

## DATA MANAGEMENT PLAN

PROJECT	
Project number:	[101109679]
Project acronym:	[TAXMORALBIAS]
Project name:	Tax Morale and Social Desirability Bias: Examining the Shadow Economy Experimentally

DATA MANAGEMENT PLAN	
Date:	[13/12/2024]
Version:	[Version 2.0]

### 1. Data Summary

This project will not re-use any existing datasets. The project will generate two primary types of data:

- **Survey List Experiment Data**

This dataset will include responses from a list experiment designed to assess the impact of tax morale on social desirability bias. The data will be quantitative in nature, collected via Computer Assisted Web Interviews (CAWI) and collected by a pollster company *Norstat Latvija*, ensuring high-quality data collection across North, East, and South Europe (Italy, Latvia and Denmark). The survey list experiment is expected to generate a comprehensive dataset from approximately 6,600 respondents (2,200 in each of the three countries).

- **Field Experiment Data**

This dataset will be compiled from a field experiment conducted in collaboration with the State Revenue Service to further explore the behavioural responses to tax compliance interventions. This field experiment will observe the activities of legal entities concerning the submission of Employer's Reports over the period of eight months. To evaluate the effectiveness of different gamification elements, participants will be randomly divided into five groups, each receiving distinct communication. Four groups will receive reminders about the upcoming deadline for report submission, while the fifth (control) group will not receive any communication, serving as the control group.

The study will involve legal entities registered in the Latvian Enterprise Register that have not been assigned an A taxpayer rating by the State Revenue Service and have submitted at least three Employer's Reports past the deadline in 2024. The research is expected to include approximately 4,500 legal entities, with communication being conducted with 3,700 of them as part of the study. The remaining fifth of the involved legal entities will serve as a control group, which will not receive any communication within the study.

The exact number of study participants will be identified on February 1, 2025, when the State Revenue Service will extract data on the volume of legal entities meeting the study's inclusion criteria. If the number of eligible participants exceeds 5,000, communication will be limited to no more than 4,000 legal entities, while the remaining participants will be included in the control group, meaning they will not receive any communication as part of the study.

The generation and use of these datasets are critical to achieving the project's goals. The purpose of the survey list experiment is to causally identify the effects of tax morale on the propensity to underreport economic activity, thus contributing to a deeper understanding of factors influencing the shadow economy. The goal of

the field Experiment is to apply and test theoretical findings from the survey in a real-world setting, thereby validating the research model and contributing to the development of effective tax compliance strategies.

The data generated will be valuable not only to the academic community, focusing on economic behaviours, public policy, and tax compliance, but also to policymakers and governmental agencies. The insights from the data could help in shaping more effective tax policies and compliance strategies, potentially influencing international standards and practices in economic policy and governance. Additionally, the datasets could serve as a reference for comparative studies in other geopolitical regions or in longitudinal studies tracking changes over time in response to policy modifications.

## 2. FAIR data

### 2.1. Making data findable, including provisions for metadata

In the absence of specific disciplinary standards directly applicable to all aspects of this project, the metadata schema will be designed to cover essential descriptors that facilitate the discovery and reuse of data. This schema will be aligned with the core principles and data sharing guidelines in the field of social sciences and commonly used in economic and behavioral research, ensuring compatibility and ease of integration with existing data repositories and systems.

The datasets and other project's data documentation will be posted on the designated data page on the Open Science Framework (OSF) platform (see [osf.io](https://osf.io)). Each dataset generated from this project will be assigned a Digital Object Identifier (DOI). This unique identifier ensures that each dataset can be cited properly in other research work and provides a permanent link to its location on the internet.

To enhance the findability and usability of the data, rich metadata will be provided for each dataset. The metadata will include:

**Title:** Clearly stating the project acronym to maintain consistency and ease of identification.

**Description:** Detailed account of the dataset, including the type of data, how it was collected, and its purpose within the scope of the research project.

**Date Created:** The exact date when the data was first made available.

**Date Modified:** If applicable, the date on which the dataset was last modified.

**Contributors:** Names and affiliations of all contributors involved in the data collection and processing.

To optimize the dataset for discovery and potential re-use, search keywords will be included in the metadata. These keywords will be carefully selected to reflect the core themes and methods of the project, such as "tax morale," "social desirability bias," "shadow economy," "list experiment," and "field experiment."

Metadata will be structured in a way that allows for easy harvesting and indexing by search engines and academic databases. OSF supports the integration of metadata with global scholarly infrastructure systems like CrossRef and DataCite, which facilitates wider dissemination and accessibility. The Open Science Framework platform not only hosts the datasets but also provides tools for managing the research project through its lifecycle. This includes features for collaboration, version control, and linking together related projects, making it an ideal ecosystem for maintaining the integrity of the research data and ensuring its long-term sustainability and accessibility.

By implementing these measures, the project aims to fully adhere to the FAIR data principles, ensuring that the datasets are Findable, Accessible, Interoperable, and Reusable, both for the academic community and for other stakeholders such as policymakers and public agencies.

### 2.2. Making data accessible

All data generated in this project will be made openly available, adhering to the principles of open science. The

datasets, along with the necessary code to reproduce analyses, will be posted on the Open Science Framework (OSF), a platform that supports the FAIR data principles of Findability, Accessibility, Interoperability, and Reusability. This commitment ensures that there are no intentional restrictions on the accessibility of the data; all datasets can be accessed freely by the public without any need for embargo periods or restricted access due to intellectual property considerations.

The data will be accessible through a standardized and free access protocol provided by the OSF. Given the project's alignment with open science practices, there will be no restrictions on the use of the data. The accessibility setup ensures that there is no barrier to entry for individuals or entities interested in utilizing the dataset. Furthermore, there is no requirement for a data access committee, as the data does not contain personal or sensitive information that would necessitate such measures.

Metadata accompanying the datasets will be openly available and licensed under a public domain dedication CC0, in compliance with the Grant Agreement. This approach facilitates the use of metadata without any encumbrances, thereby enhancing the usability of the data. The metadata will include comprehensive details such as the research design, sample size, hypotheses tested, and data analysis plan, all of which are crucial for users to understand the context and structure of the data.

The data and accompanying metadata will remain available for the long term on the OSF platform. This availability is ensured beyond the lifetime of the project, with specific measures in place to maintain the findability and usability of the data. The OSF platform is committed to preserving academic data indefinitely, thus providing confidence that the data and metadata will continue to be accessible in the future.

Documentation necessary to understand and utilize the data will be included alongside the datasets. This will encompass any software or tools used in the analysis, ensuring that users can replicate or build upon the research findings effectively.

By implementing these measures, the project guarantees that the generated data is not only useful to the academic community but also accessible to policymakers, governmental agencies, and other stakeholders, thereby maximizing the impact of the research. The open and unrestricted access to the data and metadata ensures a transparent and collaborative approach, fostering further research and development in the fields of tax compliance and the shadow economy.

### 2.3. Making data interoperable

The goal of the data management policy for this MSCA project is to ensure full accessibility of the data and information generated for other researchers and policymakers, fostering subsequent re-use and verification. To achieve this, the project will incorporate existing international classifications and standards relevant to economic and social research whenever new data are collected. Specifically, the methodologies, survey results, statistical analyses, and other data outputs will align with standards that facilitate their integration with existing databases and frameworks commonly used in economic and behavioral sciences. As a general principle, we will manage the project in a manner that ensures all generated databases, methodologies, questionnaires, and models can be made fully accessible to the academic community and the general public by the project's conclusion.

To protect sensitive or private information, data will be standardized and anonymized in line with the ethics guidelines, submitted within this project as deliverables D.1. and D3, which adheres to the stringent requirements of EU and national data protection laws. This process will prevent any personal data from being exposed or used in automated decision-making processes, ensuring that our data management practices uphold the highest standards of privacy and data security.

### 2.4. Increase data re-use

In line with the FAIR principles, the project allocates sufficient resources for data management, ensuring data security through robust security protocols, and addressing ethical aspects particularly in relation to data involving personal information. To facilitate the validation of data analysis and enhance the re-use of the data, comprehensive documentation will be provided alongside the datasets. This documentation will include:

**Readme Files:** Detailed guides describing the methodology used in data collection and analysis, including any assumptions or limitations.

**Codebooks:** Descriptions of each dataset, including variable definitions, units of measurement, and coding schemes to help users understand and utilize the data correctly. This data will be provided in English, Latvian, Danish, Italian and Russian.

**Data Cleaning and Analyses:** Documentation of the data cleaning processes and analytical methods used, ensuring that users can replicate the studies or adapt the methods to new data.

**Variable Definitions and Units of Measurement:** Clear explanations of all variables collected, including the units of measure used and any transformation applied to the data.

All data produced will be made freely available in the public domain to ensure the widest possible re-use. The datasets will be licensed under the Creative Commons Zero (CC0) license, allowing others to use the data without any restrictions, in alignment with the open access obligations set out in the Grant Agreement.

The data produced by this project will be structured and stored in a way that ensures usability by third parties, both during and especially after the end of the project. This involves using standard data formats and widely accepted data management practices that support long-term preservation and accessibility.

The provenance of all data collected and generated within the project will be meticulously documented using appropriate metadata standards. This documentation will detail the origin of each piece of data, the methods of data collection, and any modifications made to the original data. This level of detail guarantees transparency and aids in the establishment of data lineage, crucial for validation and reuse.

### 3. Allocation of resources

This project plans a streamlined and cost-effective approach to data management by utilizing free resources where possible. The centralized management ensures that the project's data handling will be coherent, compliant, and aligned with both the project's goals and broader open science initiatives.

The commitment to FAIR principles in the management of data and other research outputs involves several considerations, including storage, archiving, reuse, and security. For this project, which is exclusively conducted by Collegio Carlo Alberto with Andris Saulītis as the principal researcher, the use of the Open Science Framework (OSF) mitigates many potential costs. OSF offers free-of-charge services for storing and accessing data, which effectively reduces the direct expenses associated with these aspects of data management.

Nevertheless, there are anticipated costs related to ensuring open access to published research findings. These costs typically arise from processing charges levied by publishers for open access articles. Such expenses will be covered by the project budget, assuming compliance with the conditions of the Horizon Europe grant agreement.

As the sole beneficiary and data manager of the project, Andris Saulītis is responsible for all aspects of data management. This includes the cooperation with data collectors (Norstat Latvija and the State Revenue Service of Latvia), documentation, storage, and preservation of data. Responsibilities also extend to ensuring that all data management practices comply with FAIR principles and legal requirements such as GDPR. Based on the data's ongoing relevance to the research community, it is intent not keep data and all the relevant documentation public indefinitely.

The responsibility for updating the Data Management Plan (DMP) and coordinating data management activities also rests with Andris Saulītis. This centralized approach ensures consistency and accountability throughout the project's duration.

### 4. Data security

To ensure the integrity and confidentiality of data collected in this project, robust security measures are in place, designed to prevent unauthorized access to personal data or the equipment used for processing. These measures encompass both physical and digital safeguards.

Access to physical equipment, such as computers and servers involved in data processing, is strictly controlled. These devices are stored in secure, monitored facilities where access is limited to authorized personnel only. As regards the digital safeguards, all systems used for data processing are equipped with state-of-the-art security software, including firewalls and antivirus programs, to defend against external threats like hacking or malware attacks. Additionally, strong password protections are enforced, with regular updates and, where feasible, two-factor authentication is implemented to add an extra layer of security. Only personnel essential to the project, who have undergone rigorous vetting and training in data handling and security practices, are granted access to sensitive data. This ensures that all individuals who interact with the data are aware of and adhere to strict security protocols. The security measures implemented are regularly reviewed and audited to ensure they meet the highest standards of data protection and to adapt to any new security challenges or requirements as they arise.

The project utilizes the Open Science Framework (OSF) for long-term data preservation and storage, which provides reliable and secure archiving services. OSF is recognized for its robust infrastructure that ensures data is not only securely stored but also remains accessible over the long term. This service is provided free of charge, which helps manage the project's budget effectively while adhering to the FAIR principles.

All data handling and storage procedures are designed to comply fully with the General Data Protection Regulation (GDPR) and other relevant data protection laws. This compliance is supported by independent Audits, as both the pollster company involved with the survey data and the State Revenue Service managing the field experiment data are required to pass independent audits (such as the ISAE 3000 audit) that confirm their adherence to GDPR and the Personal Data Act).

Although sensitive data regarding ethnicity is collected for the survey, it is handled with the utmost care to prevent any risk of harm or stigmatization. Measures outlined in deliverable D1.4 ensure that the data is processed ethically and that the privacy of vulnerable groups is protected.

Files containing potential identifiers are not shared with the principal researcher Andris Saulītis or any other person at the Collegio Carlo Alberto. These files, kept by Norstat Latvija (for the survey data) and the State Revenue Service (for the field experiment) are securely deleted six months after data collection concludes, further minimizing the risk of data breaches and ensuring that personal data is not retained longer than necessary. By integrating these comprehensive security and compliance measures, the project ensures the protection of all data collected, maintaining the trust of research participants and upholding the integrity of the research process. This approach not only satisfies legal obligations but also reinforces the project's commitment to ethical research practices.

As regards the field experiment, the Researcher receives general, non-identifiable data from the State Revenue Service on the submission deadlines of Employer's Reports by the participating entities. This data excludes the actual reports, declarations, or any commercial or personal data, including employee information. The State Revenue Service is legally obliged to provide such information under current regulations and, particularly, the Freedom of Information Law. By not processing any sensitive personal data, the project significantly reduces the risk of privacy breaches or unethical use of data.

## 5. Ethics

This project, while adhering strictly to ethical standards, must also comply with various legal requirements affecting data sharing, particularly under the GDPR framework. The protection of sensitive personal data, especially concerning the ethnicity of respondents in Latvia, requires careful handling to prevent any risk of harm or stigmatization. The comprehensive measures implemented ensure that data sharing adheres to ethical principles and legal standards, as outlined in the project's deliverables D1.4 and D1.5.

For the survey component, informed consent is a critical element. Participants are made aware of the data use and protection policies through the initial screens and consent forms provided by the pollster company. This process, detailed in deliverable D1.3, ensures participants understand their rights, including the ability to withdraw consent at any time. This consent procedure supports ethical data sharing by ensuring participants are fully informed about how their data will be used, stored, and protected.

For the field experiment, traditional informed consent is not sought due to the nature of the study, which aims to observe natural behaviours in real-world settings without influencing participants' actions. This

methodological choice is aligned with social science research standards where the integrity of natural responses is paramount. The absence of informed consent in this context is carefully considered to maintain the authenticity of the data while ensuring no additional risk or harm to participants beyond normal experiences.

After the conclusion of the field experiment, participants will be debriefed on being part of the study. The communication will occur with participants who received message(s) during the study, informing them of their participation. Additionally, participants have the option to opt out of receiving further messages at any time during the study.

The project's approach to incidental findings, as described in D1.5, ensures that any unexpected data that could potentially identify participants will be managed without compromising confidentiality or anonymity. This includes rigorous procedures for assessing and responding to such findings, with oversight from the Ethical Advisory Board at Collegio Carlo Alberto. This applies also to cases where incidental findings have been found after the project expiration data or identified by those who have re-used data by downloading it from the OSF.

HISTORY OF CHANGES		
VERSION	PUBLICATION DATE	CHANGE
1.0	14.04.2024	Initial version
2.0	13.12.2024	Update relating to the field experiment