



Hariduse andme- tarkus

2.0 Sissejuhatus 89

2.1 **Andmetest otsusteni. Kuidas aitab Statistikaamet Eesti ühiskonda targemaks muuta?** 93

Terje Trasberg, Marre Karu, Liina Osila ja Kadri Rootalu, peegeldaja: Dan Bogdanov

2.2 **PISA-pimedus. Miks Eesti ei kasuta oma registriandmete potentsiaali?** 101

Eneli Kindsiko ja Liis Roosaar, peegeldaja: Liiri Oja

2.0 Sissejuhatus

Hariduse andmetarkus

Mihkel Solvak

PÕHISÕNUMID

- **Eesti puhul ei kehti truism, et registriandmed on laia katvusega, aga infovaesed, ning uuringud on väikese katvusega, aga inforikkad.** Me saame registriandmetega teha otse väga kvaliteetseid uuringuid ning uuringuid registriandmete abil parendada.
- **Tugevama andmekaitsega hariduses toome tahtmatult kaasa „kaitse“ ka uute võimaluste eest tulevikus.** Kui me ei võimalda tundlikke registriandmeid haridussüsteemi andmetega liidestada, tekib oht, et alahindame süsteemis esinevaid probleeme lausa mitmekordselt.
- **Registriandmete alakasutus tekitab rahalist püsikulu ning uute uuringute tegemine omakorda alternatiivkulu.** Registripõhiste uuringute jaoks loodud virtuaalsete andmekogude arhiveerimine lubaks vähendada uuringukulusid tulevikus.
- **Mikroandmete revolutsioonilisele kasutuselevõtule peaks järgnema samaväärne muutus poliitika mõjuhindamisel ja seires.** Läbilõikeuuringute kõrval tuleb rohkem tähtsustada poliitikavalikute mõjuhindamise automatiseerimist.

SISSEJUHATUS

Andmepõhise otsustamise ja valitsemise eelduseks on andmete korje ja hoiustamine, sisuks aga väärindatud andmete ärakasutamine paremate otsuste tegemiseks. Andmed pole eesmärk iseeneses, vaid vahend tuleviku otsuste tegemisel ebakindluse vähendamiseks ning juba tehtud otsuste tõhususe ja mõjususe hindamiseks, et saaks targemini ümber otsustada. Teoorias kõlab see väga lihtsalt, aga praktikas on asi alati keerulisem. Siinse peatüki esimene artikkel (Terje Trasberg, Marre Karu, Liina Osila ja Kadri Rootalu) piltlikustab, kuidas võiks andmeid paremini kasutada, ning teine

artikkel (Eneli Kindsiko ja Liis Roosaar) käsitleb võimalust, mis juhtub siis, kui jätame väärt andmed kasutamata. Esimene räägib sellest üldisemate näidete varal, teine võtab konkreetse juhtumi ning demonstreerib uute andmete abil, kuidas me võime ka muidu kvaliteetse uuringuga haridussüsteemis toimuvat drastiliselt alahinnata.

Mõlema artikli laiem eesmärk on näidata, kuivõrd suur on Eesti registriandmete potentsiaal parandada poliitikakujundamist.

REGISTRITE LIIDESTAMINE JA UURINGUTE VÄÄRINDAMINE

Registriandmeid korjatakse administratiivülesannete täitmiseks ja need katavad populatsiooni laialt. Sihipäraste uuringute vaatekohast on need andmed aga sekundaarsed ja seega reeglina pigem infovaesed, sisaldades vaid kaudset või piiratud hulgal uuringu põhieesmärkide täitmiseks vajalikku infot. Uuringuid seevastu viiakse läbi spetsiifilistele uurimisküsimustele vastamiseks ning neis korjatakse üldjuhul väga inforikkaid andmeid. See rikas info katab aga vaid uuringu valimit, mille alusel peame üldistama tulemusi kogu elanikkonnale. Registriandmed tekivad automaatselt riigi eri protsesside toimimise kaudu, uuringuandmeid korjatakse suure lisakuluga pigem vastavalt tajutud vajadusele uuringuid läbi viies. Tihtilugu ollakse olukorras, kus ühelt poolt on olemas väga laia elanikkonda katvad infovaesed andmed, mis pidevalt uuenevad, teiselt poolt aga väga kitsa katvusega inforikkad andmed, mis kas ei uuene üldse või uuenevad kordusuuringute kaudu harva.

Mõlemat tüüpi andmete kasutamisega kaasnevad erinevad, kuid olulised probleemid. Registriandmete puhul on probleemne vajadus opereerida lähenditega, kui register ei sisalda täpset sihipärast indikaatorit, samuti kajastavad need reeglina inimeste käitumise tagajärgi, kuid mitte käitumise motiive ja põhjendusi. Uuringuandmete puhul tekitab küsimusi võimalik valimi nihe, andmete kiire vananemine ning küsitlusandmete puhul ka teatud indikaatorite ebatäpsus võrreldes registripildiga.

Siinses peatükis näitame, et need üldistused aga Eesti puhul ei kehti, sest meil on olemas hea teoreetiline ja tehniline võimekus registriandmeid nii omavahel kui ka erinevate uuringutega liidestada. Terje Trasbergi, Marre Karu, Liina Osila ja Kadri Rootalu artiklis Statistikaameti võimaluste kohta on rida näiteid, kuidas nii registrite kui ka eraldi läbiviidavate uuringute lüngad liidestamise abil lahendatakse ning andmeid heas mõttes väärindatakse. Näiteks määrati registripõhises rahvaloenduses elukoht ligi 20 andmeallika alusel ehk teisisõnu loodi lähendite abil parem indikaator. Teine lahendus

on liidestada eesmärgipärane uuring registriga selleks, et korjata eraldi vaid registris puuduvaid andmeid ning lisada otse need, mis on juba sobiva täpse indikaatorina registris olemas. Näiteks haaratakse Eesti tööjõu-uuringusse Eesti Hariduse Infosüsteemi (EHIS) andmed inimese käimasolevate õpingute kohta, selleks et vähenendada vastaja vajadust neid andmeid täies mahus uuesti esitada.

Liidestamine saab olla mitte ainult registrite või uuringute parendamise vahend, vaid vaikimisi rakendatud põhimõte, et mitte samu andmeid korduvalt korjata.

Avaliku teabe seadus sätestab, et andmekogudesse oma avalike ülesannete täitmiseks andmeid korjav asutus peab lähtuma ühekordse küsimise printsiibist ning erandeid sellest eraldi põhjendama. Aga millega siis põhjendame seda, et nende samade asutuste tellitud uuringute puhul on andmete mitmekordne korje tavapraktika ning registriga liidestamist ehk ühekordse küsimise põhimõtte rakendamist peab eraldi taotlema ja põhjendama? Niisiis võiksime ka uuringute puhul laiemalt rakendada põhimõtet, et samu andmeid küsime vaid ühe korra ja kui andmed on registris olemas, siis tuleks põhjendada nende liidestamata jätmist, mitte nende liidestamist.

ANDMEKAITSE EI PEAKS „KAITSMA“ PAREMA TULEVIKU EEST

Eneli Kindsiko ja Liis Roosaare artiklis PISA tulemustes peituvast hariduslõhest on huvitav näide: ebatäpse indikaatori tulemusena on uuringus sotsiaalmajandusliku tausta mõju ulatust alahinnatud kahekordselt. Kui viimases PISA uuringus, mille metoodilises kvaliteedis üldiselt kahelda ei saa, ilmneb, et matemaatika-testi tulemuste varieeruvusest 13% on seletatav vanemate taustaga, siis matemaatika-eksamite tulemuste ja laste vanemate sissetuleku registriandmete liidestamisel agregeeritud tasandil näeme, et ema või isa sissetulek seletab vastavalt 24–26% tulemuste variatsioonist ehk kaks korda rohkem (vt artikli 2.2 tabel 2.2.2). Need tulemused on veelgi drastilisemad, kui vaadata eraldi Tallinna, kus on sel viisil seletatud variatsioon ligi 60% (samam, tabel 2.2.3). Kuidas

Ka uuringute puhul võiksime laiemalt rakendada põhimõtet, et samu andmeid küsime vaid ühe korra ja kui andmed on registris olemas, siis tuleks põhjendada nende liidestamata jätmist, mitte nende liidestamist.

saab PISA tulemus ja meie registriandmete pilt olla niivõrd erinev? Seletus on viisis, kuidas PISA mõõdab vanemate sissetulekut: parema lähene-mise puudumisel küsitakse õpilase käest koduse vara olemasolu kohta. Kodune vara võib kohati olla hea lähend perekonna jõukust mõõta, kuid olukorras, kus meie registriandmed sisaldavad vanemate tegelikku sissetulekut ja haridus-taset, oleks mõistlik võimaldada neid andmeid kasutada, selmet sõltuda õpilase hinnangust jõukuse lähendile. Artiklis jõutakse järeldusele, et „Eesti hariduslõhe on väheste ressurssidega koolis õppiva ja nõrga sotsiaalmajandusliku taustaga lapse nägu“. Kuigi saame põhjusega olla uhked Eesti rahvusvaheliselt kõrgel tasemel olevate PISA tulemuste üle, mida keegi kahtluse alla ei sea, näeme neis tulemustes peituvat lõhet otsekui läbi prillide, mis ei suuda pilti kuidagi ära fookustada. Kui me andmete liidestamist laste ja nende vanemate privaatsuse riive vältimiseks kardame ega võimalda, siis paradoksaalselt „kaitseme“ neid lapsi ka parema hariduse eest tulevikus. Kahekordselt alahinnatud ja kasvava lõhe ulatus määrab otseselt selle, kas ja kuidas me selle muutmiseks haridus- ja sotsiaalpoliitikat suuname.

ALAKASUTATUD ANDMED, KORDUV KULU JA KORDUV TULU

Käesolev inimarengu aruanne on andmepõhine, tuues välja mitmeid seni vähe kasutatust leidnud andmetel põhinevaid leide hariduse valdkonnas. Andmepõhise otsustamise ja valitsemise üks keskseid postulaate on, et andmete tegelik väärtus tuleb nende mitmekülgsest kasutamisest. Andmete korje ja hoiustamine on alati puhas kulu, erinevate andmestike liidestamine on aga väärindamine ja selle alusel paremate otsuste langetamine juba puhas tulu. Terje Trasberg, Marre Karu, Liina Osila ja Kadri Rootalu

toovad oma artiklis näiteid, kuidas nn programmi-välise statistikatööna täielikult või suures mahus registripõhiseid uuringuid tehakse. Selliste tööde tulemusena tekivad ajutised unikaalsed virtuaalsed liidestatud andmekogud, mis pärast uuringu läbiviimist reeglina kustutatakse. Kustutamisel on rida häid põhjusi, kuid kustutamise paratamatuks tagajärjeks on lisaks algselt kantud kulule korduva kulu tekitamine iga järgneva sarnase või samu andmeid kasutava uuringu läbiviimisel. Sellega tekib pidev arvestatav alternatiivkulu nii ajas kui ka rahas, sest sama virtuaalse registri uuesti loomine tähendab alati uut lubade taotlemist ja andmekoosseisu loomist olukorras, kus võiks seda ressursi muudel otstarvetel kasutada. Lahendus oleks tekkivate virtuaalsete andmekogude arhiveerimine ja neile vastava eetikakomitee loa alusel juurdepääsu võimaldamine – sel viisil ei oleks meil nii suurt andmete kasutuse korduvkulu ning saaksime rohkem korduvtulu.

Kui me andmete liidestamist laste ja nende vanemate privaatsuse riive vältimiseks kardame ega võimalda, siis paradoksaalselt „kaitseme“ neid lapsi ka parema hariduse eest tulevikus.

LÄBILÕIKELISEST MÕÖTMISEST PIDEVA SEIRENI

Milton Friedman on öelnud, et ärgem hinnakem poliitikaid nende eesmärkide, vaid tulemuste alusel.¹ Kui tahes ülla eesmärgiga, kuid ebaefektiivne või vastupidiste tulemustega poliitika on lõpuks siiski lihtsalt halb poliitika. Alates 2010. aastatest võib Euroopas rääkida nn mikroandmete revolutsioonist, sest just sel aastakümnel hakati rohkem registre toorandmeid liidestama ja analüüsima kasutamaks, eelkõige erinevate poliitikameetmete mõju hindamiseks. Samas on ka liidestatud andmekoosseisu pealt tehtud igasuguse mõju hinnang staatiline raport, millele poliitikakujundaja või -omanik peab uue hindamise tellima.

Eneli Kindsiko ja Liis Roosaar toovad oma artikli kokkuvõttes välja praktilisi samme, kuidas minna edasi tuvastatud hariduslõhe lahendamisel. Eeltingimus selliste muudatuste sisseviimiseks on aga ka mõjude hindamise andmepõhine käsitlemine. Registriandmed loovad end ise eri asutuste igapäevaste tööprotsesside käigus, mis tähendab, et registritel põhineva poliitika mõjude hindamisel andmekorje kulusid ei teki või on need minimaalsed ja kaetud juba registriomaniku poolt. Me peaksime vähem tellima staatilisi läbilõikeuuringuid ja rohkem nõudma nendes loodud teadmiste automaatset genereerimist.

Hariduslõhe ulatuse täpne tuvastamine määrab meie võimaliku poliitikasekkumise viisi ja ulatuse, sekkumise enda tõhusus ja mõjus ilmneb aga selle mõju pidevas automatiseeritud seires. Halb oleks näiteks kolm või neli aastat hiljem ning miljoneid eurosid kulutades teada saada, et kuigi poliitika oli üllas, on selle päriselulised mõjud olnud nõrgad või soovitud vastupidised. Kursimuutuse vajalikkus tuleb tuvastada enne, kui on tekkinud suured pöördumatud kulud.

Peaksime vähem tellima staatilisi läbilõikeuuringuid ja rohkem nõudma nendes loodud teadmiste automaatset genereerimist.

Selle peatüki artiklite peegeldustes arutatakse, kuidas nende probleemide valguses edasi minna. Dan Bogdanov ütleb, et statistikaseadus lubab ja ka ei luba andmeid töödelda, Statistikaamet seevastu saab hästi andmeid töödelda, aga ei saa hästi andmeid edastada. Liiri Oja otsib tasakaalu andmete kaitse ja kasutamise vahel nii proportsionaalsuse printsiibi mõistliku rakendamise kui ka tehnoloogiate abil. Mõnikord on muudatuse saavutamiseks vaja vana hea asi uue nimega nimetada ja seda võimalust kasutada sisu ümberdefineerimiseks ja mandaadi laiendamiseks. Dan Bogdanov viskab siin õhku mõtte teha Statistikaametist Andmeamet. Väärrib arutelu.

VIIDATUD ALLIKAD

1. M. Friedman, [Living within our means](#). Interview by Richard D. Heffner. – The Open Mind, PBS Thirteen, 7.12.1975. Transcript. Open Mind Archive.