautuen.

Satunnaiskokeet ovat klassinen tilastotieteellinen ratkaisu arviointiongelmaan. Jo Fisher (1928) esitti, että uskottava koe- ja kontrolliryhmän vertailu on mahdollista vain silloin, kun yksilöt määrätään näihin ryhmiin ”wholly at random”. Toisen maailmansodan jälkeen satunnaiskokeista on tullut vakiintunut käytäntö esimerkiksi lääketieteellisessä tutkimuksessa, missä uusien lääkkeiden on läpäistävä satunnaistetut potilastutkimukset päästäkseen markkinoille.

Satunnaiskokeista on tullut viimeaikoina suosittu arviointimenetelmä myös taloustieteilijöiden keskuudessa ja niistä saaduilla tuloksilla on ollut selviä vaikutuksia myös käytännön politiikkaan. Esimerkiksi Yhdysvaltain kongressi päätti vähentää nuorisolle suunnatun työmarkkinakoulutusohjelman (Job Training Partnership Act) rahoitusta sen jälkeen, kun Bloom et al (1993) olivat osoittaneet satunnaiskokeella, että ohjelman tulo- ja työllistämisvaikutukset saattavat olla jopa negatiivisia koulutuksensa kesken jättäneiden nuorten kohdalla. Myös Maailmanpankki suosittelee nykyään kehitysapuprojektiensa vaikutusten arviointia satunnaiskokeilla (Baker, 2000). Suomessa Työterveyslaitos toteutti kaksi työnhakukoulutusta koskevaa kenttäkoetta vuosina 1996 ja 1998. Taloustieteilijöistä näitä aineistoja ovat käyttäneet vain Hämäläinen ja Uusitalo (2005).

Satunnaiskokeisiin liittyy kuitenkin myös ongelmia. Ensinnäkin, satunnaiskokeilla on vaikea arvioida jo olemassa olevien ohjelmien vaikutusta. Satunnaistaminen voi muuttaa vakiintuneita byrokraattisia käytäntöjä ja näin vaikuttaa suoraan osallistujien tai kontrolliryhmän tulemiin. Tällöin satunnaistettu koe ei varsinaisesti mittaa itse toimenpiteen vaikutusta vaan toimenpidevaikutuksen ja satunnaistamisharhan (engl. randomisation bias) yhteisvaikutusta. Toiseksi, vaikutuksen mittaamiseen kuluu aikaa ja osallistujien joukosta tai kontrolliryhmästä poistuu usein yksilöitä ennen kuin mittaus pystytään suorittamaan. Tällöin satunnaiskoe tuottaa harhattoman tuloksen vain, jos tämä aineistosta poistuminen on myös satunnaista. Yhteiskuntatieteissä tällainen kokeen täydellinen kontrollointi on vaikeaa, vaikka kokeeseen osallistuminen saataisiinkin satunnaistettua.

Lisäksi yhteiskuntatieteellisissä sovelluksissa kontrolliryhmän tekemisiä on vaikea kontrolloida. Usein myös toimenpiteeseen osallistumattomille on olemassa vaihtoehtoisia ohjelmia. Tällöin on vaikea arvioida, mitä koe- ja kontrolliryhmän vertailu oikein mittaa. Tämä ongelma on erityisen vaikea arvioitaessa aktiivisten työvoimapoliittisten ohjelmien vaikutuksia pohjoismaissa, joissa käytännöllisesti katsoen kaikki työllisyyden ulkopuolella olevat työikäiset ovat jonkun ”ohjelman” piirissä. On mahdollista, että nämä vaihtoehtoiset ohjelmat tuottavat keskimäärin saman tuleman kuin kiinnostuksen kohteena oleva ohjelma ja tällöin satunnaiskoe antaa ymmärtää, että ohjelmalla ei ole vaikutusta

Ehkä kuitenkin merkittävin syy siihen, että satunnaiskokeet eivät ole täysin syrjäyttäneet ei-kokeellisia menetelmiä taloustieteellisessä tutkimuksessa on yksinkertaisesti se, että liian monen taloustieteellisesti tärkeän kysymyksen selvittämiseksi on mahdotonta järjestää satunnaiskokeita. Työn taloustieteessä kahden vuosikymmenen ajan kuumana käynyt keskustelu koulutuksen tuotoista on malliesimerkki tällaisesta kysymyksestä. On mahdotonta järjestää kokeita, joissa ihmiset ohjattaisiin eri koulutusurille satunnaisesti ja heidän elinkaaritulojaan mitattaisiin 30–40 vuoden kuluttua kokeen alkamisesta.

4. Ei-kokeelliset menetelmät

Ei-kokeellisilla menetelmillä on siis edelleen vahva asema taloustieteellisessä arviointitutkimuksessa. Nämä menetelmät kuitenkin estimoivat kiinnostuksen kohteena olevat parametrit vain tiettyjen oletusten nojalla. Nämä identifioivat oletukset erottavat eri ei-kokeelliset menetelmät toisistaan ja se, kuinka uskottavia käytetyt menetelmät ovat, riippuu tutkimuksen kohteena olevasta kysymyksestä sekä käytettävissä olevasta aineistosta.

Karkeasti ottaen ei-kokeelliset menetelmät voidaan jakaa kahteen ryhmään sen mukaan, olettavatko ne toimenpiteeseen valikoitumisen tapahtuvan havaittujen tai havaitsemattomien tekijöiden perusteella. Esittelen tässä katsauksessa kolme vakiintunutta ei-kokeellista menetelmää, joista vertaistaminen (engl. matching) kuuluu ensin mainittuun ryhmään. Instrumenttimuuttuja- (engl. instrumental variables) ja erotus erotuksissa -menetelmä (engl. differences in differences) kuuluvat puolestaan jälkimmäisenä mainittuun ryhmään.

4.1. Vertaistaminen

Vertaistamisen perusajatus on korvata satunnaistaminen havaittavilta ominaisuuksiltaan samanlaisten yksilöiden vertailulla. Lähtökohtana on oletus siitä, että samanlaisten yksilöiden välillä toimenpiteeseen osallistuminen on satunnaista. Koeryhmän yksilöiden vaihtoehtoinen tulema estimoidaan etsimällä kontrolliryhmästä vertailukohta, joka muistuttaa havaituilta omainaisuuksiltaan mahdollisimman paljon toimenpiteeseen osallistuvaa yksilöä.

Vertaistaminen nojaa kahteen tärkeään oletukseen. Ensinnäkin, valikoitumisen täytyy tapahtua havaittujen ominaisuuksien perusteella. Tämä tarkoittaa, että ehdollistettuna näille ominaisuuksille toimenpiteeseen osallistuminen ja vaihtoehtoiset tulemat ovat riippumattomia toisistaan. Jos havaittujen ominaisuuksien vektoria merkitään X:llä, tämä oletus voidaan kirjoittaa muotoon:

$$Y(0) \perp D \mid X.$$

Tätä oletusta kutsutaan ns. ehdolliseksi riippumattomuusoletukseksi (engl. conditional independence assumption). Esimerkiksi johdannossa esitellyn koulutusohjelman tapauksessa tämä oletus tarkoittaa, että vaikkapa iältään ja työmarkkinakokemukseltaan samankaltaisten yksilöiden keskuudessa koulutusohjelmaan valikoituminen on satunnaista.

Toiseksi, jotta vertailu olisi mahdollista, on X-muuttujien jakaumien oltava sellaisia, että mahdollisimman suurelle osaa X:n arvoista löytyy havaintoja sekä koe- että kontrolliryhmästä. X-muuttujien jakaumien on siis oltava sopivalla tavalla päällekkäisiä. Edellisessä esimerkissä tämä tarkoittaa, että iältään ja työkokemukseltaan samankaltaisten yksilöiden joukosta on löydyttävä sekä koulutusohjelmaan osallistuvia että sen ulkopuolelle jääviä.

Näiden kahden oletuksen aineistolle asettamat vaatimukset ovat ristiriidassa keskenään. Rikas X-muuttujien valikoima tekee ehdollisesta riippumattomuudesta uskottavampaa, mutta toisaalta vertaistamisesta tulee samanaikaisesti teknisesti vaikeaa, koska parien löytyminen kaikille X-muuttujien arvoille on epätodennäköisempää.

Rosenbaum ja Rubin (1983) ovat esittäneet ratkaisun tähän ongelmaan. Ns. propensity score -vertaistamisessa estimoidaan ensin malli, jossa osallistumisen todennäköisyyttä selitetään X-muuttujilla: $P(X_i) = P(D_i = 1 \mid X_i)$. Tämä malli estimoidaan yleensä probit- tai logit-reg-

ressioilla ja se tuottaa jokaiselle yksilölle propensity score -indeksin, $P(X_i)$, joka mittaa yksilön osallistumistodennäköisyyttä.

Rosenbaum ja Rubin (1983) osoittivat, että osallistumisen ja vaihtoehtoisen tuleman riippumattomuus voidaan ehdollistaa myös estimoidulle propensity score -indeksille. Ehdollinen riippumattomuusoletus saa nyt muodon $Y(0) \perp D | P(X)$. Vertaistaminen voidaan siis tehdä pelkästään propensity score -indeksin arvojen perusteella, jolloin verrataan osallistumistodennäköisyyksiltään samankaltaisten yksilöiden tulemia. Kunkin toimenpiteeseen osallistuneen yksilön kohdalla siis etsitään ei-osallistujien joukosta vertailuhavainto, jonka $P(X_i)$ on arvoltaan mahdollisimman läheinen. Toimenpiteen vaikutus tämän yksilön tulemaan on osallistujan tuleman ja vertailukohdan tuleman välinen erotus. Koko toimenpiteen vaikutus on näiden erotuksien keskiarvo osallistujien joukossa. Propensity score -vertaistaminen siis estimoi toimenpiteen vaikutuksen siihen osallistuneille (ATT), jos ehdollinen riippumattomuusoletus on voimassa.

Propensity score -menetelmästä on tullut suosittu viimeaikaisessa arviointikirjallisuudessa ja sitä on sovellettu moniin kysymyksiin myös taloustieteessä. Toisin kuin perinteiset regressiomenetelmät, jotka myös olettavat valikoitumisen tapahtuvan havaittavien muuttujien perusteella, vertaistaminen on ei-parametrinen menetelmä, joka ei edellytä funktiomuoto-oletuksia. Lisäksi, vertaistaminen ei ekstrapoloi tuloksia niille X-muuttujien arvoille, joilla löytyy vain osallistujia tai ei-osallistujia. Angristin (1998) tutkimus vapaaehtoisen asepalveluksen vaikutuksesta työtuloihin on havainnollistava esimerkki regression ja vertaistamisen eroista. Hyvä esimerkki vertaistamisen käytöstä suomalaisessa tutkimuksessa on Korkeamäen ja Uusitalon (2004) tutkimus sosiaaliturvamaksujen vähentämisen vaikutuksesta työn kysyntään Pohjois-Suomessa.

Mutta vaikka vertaistamista usein markkinoidaan teoriavapaana menetelmänä, on X-muuttujien valikoimisella usein ratkaiseva vaikutus saatuihin tuloksiin. Näiden muuttujien valitsemiseen ei ole olemassa yleisesti hyväksyttyä algoritmiä ja usein valikointi tapahtuukin subjektiivisesti. Lisäksi uskottava vertaistaminen vaatii sen verran rikkaan aineiston X-muuttujista, ettei menetelmää voida käyttää kaikkien kysymysten tutkimiseen.

Itse vertaistamiseen on olemassa myös erilaisia menetelmiä ja ainakin pienissä aineistoissa arvioinnin tulokset voivat vaihdella menetelmästä riippuen. Yleisesti käytetyin menetelmä kirjallisuudessa on ns. lähimmän naapurin vertaistaminen (engl. nearest neighbour matching), jossa kaikille osallistujille haetaan propensity score -indeksiltään lähin vertailukohta. Vaihtoehtoinen menetelmä on esimerkiksi kernel-vertaistaminen (engl. kernel matching), jossa vertailuhavaintona toimii propensity score -indeksien mukaan painotettu keskiarvo ei-osallistujien tulemista. Kirjallisuudessa ei ole mitään kriteeriä, millä näiden menetelmien välillä voisi valita.

4.2. Instrumenttimuuttujat

Instrumenttimuuttujat ovat havaittavia muuttujia, jotka vaikuttavat joko positiivisesti tai negatiivisesti toimenpiteeseen osallistumiseen, vaikuttamatta kuitenkaan suoraan kiinnostuksen kohteena oleviin tulemiin. Instrumenttimuuttujat siis aiheuttavat tulemista riippumatonta variaatiota toimenpiteeseen osallistumisessa. Koska instrumenttimuuttujat ikään kuin satunnaistavat osallistumisen, voi toimenpitee-

seen valikoituminen instrumenttimuuttujia käytettäessä tapahtua myös havaitsemattomien tekijöiden perusteella.

Arviointikirjallisuudessa puhutaan usein ”luonnollisista koetilanteista”, joissa osallistuminen on satunnaista, mutta satunnaistaminen ei ole tutkijan kontrolloitavissa. Esimerkiksi koulutusohjelmien arvioinnissa usein käytetty instrumentti on asuinpaikan etäisyys koulutuksen toteuttamispaikasta. Asumispaikalla on selvä vaikutus osallistumistodennäköisyyteen, mutta monet tutkijat ovat olleet halukkaita olettamaan, että sillä ei ole suoraa vaikutusta kognitiivisiin kykyihin tai tuloihin.

Aina 1990-luvun lopulle asti instrumenttimuuttujia pidettiin patenttiratkaisuna valikoitumisesta aiheutuviin ongelmiin. Viimeaikainen arviointitutkimus on kuitenkin ratkaisevasti lisännyt ymmärrystä siitä, mitä parametreja instrumenttimuuttujat oikein estimoivat ja johtanut huomattavaan varovaisuuteen niiden käytössä.

Instrumenttimuuttujien käyttöä on helpointa havainnollistaa yksinkertaisella dummy-instrumenttimuuttujalla, Z_i, joka saa arvoja 1 tai 0. Z_i voidaan tulkita ”toimenpiteeseen osoittavaksi” muuttujaksi. Jos esimerkiksi Z_i saa arvon 1, yksilö i ohjataan osallistumaan toimenpiteeseen ja jos Z_i saa arvon 0, hänet ohjataan toimenpiteen ulkopuolelle. Varsinainen osallistuminen ei kuitenkaan ole tutkijan kontrolloitavissa.

Käyttökelpoisella instrumentilla pitää olla vaikutus yksilön i toimenpiteeseen osallistumiseen, $D_i(Z_i)$, sekä osallistumisen kautta vaikutus tulemaan, $Y_i(Z_i, D_i)$. Näin ollen validin instrumentin on täytettävä seuraavat kaksi ehtoa:

(i): $E[D_i(1)–D_i(0)] \neq 0$
(ii): $Y_i(1,D_i) = Y_i(0,D_i) = Y_i(D_i)$.

Näistä ehdoista ensimmäinen vaatii, että instrumenttimuuttujalla on jokin vaikutus osallistumistodennäköisyyteen. Hyvä instrumenttimuuttuja on sellainen, joka aidosti osoittaa mahdollisimman tuntuvan joukon yksilöitä toimenpiteeseen. Tämän ehdon paikkansapitävyys on yksinkertaista testata tavanomaisilla aineistoilla. Sen sijaan toinen ehto ei ole testattavissa. Tämä ehto, jota kutsutaan kirjallisuudessa poissulkemisrajoitukseksi (engl. exclusion restriction), vaati, että instrumenttimuuttuja vaikuttaa tulemaan vain toimenpiteeseen osallistumisen, D_i, kautta. Nimenomaan tämä ehto takaa, että instrumenttimuuttujan osoittama osallistuminen on ikään kuin satunnaista tulemien kannalta.

Jos instrumenttimuuttuja Z täyttää ehdot (i) ja (ii), voidaan toimenpiteeseen osoitettujen ja osallistuneiden tulemia, ($Z_i = 1$ ja $D_i = 1$), verrata toimenpiteen ulkopuolelle osoitettujen ja ei-osallistuvien ($Z_i = 0$ ja $D_i = 0$) tulemiin. Tällöin instrumenttimuuttujat estimoivat seuraavan vaikutuksen:

$$\hat{\Delta}_{IV} = \bar{Y}_i(1, D_i(1)) – \bar{Y}_i(0, D_i(0)).$$

Instrumenttimuuttuja siis estimoi toimenpiteen vaikutuksen niille henkilöille, jotka muuttavat osallistumistaan instrumenttimuuttujan vaikutuksesta.

Kun toimenpiteen vaikutukset ovat heterogeenisiä, instrumenttimuuttujan estimoivan vaikutuksen tulkinta on vaikeaa ilman lisäoletuksia. Imbens ja Angrist (1994) ovat osoittaneet, että tämä estimaatti on tulkittavissa vain, jos instrumentti vaikuttaa osallistumistodennäköisyyteen monotonisesti, ainoastaan joko lisäten tai vähentäen sitä. Tällöinkään instrumenttimuuttuja ei estimoi toimenpiteen keskimääräistä vaikutusta (ATE) tai sen vaikutusta osallistujiin (ATT). Sen sijaan instrumenttimuuttu-

ja estimoi paikallisen version näistä parametreista. Tämä on toimenpiteen vaikutus niihin yksilöihin, jotka osallistuvat toimenpiteeseen vain jos $Z_i=1$ ja ovat osallistumatta vain jos $Z_i=0$. Se kuinka lähellä tämä estimoitu vaikutus on kiinnostuksen kohteena olevia parametreja riippuu käytetyn instrumentin ominaisuuksista. Imbens ja Angrist (1994) kutsuvat tätä uutta parametria toimenpiteen paikalliseksi keskimääräiseksi vaikutukseksi (engl. local average treatment effect, LATE).

Instrumenttimuuttujan epäonnistuminen ATE:n tai ATT:n estimoimisessa on seurausta nimenomaan koetilanteen luonnollisuudesta. Koska kyseessä ei ole aito koetilanne, on tutkijan mahdotonta kontrolloida toimenpiteeseen osallistumista. Populaatiossa voi aina olla yksilöitä, jotka osallistuvat tai jättävät osallistumatta toimenpiteeseen instrumenttimuuttujasta riippumatta. Esimerkiksi jos asuinpaikan etäisyyttä koulutuspaikasta käytetään instrumenttina koulutusohjelman vaikutuksen arvioinnissa, estimoi tämä instrumentti ohjelman vaikutuksen vain niille henkilöille, jotka osallistuvat ohjelmaan, koska he asuvat lähellä ja eivät osallistuisi ohjelmaan, jos he asuisivat kaukana. Tästä huolimatta jotkut yksilöt voivat osallistua ohjelmaan täysin etäisyydestä riippumatta. On vaikea uskoa, että ohjelman vaikutus näihin kahteen ryhmään olisi sama.

Nykyaikaisessa arviointikirjallisuudessa instrumenttimuuttujia käytetään silloin, kun LATE-parametrilla on jokin kiinnostava tulkinta. Cardin (1999) katsaus koulutuksen tuotoista käytyyn keskusteluun työn taloustieteessä 1990-luvulla on hyvä esimerkki siitä, miten parempi ymmärrys instrumenttimuuttujien toiminnasta on auttanut tulkitsemaan saatuja tuloksia. Hyvä esimerkki suomalaisesta arviointitutkimuksesta, jossa käytetään instrumenttimuuttujia, on Hämäläisen (2002) tutkimus työvoimapoliittisten toimenpiteiden vaikutuksista. Hämäläinen käytti toimenpiteiden alueellista tarjontaa instrumenttina toimenpiteeseen osallistumiselle.

4.3. Erotukset erotuksissa (EE)

Erotuksia erotuksissa -menetelmä (engl. differences in differences) käyttää hyväksi luonnollisia koetilanteita paneeliaineistojen avulla. Paneeliaineistossa yksilöitä seurataan ajassa ja tämä aikadimensio mahdollistaa instrumenttimuuttujien vaatiman poissulkemisrajoituksen heikentämisen. EE-menetelmän pääajatus on, että toimenpiteeseen valikoitumista ohjaa jokin ajassa muuttumaton havaitsematon tekijä. Tämän havaitsemattoman tekijän vaikutus tulemiin voidaan poistaa vertailemalla koe- ja kontrolliryhmien tulemien muutoksia. Esimerkiksi jos koulutusohjelmaan valikoituminen tapahtuu tutkijalle havaitsemattoman synnynnäisen kyvykkyyden perusteella, voidaan ohjelman vaikutusta arvioida vertailemalla osallistujien ja ei-osallistujien kognitiivisten kykyjen ja tulojen muutoksia.

EE-menetelmässä käytetään paneeliaineistoa hyväksi siten, että ensimmäisellä havaintoperiodilla, t_0, sekä koe- että kontrolliryhmä ovat toimenpiteen ulkopuolella. Vastaavasti toisella havaintoperiodilla, t_1, koeryhmä on käynyt läpi toimenpiteen, kun taas kontrolliryhmä on ollut sen ulkopuolella. Ideana on, että periodin t_1 tuleman vähentäminen periodin t_0 tulemasta kunkin ryhmän sisällä poistaa ryhmän sisäisen keskimääräisen kiinteän havaitsemattoman vaikutuksen tulemaan. Koe- ja kontrolliryhmän erotusten vertailu taas poistaa toimenpiteestä riippumattoman aikatrendin vaikutuksen tulemaan.

Tämän menetelmän käyttöä voidaan havainnollistaa kirjoittamalla koeryhmän tulemat periodeilla t_1 ja t_0 seuraavasti

$$Y^T_{it_1} = \phi_T + \theta_{t_1} + \Delta$$

$$Y^T_{it_10} = \phi_T + \theta_{t_0},$$

missä ϕ_T on koeryhmän havaitsematon vakio ja θ_t ovat tulemien ajassa tapahtuvia muutoksia. Toimenpiteen aito vaikutus on Δ. Kontrolliryhmän tulemat ovat vastaavasti

$$Y^C_{it_1} = \phi_C + \theta_{t_1}$$

$$Y^C_{it_10} = \phi_C + \theta_{t_0},$$

missä taas ϕ_C on kontrolliryhmän havaitsematon vakio.

EE-menetelmän identifoivat oletukset ovat, että yksilöiden havaitsemattomat tekijät ovat ajassa kiinteitä ja että tulemien trendi on sama koe- ja kontrolliryhmissä. EE-menetelmä estimoi toimenpiteen vaikutuksen vertailemalla koe- ja kontrolliryhmän tulemien erotuksia:

$$\hat{\Delta}_{EE} = (\bar{Y}^T_{t_1} - \bar{Y}^T_{t_0}) - (\bar{Y}^C_{t_1} - \bar{Y}^C_{t_0}) = \Delta,$$

missä $\bar{Y}$ viittaa kunkin ryhmän keskiarvotulemaan. Koeryhmän tulemien erotus, $\bar{Y}^T_{t_1} - \bar{Y}^T_{t_0}$, on aikatrendin ja toimenpiteen aidon vaikutuksen summa, kun taas kontrolliryhmän tulemien erotus, $\bar{Y}^C_{t_1} - \bar{Y}^C_{t_0}$, vastaa pelkkää aikatrendiä. Näiden kahden erotuksen erotus on yhtä kuin toimenpiteen aito vaikutus Δ. Tästä nimitys erotukset erotuksissa.

Kun toimenpiteen vaikutukset ovat heterogeenisiä, EE-estimaattori estimoi toimenpiteen vaikutuksen toimenpiteeseen osallistuneille (ATT):

$$E(\hat{\Delta}_{EE}) = E[\Delta_i \mid D_i = 1].$$

EE-menetelmästä on tullut hyvin suosittu arviointimenetelmä, koska se nojaa läpinäkyviin ja helposti ymmärrettäviin oletuksiin. Hyvä esimerkki erittäin vaikutusvaltaisesta, joskin myös kiistellystä, EE-menetelmää käyttäneestä tutkimuksesta on Cardin ja Kruegerin (1994) tutkimus minimipalkkojen vaikutuksesta pikaruokaloiden työllisyyteen. Card ja Krueger käyttivät Pennsylvanian pikaruokaloita kontrolliryhmänä ja tutkivat New Jerseyssä tehdyn minimipalkan korotuksen vaikutuksia.

EE-menetelmää käytettäessä oletetaan, että koe- ja kontrolliryhmien tulemissa on jokin kiinteä, ajasta riippumaton ero. Monissa käytännön sovellutuksissa on kuitenkin havaittu, että näiden ryhmien väliset erot vaihtelevat ajassa. Erityisen tuhoisaa EE-menetelmän kannalta on Ashenfelterin (1978) ja Ashenfelterin ja Cardin (1985) esille tuoma ns. "Ashenfelterin kuoppa", jolla viitataan tilanteisiin, joissa koeryhmän tulemat ovat juuri ennen toimenpidettä kontrolliryhmää selvästi heikommat. Tämä on tyypillinen tilanne esimerkiksi arvioitaessa koulutusohjelmia, joissa koulutus kohdennetaan sitä eniten tarvitseville. Tällaiset ihmiset ovat usein kokeneet jonkinlaisen tuloihin kohdistuvan negatiivisen sokin juuri ennen koulutusta ja tällöin EE-menetelmän estimoivat vaikutukset liioittelevat toimenpiteen vaikutusta.

Paneeliaineistoa voidaan käyttää hyväksi myös vertaistamisessa. Heckman, Ichimura ja Todd (1998) ovat kehittäneet menetelmän (engl. differences in differences matching), jossa vertaistetaan kontrollimuuttujien toimenpidettä edeltävien arvojen perusteella samankaltaisten yksilöiden tulemien erotuksia. Näin oletetaan, että toimenpiteeseen valikoituminen on satunnaista ehdollistettuna X-muuttujille ja havaitsemattomille kiinteille vaikutuksille ϕ_i. Blundell et al (2001) ja Huttunen (2006) ovat

Taulukko 1. NSW-koulutusohjelman vaikutukset satunnaiskokeella ja ei-kokeellisilla menetelmillä

	NSW:n estimoitu vaikutus	Satunnaisvirhe
LaLonde (1986)		
Satunnaiskoe	886	476
Koeryhmän ja CPS-kontrollin erotus	– 8 870	562
Havaittujen muuttujien kontrollointi	– 4 416	557
Erotukset erotuksissa	1 102	450
Dehejia ja Wahba (1999)		
Satunnaiskoe	1 794	663
Propensity score -vertaistaminen	1 713	1 115
Smith ja Todd (2005)		
Erotukset erotuksissa -vertaistaminen	1 849	781

Numerot viittaavat vuositulojen erotuksiin dollareissa. CPS viittaa Current Populatio Survey -aineistoon.

esimerkkejä tätä menetelmää soveltavista tutkimuksista.

5. Kokeellisten ja ei-kokeellisten menetelmien vertailu

Kuten edellä on jo todettu, taloustieteessä joudutaan vielä usein turvautumaan ei-kokeellisiin menetelmiin. Menetelmän valinta on kuitenkin vaikeaa, koska kaikki edellä esitellyt ei-kokeelliset menetelmät toimivat varsin erilaisten oletusten vallitessa. Yksi tapa vertailla näitä menetelmiä on tutkia, kuinka hyvin ei-kokeelliset menetelmät pystyvät replikoimaan satunnaiskokeilla saatuja tuloksia.

LaLonden (1986) tutkimus on kuuluisa esimerkki tällaisesta vertailusta. Hän käytti Yhdysvalloissa toteutetun National Supported Work (NSW) -ohjelman aineistoa. NSW oli ensimmäisiä työvoimakoulutusohjelmia, jonka vaikutukset arvioitiin satunnaiskokeilla. Viranomaiset valitsivat ensin ohjelmaan sopivat henkilöt ja osoittivat heidät sitten satunnaisesti joko koe- tai kontrolliryhmään. Koeryhmän jäsenille tarjottiin 9–18 kuukauden mittainen työharjoittelu. Kokeeseen osallistuneet henkilöt olivat pääasiassa huonosti työmarkkinoilla pärjääviä toimeentulonsaajia, entisiä huumeiden väärinkäyttäjiä tai koulunsa keskeyttäneitä.

Kokeen tarkoituksena oli tutkia ohjelman vaikutus osallistuneiden ohjelman jälkeisiin tuloihin. LaLonde (1986) laski tämän vaikutuksen vertaamalla koe- ja kontrolliryhmän tuloja kokeen jälkeen. Ohjelma nosti osallistujien vuosituloja 886 dollarilla. Tämän jälkeen hän tutki, miten hyvin eri ei-kokeelliset menetelmät pystyivät replikoimaan tämän tuloksen. Tätä varten satunnaiskokeessa käytetty kontrolliryhmä korvattiin Yhdysvaltain väestölaskennasta (Current population survey) poimitulla koeryhmää muistuttavalla kontrolliryhmällä.

Eri ei-kokeellisten menetelmien tulokset on raportoitu taulukossa 1. Kuten taulukosta ilmenee, ei-kokeelliset menetelmät antoivat hyvin erilaisia arvioita ohjelman vaikutuksesta. Vain EE-metelmällä saadut tulokset ovat lähelläkään satunnaiskokeen tuloksia ja LaLonde raportoi, että myös EE-tulokset vaihtelivat suuresti eri kontrolliryhmien välillä. LaLonden

tulokset johtivat siihen, että ei-kokeellisiin evaluointituloksiin alettiin suhtautua hyvin skeptisesti.

Dehejia ja Wahba (1999) käyttivät pienempää otosta NSW-aineistosta ja tutkivat propensity score -vertaistamisen kykyä replikoida satunnaiskokeiden tuloksia. Taulukossa 1 on raportoitu koe- ja kontrolliryhmän erotus Deheijan ja Wahban otoksessa, 1797 dollaria, ja vertaistamisella saatu tulos, 1713 dollaria. Vertaistaminen pääsee siis Deheijan ja Wahban otoksessa hyvin lähelle kokeellisia tuloksia. Dehejian ja Wahban tulokset johtivat vertaistamisen suosion huomattavaan kasvuun taloustieteilijöiden parissa.

LaLonden (1986) ja Dehejian ja Wahban (1999) lähestymistapaa on myös kritisoitu. On katsottu, että on epärealistista odottaa, että väestölaskennasta poimitut ei-kokeelliset kontrolliryhmät voisivat replikoida satunnaiskokeiden tulokset. Todennäköisesti ei-kokeellista arviointitutkimusta ei koskaan tehtäisi tällä tavoin. Etenkin Heckman et al (1998) ovat huomauttaneet, että LaLonden käyttämät kokeelliset ja ei-kokeelliset kontrolliryhmät ovat peräisin eri alueellisilta työmarkkinoilta ja näitä ryhmiä koskeva aineisto on kerätty eri kyselyillä.

Smith ja Todd (2005) kuitenkin muistuttavat, että erot alueellisten työmarkkinoiden ja kyselylomakkeiden välillä ovat ajassa kiinteitä havaitsemattomia tekijöitä. NSW-aineiston paneeliominaisuutta voidaan käyttää näiden ongelmien ratkaisemiseksi. Smith ja Todd (2005) käyttävät samaa aineistoa kuin Dehejia ja Wahba (1999) ja osoittavat, että erotuksia erotuksissa hyväksi käyttävällä vertaistamisella päästään hyvin lähelle satunnaiskokeen tuloksia. Lisäksi tällä menetelmällä saadut tulokset ovat huomattavasti vähemmän herkkiä vertaistamisessa käytettävien muuttujien tai aineiston vaihteluun kuin Dehejian ja Wahban (1999) käyttämä poikkileikkausvertaistaminen. Smithin ja Toddin (2005) tutkimus on hyvä esimerkki siitä, miten aineiston vahvuuksia voidaan käyttää hyväksi ei-kokeellisten menetelmien luotettavuuden parantamiseksi.

6. Johtopäätökset

Edellä esitetyt arviointimenetelmät ovat esimerkkejä tällä hetkellä suosituimmista menetelmistä taloustieteessä. Toisin kuin monissa muissa tieteissä, satunnaiskokeet eivät ole syrjättäneet ei-kokeellisia menetelmiä taloustieteellisessä tutkimuksessa. Tämä on seurausta kontrolloitujen kokeiden järjestämisen vaikeudesta yhteiskuntatieteissä.

Koetilanteet vastaavat taloustieteessä harvoin laboratorio-olosuhteita ja niinpä satunnaiskokeet ovat haavoittuvaisia satunnaistamisharhalle ja aineistosta poistumiselle. Lisäksi pelkkiin satunnaiskokeisiin nojautuminen sulkisi liian monia kiinnostavia kysymyksiä taloustieteellisen evaluointitutkimuksen ulkopuolelle.

Ei-kokeelliset menetelmät nojaavat puolestaan vahvoihin ja vaikeasti testattaviin oletuksiin. Nämä menetelmät asettavat myös hyvin erilaisia vaatimuksia aineistolle. Vertaistaminen edellyttää rikasta aineistoa taustamuuttujista. Instrumenttimuuttujat ja erotuksia erotuksissa -menetelmä taas edellyttävät sopivan luonnollisen koetilanteen löytymistä.

Arviointimenetelmiä käytettäessä on muistettava mihin oletuksiin nämä menetelmät nojaavat ja minkälaisia vaatimuksia ne asettavat aineistolle. Viimeaikainen arviointitutkimus on merkittävästi lisännyt ymmärrystä siitä, mitä parametrejä menetelmät estimoivat. Ratkaise-

vaa tälle kehitykselle on ollut lähteminen liikkeelle heterogeenisten vaikutusten oletuksesta. □

Kirjallisuus

Angrist, J. (1998), "Estimating the labor market impact of voluntary military service using social security data on military applicants", *Econometrica* 50: 1009–1027.

Ashenfelter, O. (1978), "Estimating the effect of training programs on earnings", *The Review of Economics and Statistics* 60: 47–57.

Ashenfelter, O. ja Card, D. (1985), "Using longitudinal structure of earnings to estimate the effect of training programs", *The Review of Economics and Statistics* 67: 648–660.

Bloom, H.S., Orr, L.L., Cave, G., Bell, S.H. ja Doolittle, F. (1993), *The National JTPA Study: Title II-A Impacts on Earnings and Employment at 18 Months*, Bethedsa, Abt Associates.

Baker, J. (2000), *Evaluating the Impact of Development Projects on Poverty: A Handbook for Practitioners*, Directions in Development, World Bank, Washington, D.C.

Blundell, R. ja Costas Dias, M. (2002), "Alternative approaches to evaluation in empirical microeconomics", Cemmap Working Paper, No CWP10/02.

Blundell, R., Costas Dias, M., Meghir, C. ja Van Reenen, J. (2004), "Evaluating the employment effect of mandatory job search program", *Journal of the European Economic Association*.

Blundell, R., Dearden, L. ja Sianesi, B. (2005), "Evaluating the effect of education: Models, methods and results from the National Child Development Survey", *Journal of the Royal Statistical Society*, Series A, 168: 473–512.

Card, D. (1999), "The causal effect of education on earnings", teoksessa Ashenfelter, O. ja Card, D. (toim), *Handbook of Labor Economics vol III*, Amsterdam, North Holland.

Card, D. ja Krueger, A. (1994), "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania", *American Economic Review* 84: 772–784.

Carneiro, P., Heckman, J.J. ja Vytlacil, E. (2005), "Understanding what instrumental variables estimate: Estimating marginal and average returns to education", julkaisematon.

Dehejia, R.H. ja Wahba, S. (1999), "Reevaluating the evaluation of training programs", *Journal of the American Statistical Association* 94: 1053–1062.

Fisher, R.A., (1935), *The Design of Experiments*, New York, Hafner.

Heckman, J.J., Ichimura, H. ja Todd, P. (1998), "Matching as an econometric evaluation estimator", *Review of Economic Studies* 65: 261–294.

Heckman, J., Ichimura, H., Smith, J. ja Todd, P. (1998), "Characterizing selection bias using experimental data", *Econometrica* 66: 1017–1098.

Heckman, J.J., LaLonde, R.J. ja Smith, J.A. (1999), "The economics and econometrics of active labor market programmes", teoksessa Ashenfelter, O. ja Card, D. (toim.), *Handbook of Labor Economics vol III*, Amsterdam, North Holland.

Huttunen, K. (2006), "The Effect of Foreign Acquisition on Employment and Wages: Evidence from Finnish Establishments", *The Review of Economics and Statistics*, tulossa.

Hämäläinen, K. (2002), "The Effectiveness of labour market training in different eras of unemployment", teoksessa Ilmakunnas, S. ja Koskela, E. (toim.), *Towards Higher Employment – the Role of Labour Market Institutions,* VATT-julkaisuja 32.

Hämäläinen, K. (2005), "Työvoimapolitiikan arvioinnin sietämätön vaikeus", teoksessa Hämäläinen, K., Taimio, H. ja Uusitalo, R. (2006), *Työttömyys – taloustieteellisiä puheenvuoroja*, Helsinki, Edita.

Hämäläinen, K. ja Uusitalo, R. (2005), *Kannattaisi kokeilla – Kokeelliset menetelmät työvoimapoliit-*

tisten toimenpiteitten vaikutusten arvioinnissa, Työpoliittinen tutkimus, no 285.

Imbens, G. ja Angrist, J. (1994), "Identification and estimation of local average treatment effects", *Econometrica* 62: 467–475.

Korkeamäki, O. ja Uusitalo, R. (2004), "Employment effects of a payroll tax cut – Evidence from a regional tax subsidy experiment", julkaisematon.

LaLonde, R. (1986), "Evaluating the econometric evaluations of training programs with experimental data", *American Economic Review* 76: 604–620.

Rosenbaum, P. ja Rubin, D. (1983), "The central role of the propensity score in observational studies for causal effects", *Biometrika* 70: 41–55.

Smith, J. ja Todd, P. (2005), "Does matching overcome LaLonde's critique of nonexperimental estimators?", *Journal of Econometrics* 125: 305–353.

Vella, F. (1998), "Estimating models with sample selection bias: A survey", *Journal of Human Resources* 34: 127–169.