

2.0 Introduction

Data literacy in education

Mihkel Solvak

KEY MESSAGES

- **In Estonia, the truism that administrative register data offer broad coverage but limited depth while surveys provide rich detail but limited coverage does not hold.** High-quality research can be conducted directly using register data, and survey research can be strengthened through their integration.
- **Stronger data protection in education may inadvertently shield us from future opportunities.** If sensitive administrative register data cannot be linked with education system data, we risk substantially underestimating problems within the system.
- **The underuse of register data generates recurring financial costs, while commissioning new surveys entails opportunity costs.** Archiving virtual datasets created for register-based research would help reduce future research expenses.
- **The revolutionary use of microdata should be accompanied by a corresponding shift in policy impact assessment and monitoring.** Alongside cross-sectional surveys, greater emphasis should be placed on automated evaluation of policy measures.

INTRODUCTION

Data-driven decision-making and governance require the collection and storage of data, but their real value lies in using enhanced data to make better decisions. Data are not an end in themselves but a means of reducing uncertainty in future decisions and of assessing the efficiency and effectiveness of those already taken, allowing course corrections where necessary. In theory, this sounds straightforward; in practice, it is more complex. The first article in this chapter

(Terje Trasberg, Marre Karu, Liina Osila and Kadri Rootalu) illustrates how data can be used more effectively, while the second (Eneli Kindsiko and Liis Roosaar) examines what happens when valuable data are left unused. The first draws on broader examples; the second focuses on a specific case and shows, using new data, how even a high-quality study may significantly underestimate developments in the education system.



The broader aim of both articles is to demonstrate the substantial potential of Estonia's administrative register data to improve policy-making.

LINKING REGISTERS AND ENHANCING SURVEYS

Administrative register data are collected to carry out public functions and typically cover the population as a whole. However, from the perspective of targeted research, such data are secondary and often limited in depth, containing only indirect or partial information relevant to specific research questions. Surveys, by contrast, are designed to answer clearly defined questions and usually gather highly detailed data. This richness, however, applies only to the survey sample, and findings must therefore be generalised to the wider population. Register data are generated automatically through the operation of state systems, whereas survey data are collected at considerable additional cost in response to identified research needs. In practice, this often means that data with broad population coverage but less detail are continuously updated, while more detailed data cover only a narrow sample and are either not updated at all or updated only infrequently through repeat surveys.

Both types of data present distinct challenges. With register data, researchers may need to rely on proxy indicators where precise, purpose-built measures are unavailable. Registers also tend to capture behavioural outcomes rather than the motivations behind them. Survey data, in turn, raise concerns about sampling bias, the rapid ageing of data and, in the case of questionnaires, potential inaccuracies in self-reported information compared with administrative register data.

In this chapter, we show that these generalisations do not apply in the Estonian context, as we have both the theoretical and technical capacity to link register data with one another and with survey data. The article by Terje Trasberg, Marre Karu, Liina Osila and Kadri Rootalu outlines several examples of how the limitations of both registers and standalone surveys can be addressed through linkage, thereby enhancing

the value of the data. For example, in the recent register-based population census, place of residence was determined using around 20 data sources – that is, a more accurate indicator was constructed through the combination of proxy measures. Another approach is to link a purpose-designed survey with a register in order to collect separately only the data missing from the register and to incorporate directly those already available there as reliable indicators. For example, data from the Estonian Education Information System (EHIS) on ongoing studies are integrated into the Estonian Labour Force Survey to reduce the need for respondents to provide the same information again in full.

Linkage can serve not only as a way to improve registers or surveys but also as a default principle to avoid collecting the same data repeatedly.

The Public Information Act requires institutions collecting data for the performance of public functions to follow the once-only principle and to justify any exceptions. However, how can it be justified that, in surveys commissioned by these same institutions, the repeated collection of identical data remains common practice, while linkage with registers – that is, the application of the once-only principle – must be separately requested and justified? It would be reasonable to apply the once-only principle more consistently. Where data already exist in registers, it should be the decision not to link them that requires justification, rather than the linkage itself.

DATA PROTECTION SHOULD NOT SHIELD US FROM A BETTER FUTURE

The article by Eneli Kindsiko and Liis Roosaar on the educational inequality hidden in PISA results offers a striking example: because an imprecise indicator is used, the impact of socioeconomic background is underestimated by roughly half. In the most recent PISA study – whose methodological quality is generally not in doubt – 13% of the variance in mathematics test results is attributed to parental background. However, when mathematics examination results are linked with administrative register data on parents' income

at aggregate level, mothers' or fathers' income explains 24–26% of the variance – approximately twice as much (see Article 2.2, Table 2.2.2). The difference is even more pronounced in Tallinn, where nearly 60% of the variance is explained in this way (see Table 2.2.3).

How can the PISA results and the picture derived from administrative register data differ so markedly? The explanation lies in how PISA measures parental income: in the absence of a better option, pupils are asked about household possessions. While possessions may at times serve as a proxy for family wealth, in a context where administrative registers contain parents' actual income and education levels, it would be more appropriate to enable the use of these data rather than rely on pupils' assessments of proxy indicators. The article concludes: 'Educational inequality in Estonia has the face of a child who studies in an under-resourced school and comes from a disadvantaged socioeconomic background.' Although Estonia's internationally high PISA results – which are not in dispute – are a source of justified pride, the underlying inequality appears blurred, as if viewed through lenses that fail to bring the picture into focus. If, out of concern for children's and parents' privacy, we refrain from enabling data linkage, we may paradoxically be shielding these children from better educational opportunities in the future. The scale of educational inequality that is both growing and underestimated by roughly half directly influences whether and how education and social policy are directed to address it.

UNDERUSED DATA, RECURRING COSTS AND RECURRING RETURNS

This Human Development Report is data-driven and presents several findings in the field of education based on data that have so far been underused. A central premise of data-driven decision-making and governance is that the real value of data lies in their repeated and varied use. Data collection and storage are pure costs, whereas linking datasets creates added value, and using them to inform better decisions generates a return. Terje Trasberg, Marre Karu, Liina Osila and Kadri Rootalu describe studies conducted entirely or largely

on the basis of administrative registers as so-called extra-programme statistical work. Such projects produce temporary, unique virtual linked datasets, which are typically deleted once the study is completed. There are valid reasons for deletion, but an unavoidable consequence – beyond the initial cost – is that each similar study, or any study relying on the same data, generates new costs. This creates a recurring and significant opportunity cost in both time and money, as recreating the same virtual dataset requires new authorisations and reconstruction of the data, even though those resources could be used elsewhere. One solution would be to archive the virtual datasets created and, subject to approval by the relevant ethics committee, allow controlled access to them. This would reduce recurring data-use costs and increase the returns from existing data.

FROM CROSS-SECTIONAL MEASUREMENT TO CONTINUOUS MONITORING

Milton Friedman argued that policies should be judged not by their intentions but by their results.¹ However noble the objective, a policy that is ineffective or produces the opposite outcome is ultimately a poor policy. Since the 2010s, Europe has witnessed what can be described as a microdata revolution: raw administrative register data have increasingly been linked and used for analysis, particularly to assess the impact of policy measures. Yet even when based on linked datasets, impact assessments typically take the form of static reports, and each new question requires a separate evaluation.

In the conclusion of their article, Eneli Kindsiko and Liis Roosaar propose practical steps to address the identified educational inequality. Implementing such measures, however, presupposes a data-driven approach to impact evaluation. Administrative register data are generated automatically through the routine activities of public institutions, meaning that when policy impacts are assessed using register data, additional data collection costs are minimal or already covered by the register owner. We should therefore commission fewer static cross-



sectional studies and place greater emphasis on the automated generation of knowledge from data that already exist.

Accurately identifying the scale of educational inequality determines the form and scope of possible policy interventions, while their effectiveness can be assessed only through continuous, automated monitoring. It would be problematic to discover three or four years later, after spending millions of euros, that although a policy was well-intentioned, its real-world effects were limited or even counterproductive. The need for course correction must be identified before substantial and irreversible costs are incurred.

The reflections accompanying the articles in this chapter consider how to move forward in light of these challenges. Dan Bogdanov notes that the Official Statistics Act both permits and restricts data processing: Statistics Estonia is well equipped to process data but less able to share them. Liiri Oja explores how to balance data protection and data use through the proportionate application of legal principles and the use of technological solutions. At times, achieving change may require renaming an established institution and using that moment to redefine its substance and expand its mandate. Dan Bogdanov raises the idea of transforming Statistics Estonia into a Data Agency – a proposal that warrants discussion.

REFERENCES

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