Challenges for the management of qualitative and quantitative data: The example of social policy-related data collections

Andreas Heinrich and Eduard Klein
University of Bremen, Germany

Social policy research often uses and/or generates a huge amount of research data. This poses two problems that have gained increasing prominence in recent social science debates: the quality of research data and, as a means of improving it, enhancing data transparency (i.e. the free availability of the relevant original research data). In order to improve one’s research, how can a researcher assess the quality of their data and how can data transparency be increased?

Naturally, the specific challenges related to data quality assessments depend on the academic discipline and the object of study. In this article, we use the example of social science data on post-Soviet area studies. However, this in no way implies that related problems are less pressing or necessarily of a different nature in other disciplines or for other parts of the world.

Overall, the country context affects data quality in multiple ways. Compared with the Organisation for Economic Co-operation and Development (OECD) region, the majority of post-Soviet countries possess deficient information infrastructures, as many resources and services provided centrally in the Soviet Union have been disconnected or can no longer be provided by the successor states (Johnson, 2014). Therefore, the availability and quality of statistical data are limited in many regards (Bessonov, 2013), and access to official data is often impeded. Moreover, in authoritarian countries, conducting interviews or collecting data on politically sensitive topics can lead to ethical and legal problems.
Problems of data quality

Based on the underlying cause, Heinrich et al. (2019) differentiate three quality problems concerning research data: (1) intentionally falsified data, (2) unintended mistakes in the data and (3) incomplete and thus misleading data.

1. The systematic falsification of research data is only detectable on a case-by-case basis; it appears most often in non-academic sources, for example, official state agencies. Especially in authoritarian regimes, state organs may simply produce the politically desired information. This often concerns social or economic statistics. Discrepancies in Russian official statistics, for instance, can – at least partly – be explained by deliberate misreporting. For example, Russia’s state statistical agency, Rosstat, reported in 2019 that the percentage of people living below the poverty line in 2018 decreased by 0.3% compared to 2017. However, according to the Institute for Social Analysis and Forecasting (ISAP), this decline can be attributed to Rosstat’s new methodology for calculating income, which reduces the poverty rate by overstating income increases, rather than any true reduction in poverty. ISAP calculated that incomes decreased by 2.3%, leading to an increase in poverty instead of the officially proclaimed – and politically favoured – decrease.

2. One can assume that unintended mistakes (on the part of the data collector) occur rather frequently in public opinion polls on politically sensitive issues in authoritarian regimes, where freedom of expression is restricted and may lead to negative consequences. In a survey conducted in Russia in 2016, only 30% of respondents stated that they would always honestly answer questions related to politics and only 12% of them assumed that other people would do so (Levada Center, 2016). In addition, with a high rejection rate in public opinion polls, only a small part of the Russian populace (between 10% and 30%) seems to be willing to take part in such inquiries (Napeenko, 2017). In this context, the validity of opinion poll results is questionable.

In a related case, tax evasion has influenced the Gini coefficient produced by the ‘Azerbaijan Household Income and Expenditure Survey’, which has been included in global datasets. Azerbaijan’s Gini coefficient had a low value, indicating a low degree of social inequality in the country. In fact, the low value was partly caused because better off, middle-class households did not participate in the survey for fear that their undeclared income would be detected and eventually taxed (Ersado, 2006).

3. Quantitative analyses often simply ignore incomplete data, thus potentially introducing bias. The International Federation of Human Rights complains about the lack of reliable statistics on migratory flows within the post-Soviet region. In Kyrgyzstan, for example, the ‘lack of disaggregated statistics specifically on the movement of women and children at [the] national and regional levels’ leads experts to believe that these data underestimate the number of Kyrgyz labour migrants by up to one million. Due to insufficient data recording
at border crossings, the majority of Kyrgyz migrant workers are undocumented. As a result, statistics from the home and host countries, as well as expert estimates, do not match (International Federation of Human Rights, 2016: 9). In another example, Turkmenistan and Uzbekistan withhold basic socio-economic data from household surveys, making it impossible to calculate – among other indicators – their Gini coefficient (the latest available Gini index for Turkmenistan is from 1998, for Uzbekistan from 2003). In the case of qualitative data, incomplete data are often harder to identify, and the implications for the conclusions drawn are less obvious. During the selection of interview partners for expert or elite interviews, for instance, the following biases are likely to occur: More important people (in terms of relevant responsibilities and knowledge) tend to delegate interviews to less important people. Moreover, in the ‘snowball approach’ (in which interviewees are found based on suggestions from earlier interview partners or from those who declined themselves), interviewees will most likely suggest like-minded people for interviews. In addition, in authoritarian regimes, respondents may be discouraged from talking to researchers or may self-censor their answers (Shih, 2015).

In all these cases, data collections based on these data should only be used after a proper assessment of their shortcomings – which does not always take place.

**Problems of data interpretation**

Even correct and complete datasets do not negate the need for data interpretation. Such problems can also occur when researchers assign the data more reliability than is permissible. Quantitative data may especially suggest an accuracy that may not be supported by the underlying information. In his article in this special issue, Brand looks at data on poverty, which imply a certain normative understanding of the object. Official Russian poverty statistics narrow the phenomenon down to a minimum subsistence level, whereas surveys and sociological studies suggest that poverty is much more widespread. These differences are not caused by poor data quality, but by various normative assumptions about poverty, which should always be made explicit when using such data.

The questionnaire design of surveys can also strongly influence the answers of the respondents. When asked about the desirability of democracy as a regime type, large parts of the populations in the post-Soviet region do not think of a political ideal type but of their own experiences with democratically elected governments in the 1990s. However, this period was characterised by economic hardships and social disruption (Carnaghan, 2011). Thus, the question is often understood as a question about the desirability of a ‘return to the 1990s’. Such surveys have to be interpreted and contextualised accordingly and cannot simply be included in global comparisons about perceptions of democracy.

In authoritarian regimes, the possibility of repression by state agencies also fosters self-censorship within the public, mass media and social media (Goode, 2010; Malthaner, 2014). Therefore, all forms of content and discourse analysis might include self-censored and actually censored forms of expression. To avoid misinterpretation, the country-specific context of data production has to be kept in mind.
Data transparency and privacy protection

In social science, data transparency is considered a solution to problems of data quality and data interpretation. By making research data publicly available, researchers receive the opportunity to assess not only the research results in academic publications but also the underlying raw data.

However, at least two factors render the expectation that publicly available data collections are quickly assessed for mistakes, and problems are easily spotted and immediately corrected rather over-optimistic. First, the problem is often related to the validity, applicability and contextuality of research data, which are challenged by a number of complex arguments. Second, many of the problems of data interpretation are specific to individual analyses and related academic publications. An assessment of the reliability of data collection and data interpretation is often not available to the broader academic community (Heinrich et al., 2019: 140). In addition, data collections related to qualitative research methods often cannot (easily) be prepared for online publication in the context of transparency initiatives. Thus, mistakes and misinterpretations are much harder to substantiate than in disciplines that work solely with quantitative methods.

With the upload of qualitative data collections (e.g. medical records, social security information), the question of privacy protection requires special attention (for more information, see Heinrich et al., 2019: 141–143). In contrast to quantitative research, qualitative research is often based on rather detailed profiles of research participants. It can be challenging to publish direct quotes anonymously, as these will be searchable on the Internet. Pseudonyms or nicknames may also be identifiable because they may be used in various contexts online and hence function as a digital identity (Utaaker Segadal, 2015: 43).

In politically sensitive regions such as the former Soviet Union, anonymity and privacy protection are of particular importance (Côté, 2013). Authoritarian regimes have increasingly learned to use modern technologies to identify people and suppress opposing views and criticism. Roberts (2013: 348) points out, ‘there can be no limit on the provision of anonymity and care in handling data; even in cases when the respondent does not ask for that provision’.

Consequently, the online availability of research data is not sufficient to mitigate the problems of data quality and data interpretation. To improve social policy research on the post-Soviet region, it would be desirable to additionally link the data to a peer discussion and/or the relevant literature discussing the respective data collection.

Conclusion: social welfare data on Discuss Data

One project that tries to tackle the above-mentioned problems through providing a single platform where research data are not only shared, but actively discussed and assessed, is ‘Discuss Data’, an ‘Open Platform for the Interactive Discussion of Research Data Quality (on the example of area studies on the post-Soviet region)’ created and operated by the Göttingen State and University Library and the Research Centre for East European Studies at the University of Bremen. As Heinrich et al. (2019: 141) state,

Discuss Data aims to create an online platform that combines the publication of research data not only with a documentation of the data collection process but also with an interactive place of communication to discuss, evaluate, and contextualise these research data. The expert
community will be enabled to indicate faulty or misleading data, to recommend complementary datasets (in case of gaps in the data collection) and to discuss extensively the validity, applicability, and interpretation of the data. This platform creates the opportunity to gather – in a structured way – the feedback to research data that is currently scattered among journal articles, conference papers, and blog posts or has not been published at all.

Discuss Data believes that the quality assessment and contextualisation of research data are best served through the publication of these data collections, their metadata (i.e. detailed descriptions of the data) and documentation describing the process of data collection in a single place. This enables a quality assessment by experts who are familiar with the content, method and/or context of the dataset. To protect privacy, Discuss Data offers every data provider the option to specify the extent to which data collections can become available online (Heinrich et al., 2019: 143).

In summary, the availability and quality of research data are limited in many regards: The challenges range from intended falsification of data, unintended mistakes, and incomplete datasets to the over- or misinterpretation of correct (and complete) datasets. Moreover, the publication of qualitative social science research data is often a cause for ethical concerns, most importantly regarding privacy protection. For the often-demanded transparency in academic knowledge production, the careful publication of the underlying research data is a necessary first step. The active discussion of these data is the second, even more vital step. However, academic fora that enable such a professional discussion do not yet exist. By addressing these problems, Discuss Data creates a digital infrastructure that functions as a virtual communication platform, enabling the discussion of publicly available research data. Sharing, using and discussing social policy data with Discuss Data would solve many of the aforementioned challenges that social policy researchers face – not only when working on the post-Soviet region.

Acknowledgements
Discuss Data is jointly conducted by the Göttingen State and University Library and the Research Centre for East European Studies at the University of Bremen. A first version of the Discuss Data platform is available online since autumn 2020.

Funding
The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) in the context of the Collaborative Research Centre 1342 ‘Global Dynamics of Social Policy’ (Project No. 374666841) as part of Subproject B06 ‘External reform models and internal debates on the new conceptualization of social policy in the post-Soviet region’.

Notes
1. For more information, see Heinrich et al. (2019).
2. According to Schneider (2012), Azerbaijan has one of the largest shadow economies in the world: Between 1999 and 2007, on average, 58% of Azerbaijan’s GDP (gross domestic product) was generated in the shadow economy.
3. See https://www.discuss-data.net/
References

Bessonov VA (2013) On the problems of Russian statistics. Problems of Economic Transition 55(11): 36–49.

Carnaghan E (2011) The difficulty of measuring support for democracy in a changing society: Evidence from Russia. Democratization 18(3): 682–706.

Côté I (2013) Fieldwork in the era of social media: Opportunities and challenges. PS: Political Science & Politics 46(3): 615–619.

Ersado L (2006) Azerbaijan’s Household Survey Data: Explaining why inequality is so low. Policy Research Working Paper WPS 4009. Washington, DC: World Bank. Available at: http://documents.worldbank.org/curated/en/2006/09/7063026/azerbaijans-household-survey-data-explaining-inequality-so-low

Goode JP (2010) Redefining Russia: Hybrid regimes, fieldwork, and Russian politics. Perspectives on Politics 8(4): 1055–1075.

Heinrich A, Herrmann F and Pleines H (2019) Transparency and quality assessment of research data in post-Soviet area studies: The potential of an interactive online platform. Journal of Eurasian Studies 10(2): 136–146.

International Federation of Human Rights (2016) Women and Children from Kyrgyzstan Affected by Migration. Paris: FIDH.

Johnson IM (2014) The rehabilitation of library and information services and professional education in the post-Soviet republics: Reflections from a development project. Information Development 30(2): 130–147.

Levada Center (2016) Trust in mass media and readiness to state one’s opinion. Levada Center, 12 August. Available at: https://www.levada.ru/2016/08/12/14111/ (accessed 17 February 2018).

Malthaner S (2014) Fieldwork in the context of violent conflict and authoritarian regimes. In: Della Porta D (ed.) Methodological Practices in Social Movement Research. Oxford: OUP, pp. 173–194.

Napeenko G (2017) If you scratch on a domestic liberal, you get an educated conservative: Sociologist Grigory Yudin about deceived public polls, elites’ fear of the people and the political suicide of the intelligentsia. Colta.ru, 13 March. Available at: http://www.colta.ru/articles/raznoglasiya/14158 (accessed 22 August 2017).

Roberts SP (2013) Research in challenging environments: The case of Russia’s ‘managed democracy’. Qualitative Research 13(3): 337–351.

Schneider F (2012) The Shadow Economy and work in the shadow: What do we (not) know? IZA Discussion Paper no. 6423, March. Bonn: Institute for the Study of Labor.

Shih V (2015) Research in authoritarian regimes: Transparency tradeoffs and solutions. Qualitative & Multi-Method Research 13(1): 20–22.

Utaaker Segadal K (2015) Possibilities and limitations of Internet research: A legal framework. In: Fosseheim H and Ingierd H (eds) Internet Research Ethics. Oslo: Cappelen Damm Akademisk, pp. 35–47.

Author biographies

Andreas Heinrich is postdoctoral researcher at the Collaborative Research Centre 1342 ‘Global Dynamics of Social Policy’ and the Research Centre for East European Studies at the University of Bremen, Germany. His current research focuses on knowledge transfer induced by international organisations, with a special focus on the countries of the former Soviet Union.

Eduard Klein is a research associate at the Research Centre for East European Studies at the University of Bremen, Germany. He is part of the Discuss Data team and editor of the Ukraine- Analysen, a German-language analytical digest on Ukraine. His research interests are corruption and informality in the post-Soviet region.