Profile report for the Chair 'Data Quality'
Faculty of Social and Behavioural Sciences, Utrecht University

1. Domain of the Chair
In the digital age where data play an increasingly important role in producing scientific knowledge and policy, the quality of data is crucial for both scientific and policy-relevant purposes. The need for expertise in data quality is growing. An important task for the chair is to bridge the gap between developments in science and future applications where data quality is essential, such as data-driven governance and policy within the public sector, and semi-automated data collection processes (such as using process data to monitor behaviour).

Whether data is of high quality depends on the purpose of its use ('fit for purpose'). In scientific research, data is usually of high quality if it can provide a solid answer to the research question posed. But sometimes it is also important whether an answer can be provided quickly or is easily interpretable. Such considerations often play a more significant role in, for example, policy research. The same data can be of high quality for one type of use but of low quality for another type.

The research of the chair includes various dimensions of data quality that together determine whether data is 'fit for purpose':
- The relevance of data to the research purpose
- The accuracy of data, including validity and reliability.
- The credibility of data, including data documentation of the data collection process
- The timeliness of data: the speed and regularity of data production
- The accessibility of data and ways to improve it through open science or the application of statistical techniques to use sensitive data (disclosure control)
- The interpretability of the data for its users.

In a more practical sense, the chair deals with investigating one or more dimensions of data quality within two major thematic areas:
1. Improving the quality of data during data production
2. Analyzing or correcting the data quality of existing data and solving problems due to low data quality in statistical analyses.

In practice, both themes are interconnected. To improve the quality of a data collection process, it is often necessary to first analyse problems with existing data collection procedures. In research that is repeatedly conducted, it is often possible to evaluate whether changes in the data collection process subsequently lead to higher data quality. This allows for connecting (changes over time in) the data collection process to effects on data quality. In other types of research, however, it is not possible to intervene in the data production process. For example because the data are produced as a byproduct of another process (as is often the case with administrative data or forms of Big Data). Conversely, research is sometimes aimed at improving the data collection process without room for evaluation or correction of quality problems afterward, for example, because data are not accessible or because it takes a long time to collect data.

The theme 'improving data quality' includes a number of topics.
- How to improve research designs: this concerns situations in both experimental and non-experimental social science research where the research process is under the control of the researcher. Within the empirical circle of conducting research, this specifically concerns the operationalization phase, how to measure the concepts of interest as well as possible (validity and reliability). It also concerns minimizing representation errors (e.g. sampling, preventing or limiting nonresponse), so that the results of the study can be generalized to a larger population..
- How to optimize research designs given different dimensions of data quality, as well as time and money. This topic also includes the development and evaluation of methods to integrate data from different sources within one study during data collection (e.g.,

administrative data, big data, surveys, and qualitative data) in such a way that different dimensions of data quality are balanced.
- Investigating the concept of data quality itself is essential as a basis of later improvement: What does data quality mean in different research contexts, and specifically in research situations where new types of data are collected (Internet of Things, Artificial Intelligence)?

The theme 'analyzing or correcting for data quality problems' includes the following topics:
- Evaluating research designs using existing data. This includes the analysis of measurement quality of instruments (validity and reliability, and the analysis of representation errors (e.g. sampling, non-response, generalization) to identify problems with data quality and areas of improvement. This topic also involves the use of statistical models to assess the reliability and validity of measurement instruments, and the use of external data sources (on the sample or population) to assess and quantify representation errors.
- Evaluating dimensions of data quality of research designs and the costs associated. Beyond measurement and representation errors, methods are needed to assess and balance other dimensions of data quality, such as the timeliness, coherence and interpretability of data. In particular, methods are needed to evaluate different ways to integrate different data sources within one study after data collection (e.g., using administrative data, big data, surveys, and/or qualitative data within one study).
- Developing methods so that future scientific studies can make scientifically informed choices between various research designs in different research contexts. In many circumstances, researchers will not have to time to formally model, assess and quantify dimensions of data. This also applies to situations when researchers design a study: should the focus be on minimizing errors in measurement or representation for example? Methods are needed that allow applied researchers to assess data quality assessment in practice relatively easily.

Within the chair, collaboration will take place with researcher across the different disciplines within the Faculty of Social Sciences (FSBS) regarding the improvement of research designs. Outside the faculty, collaboration will be sought with societal partners such as CBS, SCP, RIVM, WODC and other (local) authorities that deal extensively with empirical social science data and for whom the use of good data is essential for advising policy-making. Special attention is given to data quality within official statistics. Nationally, both within and outside Utrecht University, collaboration also takes place with researchers working with new sources of data, often employing a data science approach, who encounter issues with the quality of the data they use. Internationally, collaboration occurs with other research groups at the intersection of methodology and statistics for the social sciences and data science scientists.

2. The importance of the chair

In the digital era where data plays an increasingly important role in producing scientific knowledge and policy, the quality of data is crucial for both scientific and policy-relevant purposes. The need for expertise in data quality is growing. The emergence of the discipline of 'data science' is partly a result of the growing availability of data as a byproduct of the digitization of our society, as well as the continually increasing capabilities of computers and other devices such as mobile phones to process those data. Within the discipline of data science, the emphasis is on processing and analyzing data. In recent years, there has been an increasing realization that the quality of data is often a limiting factor in the application of techniques from data science. Within data science, there is currently relatively little expertise and attention given to improving the quality of data in data processes, or the use of statistical models to correct shortcomings in such data. Policy-makers are increasingly relying on data in assisting policy-making decisions, and one of the goals of the chair is to create awareness among policy-makers that it is essential that data are of high quality. Both data producers, such as scientists and governmental researchers need to be aware of methods that ensure high data quality, but this is also true for data users, such as policy-makers, and the general public. The chair fits into the current and growing need to analyse, correct, and improve the quality of data used in research that utilizes data from new sources.

Within the social sciences, as in other fields of science, there is an increasing use of large amounts of data. Digital data, such as health data collected with smartwatches, digital behavioural data collected via the internet, or location data and in-the-moment attitudinal data collected via apps on mobile phones, are increasingly used in the social and behavioural sciences. There are specific challenges that hinder the successful use of this type of data: for example, the use of digital tools varies greatly between groups, leading to a significant risk of selection effects – and representation errors – that in turn can bias the results of studies. Or how to combine digital behavioral data with more traditional question-based methods, and how to do so in a ethical and privacy-preserving way.

Additionally, it is essential for social scientists and policy-makers to be able to link new data sources to more traditional forms of data, such as experimental data, survey data, or administrative data. Integrating data from different sources for the same individuals or for different individuals is a scientific challenge. In the coming years, the need for expertise in data quality and the integration of data from different sources will only increase. The chair meets a growing need among social scientists, societal partners, and policy-makers to effectively convert the datafication of our society into social scientific knowledge and data-driven policies. For Utrecht University, the chair is important for several reasons. In the university's education and research, data plays an increasingly important role. Knowledge of data creation, data processing, and the essential role of data quality in both research and education within the strategic themes and focus areas is crucial. Utrecht University has a strong reputation for empirical research, particularly in the social sciences. Utrecht University values open science, in which improving the accessibility of data is an important elements. The chair in data quality strengthens Utrecht University's position in these fields.

3. Positioning of the Chair

Position in the Faculty of Social and Behavioural Sciences:
The chair in Data Quality is situated within the section of Methodology & Statistics of the department of Social Sciences, which is one of the three departments within the Faculty of Social and Behavioural Sciences. The section forms the administrative unit for education and research, as well as personnel and finances. Within the section, there are chairs in Applied Bayesian Statistics, Statistics for the Social Sciences, Methodology & Statistics for the Social Sciences, Data Science in Healthcare, Longitudinal Data Analysis, Collaborative AI for the Social Sciences, and the Statistical Analysis of Incomplete Data. All of these chairs emphasize data processing and analysis. The emphasis of the new chair in data quality lies in the production phase of data, and the question of how to improve data.

Position in Utrecht University:
At Utrecht University, there are two relevant strategic themes for the chair (Dynamics of Youth and Institutions for Open Societies), and several focus areas relevant to the chair (Applied Data Science, Governing the Digital Society, and Human-centered Artificial Intelligence). Both the strategic themes and focus areas aim to stimulate multi- and interdisciplinary research and establish new collaborations, both within and outside the university. Data quality plays a role in the focus areas when it comes to understanding the digital society, improving models within data science, and enhancing the interaction between humans and (AI) models. Both the focus areas and strategic themes see an increase in the use of data from new sources such as sensors, administrative data, and Big Data. There is also a need for expertise to combine new data sources with more traditional forms of data such as survey research and experimental data. The chair meets a need within social science research to evaluate how data from new sources, sometimes in combination with other data sources, can provide better answers to research questions.

Position nationally and internationally:
Outside Utrecht University, the research of the chair falls within the research school 'Interuniversity Graduate School for Psychometrics and Sociometrics (IOPS)'. Within IOPS, the research of the chair aligns with the theme of 'sociometrics'.
There are currently no chairs in the Netherlands that conduct research and provide education on data quality in the social sciences. Various universities have chairs in statistics, some of which also deal with methodology and data quality. In recent years, several chairs in data science have been established within university departments, focusing on data integration within specific fields.

Internationally, there are chairs in the methodology of social science research (in Germany: for example University of Bremen, University of Mannheim, LMU Munich; in the United Kingdom: University of Essex, London School of Economics, University College London, University of Manchester; in the United States: for example University of Maryland, University of Michigan, University of Nebraska, University of Wisconsin, University of Berkeley, University of Washington). These chairs mainly focus on data quality in survey research, although there has been a movement in recent years to also consider other data sources. Additionally, many universities have chairs in data science, usually housed within computer science departments, primarily focusing on data processing. Data quality often plays a modest role in these chairs.

4. Tasks and embedding of the Chair/ Internal Situation

The section of Methodology & Statistics provides education in research methodology and statistics for the Faculty of Social and Behavioural Sciences and is responsible for various courses within programs of other faculties at Utrecht University. The section conducts research into the use of new research methods, data, and data analysis techniques.

The research of the section focuses on three main themes:

1. Data quality
2. Statistical modeling
3. Applied data science/Artificial intelligence

Currently, the chair in 'statistical analysis of missing data' conducts research aligned with the theme of data quality, but within the broader theme of data quality, the department does not currently have a specific chair.

Therefore, the data quality chair will have a significant role within the **research** theme of 'data quality' at the section of Methodology & Statistics. Additionally, the chair will play a crucial role in education and collaboration with societal partners within the section. The chair will establish connections and collaborate with researchers within the strategic themes of 'Institutions for Open Society' and 'Dynamics of Youth' at Utrecht University. The chairholder of the data quality chair is thus a unifying and inspiring **leader** in both research and education.

The **education** offered by the section of Methodology & Statistics comprises a diverse range of courses within various programs at Utrecht University. Together with other professors within the section, the Data Quality chair is responsible for organizing, shaping, and delivering the curriculum on methods and statistics within the following programs:

- Bachelor's Programs FSBS (in Dutch):
    - Psychology
    - Interdisciplinary Social Sciences
    - Pedagogical Sciences
    - Sociology
    - Cultural Anthropology
    - Educational Sciences
    - Academic Teacher Program for Primary Education
- The research master Methodology and Statistics for the Behavioural, Biomedical, and Social Sciences (in English)

Additionally, the chair, along with other chairs, provides education within:

- The Professional Education program of the section, including the Utrecht Summer School
- University College Utrecht
- Courses in other Research Masters offered by the Faculty of Social Sciences
- Various other courses within Utrecht University in the fields of methods, statistics, and data science, such as the minor in Applied Data Science and the Master in Applied Data Science.

Within education, collaboration with societal partners, and research, the chairholder plays a crucial role in promoting **Team** Science in research and education on data quality. The chairholder oversees colleagues, ensure that employees can utilize their talents, and promotes scientific education and research according to the principles of TRIPLE as advocated within Utrecht University.

The chair is active both nationally and internationally in the field of research methods and data quality, actively seeking collaboration with societal partners in research and policy areas directly related to the theme of data quality. The chair is active in national and international networks in disseminating knowledge and implementing standards in documenting, investigating, and improving data quality as examples of **professional performance**.