

International Journal of Population Data Science

Journal Website: www.ijpds.org



A Data Quality Framework for Integrated Data via Statistical Matching

Angelo Moretti¹, Camilla Salvatore¹, and Fabienne Kraemer²

¹Utrecht University, Utrecht, Netherlands

²GESIS - Leibniz Institute for the Social Sciences, Mannheim, Germany

Researchers often combine data from multiple sources with the goal of bringing together different types of information and creating a richer data set. With the combined data researchers can cross-validate information or explore relationships that cannot be examined using only one data source. To ensure the validity of empirical conclusions, the integrated data needs to meet certain data quality standards. However, each data source may carry specific representation or measurement errors into the integrated dataset, and the data integration process itself can introduce new biases or amplify existing ones. While quality measures have been proposed to assess data from multiple sources (De Waal et al., 2019), guidance on potential data quality issues across the stages of the integration process remains limited. We present a compre-

hensive data quality framework to guide researchers through integration via statistical matching. The framework outlines data quality measures and best practices at each stage of the process: pre-matching, matching, and post-matching. To illustrate its practical application, we apply the framework to a real-world case study integrating the 2021 German General Social Survey (ALLBUS) (probability-based survey) with data from the German Longitudinal Election Study in 2021 (non-probability survey). The integrated data are used to examine how various social attitudes relate to voting behavior and political engagement. Our work contributes to understanding data quality issues in integrated data while providing practical implementation guidance for researchers and practitioners.

