Beadle, Brian

Doctoral Thesis

The design and application of an agricultural sustainability index using item response theory

The Design and Application of an Agricultural Sustainability Index using Item Response Theory

Dissertation

To obtain the degree

Doctor of Economics (Dr. rer. pol.)

at the Faculty of Economics and Law, Martin-Luther-University Halle-Wittenberg

Submitted by: Brian Beadle

Reviewers: Prof. Dr. Alfons Balmann and Prof. Dr. Christoph Wunder

Defense date: 4 July, 2023

Halle (Saale)

July, 2023

Abstract

Agricultural sustainability (AS) is a topic that has been evolving since the mid-20th century, and is a key component in the UN's Sustainable Development Goals. Despite its wide use in both research and policy, there is still not a universally accepted method for its measurement. This is particularly the case for farm level AS assessment, with strong data requirements being one of the major obstacles to developing a useful farm level sustainability index. This research proposes using item response theory (IRT) to generate a farm-level AS index. IRT has a number of advantages over existing methods, the most important of which is that the proposed AS index is independent of the variables used in the IRT model. Use of a variable-independent model means farm level AS scores can be generated using readily-available data and compared across multiple regions with different sets of variables. The thesis is comprised of three scientific articles outlining the design and application of an AS index using IRT. The first article is the key contribution of the research and focuses specifically on the design of the AS index using data from the Farm Accountancy Data Network (FADN) and other secondary sources. Nine sustainability items are generated and used as inputs in a graded response model, and robustness tests are conducted on the model using a leave-one-out cross validation to test the model design, then items are systematically removed to test the model against missing data and simulate a scale linking procedure. The second and third articles then present applications of the index on two key topics within the literature on the sustainability of farming. The second article provides a descriptive analysis comparing differences in AS with respect to non-food crop production and producing on marginal lands

in the context of the bioeconomy, and the third article aims to identify causal links between AS and the conversion from conventional to organic farming methods. The key findings of the research are that (1) constructing an AS index with IRT may be a suitable alternative to existing methods and can ease the issue of data constraints, (2) farms producing a combination of non-food crops with food crops are more sustainable on average than those not producing non-food crops, and (3) while there is not enough evidence to suggest a causal relationship between AS and the conversion to organic production, organic farms are more sustainable on average with respect to every farm type and size in the data set.

Keywords: Sustainable agriculture, item response theory, non-food crops, bioeconomy, organic farming

Acknowledgements

I would first like to express my sincere gratitude to my supervisors. At IAMO, many thanks go to Prof. Dr. Alfons Balmann for his support, as well as his enthusiasm and dedication in making the institute a great place to work at. I would also like to thank Dr. Stephan Brosig for his attention to detail, thoughtful comments and discussions, and contributions to the research. At MLU, I am very grateful for Prof. Dr. Christoph Wunder for his seemingly unlimited patience and support, his willingness to go above and beyond the role of most supervisors, and for his invaluable contributions in the development and realization of this project.

Second, I want to thank my informal support system on both sides of the Atlantic. A big thank you goes to Trevor (i.e. "Nacho"), Nick, and Legion back in Wisconsin. In Europe, I am grateful for all of the friends and colleagues that were around for moral and emotional support throughout my time in Halle. Of course the list is too long to name everyone, but I particularly would like to thank Franziska and Laura for their support.

Most importantly, I want to thank my parents, Kathleen and Harvey (d. 2018) Beadle, for their continued support through all of the ideas and goals I have chased over the years, rational and irrational alike. None of this would have been possible without them.

Contents

List of Figures

List of Tables

List of Equations

List of Abbreviations

AC	Always conventional
ANOVA	Analysis of variance
AO	Always organic
AS	Agricultural sustainability
ATT	Average treatment effect
BMBF	Federal Ministry of Education and Research (Bundesministerium für Bildung und Forschung)
C	Fully conventional
CAP	Common Agricultural Policy
CSDiD	Callaway and Sant'Anna difference-in-differences
DiD	Difference-in-differences
DQ	Disqualified
EU	European Union
FADN	Farm Accountancy Data Network
GHG	Greenhouse gas
GRM	Graded response model
ha	Hectare
IRT	Item response theory
LCA	Life cycle analysis
LFA	Less favored area
LOO-CV	Leave-one-out cross-validation
MCMC	Markov Chain Monte Carlo
MIRT	Multi-dimensional item response theory
NFC	Non-food crop
O	Fully organic

OECD	Organisation for Economic Co-operation and Development
P/T	Partial or transitioning
Post-S	Post-conversion starter
Pre-S	Pre-conversion starter
Q	Quitters
S	Starters
SDG	Sustainable development goal
TWFE	Two-way fixed effects
UAA	Utilized agricultural area
USD	United States Dollar

1

Introduction

1.1 Background

In 1987, the release of "Our common future" (Brundtland 1987) played a key role in the emergence of sustainability into mainstream debate. Also referred to as the Brundtland Report, the popularity of the publication was centered around a key definition of sustainability as "meeting the needs and aspirations of the present generation without compromising the ability of future generations to meet their needs" (Brundtland 1987, p. 292). Debate at the time was mainly focused on the negative impacts of human activity on the environment and included, as examples, resource depletion from practices such as deforestation (see Geist and Lambin 2001) and overfishing (Murawski 2000). This time period also marks the emergence of communication on anthropogenic climate change; however, the topic at the time was a niche idea limited to a narrow scope of scientific articles and synthesis reports (Moser 2010).

Today, many of the environmental concerns introduced in the late 20^{th} century have been realized as emergent crises. Factors such as biodiversity loss and air pollution are critical issues in many parts of the world, and the effects of climate change from greenhouse gas emissions are detectable somewhere on the planet every day (Sippel

et al. 2020) in the form of e.g. floods and extreme temperatures (Stott 2016). What is more, other environmental issues that were unrealized a only few decades ago have now emerged as ever-increasing global threats. Plastic pollution, for example, is now found in almost every landscape on the planet, including deep oceans, deserts, and arctic snow (MacLeod et al. 2021).

While many of the sustainability discussions still focus on environmental crises, the concept has been extended in recent decades to address societal issues and the world's economies as well (Hajian and Kashani 2021). As shown in Figure 1.1, sustainability is now generally subdivided into three dimensions (also known as pillars or spheres). Social sustainability addresses basic human needs, changes in behavior to meet or exceed environmental and biophysical goals, and the continuance of social and cultural characteristics (Vallance, Perkins and Dixon 2011). In contrast, economic sustainability places emphasis on financial stability while reducing or eliminating negative externalities. Examples for reducing externalities include, among others: investments in the adoption of renewable resources, reductions of waste and pollution, and ensuring that economic growth is inclusive and equitable and does not leave certain groups behind.

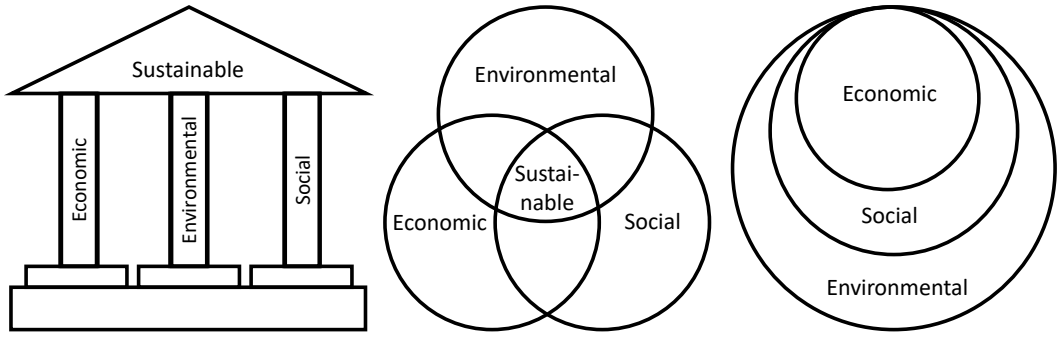

Figure 1.1: Common visualizations of sustainability and its components.

Despite clear distinctions in the goals of each dimension, an important aspect of sustainability is the idea that the individual dimensions must be considered as part of a larger holistic framework. This holistic perspective is best represented in the right panel of Figure 1.1 showing a nested hierarchy among the dimensions. From this perspective, the economy is only considered as a subset of society, and all eco-

nomic and social activities take place within a subset of the natural environmental we live in. The idea of a holistic approach to sustainability gained significant attention in 2015 when the United Nations (UN) developed a comprehensive set of 17 short- and long-term goals addressing a wide range of global sustainability concerns. Designed as an upgrade to the Millennium Development Goals (see Sachs 2012), the Sustainable Development Goals (SDGs) encompass more than 300 individual indicators among the three sustainability dimensions. The indicators cover a broad range of global goals and initiatives that include ending poverty, improving health and education, and setting targets for reducing greenhouse gas emissions and pollution (United Nations and Development 2015). A key advantage to the SDGs is the aforementioned holistic approach, where it is not possible to achieve the individual components of sustainability in isolation (Haywood et al. 2019).

Considering projections for the world population to continue growing well into the future (see Gerland et al. 2014), a particularly important aspect of the SDGs is the development of a sustainable food supply. SDG 2 aims to "end hunger, achieve food security and improved nutrition, and promote sustainable agriculture" by 2030 (United Nations and Development 2015). Within this goal, indicator 2.4.1 specifically targets agricultural sustainability (AS). The effective measurement and monitoring of AS is critical because in addition to its role as the provider of food, feed, and fiber, agriculture is a key driver of economic growth by providing employment opportunities and income generation. Further, it has a significant impact on social development and poverty reduction, particularly in rural areas. In this context, the promotion of sustainable agricultural practices can help to empower rural communities while providing for a growing population. However, agriculture is also a major contributor to many of the world's most pressing environmental challenges. For example, almost one third of global anthropogenic emissions originate in the agricultural sector, a majority of which are produced by enteric fermentation (37%) and fertilizers (29%) (Tubiello et al. 2013).

In addition to SDG indicator 2.4.1, a broad range of methods have been proposed for measuring and tracking AS in the last few decades. The methods can be categorized into four basic types: life cycle analysis (LCA), green accounting, ecological footprinting, and composite indicators (Frater and Franks 2013). SDG indicator 2.4.1 belongs to the latter group, which is perhaps the most widely used due to its flexibility in both design and application. Ewert et al. (2009), for example, proposes an assessment suited for analysis throughout Europe, and a variety of national level assessments have been developed such as the national report published by the German Agricultural Society (DLG 2016) and a group of studies out of Ireland (e.g. Hennessy et al. 2013; Ryan et al. 2016; Lynch et al. 2016; Dillon et al. 2015). Examples of smaller scale assessments include studies such as Dantsis et al. (2010) and Gómez-Limón and Riesgo (2009) who developed assessments for regions of Greece and Spain, respectively.

1.2 Problem and research question

The problem this dissertation addresses is the issue of so-called data dependency. Data dependency refers to the idea that the aforementioned methods are produced a priori, and that missing data or inconsistencies in the set of variables used for constructing the model can lead to biased results or misrepresentations of the phenomena being measured. Using the composite indicator typology as an example, these indices are generally compiled using a method such as a simple sum-score aggregation: $y = x_1 + x_2 + x_n$, where y represents the AS index and x_n are the variables used to generate it. While this format is attractive for its simplicity and transparency, the weakness is that any changes to the list of x variables will change the meaning and interpretation of y.

There are two main issues where data dependency causes problems in an AS index. The first is in the case of missing data, where one or more of the individual variables are not available for a particular observation. Missing data is a common occurrence

and can happen randomly or from a systematic flaw in the data collection design. Depending on the method used to construct the composite indicator and the severity of the problem, missing data may be imputed using some form of estimation technique. However, the accuracy of the index may be compromised if the missing data are not well handled (for an overview of missing data issues and imputation methods, see OECD 2008).

The second issue in data-dependent indices occurs when multiple data sets from different regions are merged together to create larger-scale (i.e. international) indices. This is particularly problematic for AS indices because agricultural data are generally collected at the regional or national level, and there are no international standards for data collection regarding the specific variables, measurement units, or frequencies of collection (e.g. annually, bi-annualy, etc.). The consequence is that while most agricultural data sets do share a common set of basic indicators such as farm size and output, much of the data varies across regions. This is problematic because the index produced by a data-dependent method can only contain variables that are common to all of the data sets being used.

Table 1.1 demonstrates this problem by simulating a situation where researchers would like to create a multi-regional composite indicator with data from three independent regions, each of which are responsible for collecting their own data. Individually, each region could construct relatively strong indices with four variables each: Region 1 could construct an index with a function $f(x_1, x_2, x_3, x_4)$, Region 2 could could construct an index with $f(x_2, x_3, x_4, x_5)$, and Region 3 with $f(x_3, x_4, x_5, x_6)$. However, employing a data-dependent method to construct an index for all three regions would limit the study to the function $f(x_3, x_4)$ since these are the only two variables that all three regions have in common.

Variable	Region 1	Region 2	Region 3
x_1	✓		
x_2	✓	✓	
x_3	✓	✓	✓
x_4	✓	✓	✓
x_5		✓	✓
x_6			✓

Table 1.1: Demonstration of data constraints in current AS assessment methods with multiple data sets

Considering the demonstrated issues with current AS methods, the overarching question of this research is: Can an AS assessment method be developed that eases the data dependency problem of existing methods? Corresponding to this question, the main body of the dissertation is divided into three chapters outlining the development and application of a new approach to measuring AS that is more flexible to data substitutions. The chapters are outlined in greater detail in the following section.

1.3 Structure of the dissertation

The dissertation consists of three scientific articles centered around the development of a new AS assessment method. Chapter 2 introduces the approach used to develop the proposed index and performs a series of simulations to test the robustness of the index under various data restrictions. Chapters 3 and 4 then present applications of the proposed index by addressing, respectively, the sustainability non-food crop production and producing on marginal lands in the context of the bioeconomy, and differences in sustainability between conventional and organic farms.

Chapter 2 proposes the development of an AS index using item response theory (IRT). IRT has a number of advantages over existing methods, the most important of which is that our AS index is independent of the variables used in the IRT model. This means that farm level AS scores can be estimated with readily-available data and compared across different sets of variables from multiple regions. This application uses data from the Farm Accountancy Data Network (FADN) and other secondary sources to estimate an AS index, then compares the results of the IRT estimations with known associations between the sustainability of farms relative to their type and size. In line with the literature, the model finds (1) a positive relationship between farm size and AS, (2) higher levels of sustainability for crop and mixed farming systems, and (3) below-average performance for livestock farms and vineyards. Chapter 2 further tests the sensitivity of the AS index against randomly missing data and simulate a scale linking procedure to test the flexibility in measuring multiple regions with different data sets, finding that the index is generally robust in both analyses.

After demonstrating the feasibility of the AS index in the first section of the dissertation, Chapter 3 presents the first application of the proposed index by addressing the sustainability of non-food crop (NFC) production in the context of the bioeconomy. In recent decades there has been a rapid expansion in the market for NFCs, with corresponding attention directed at producing the crops on marginal lands to mitigate competition with food. However, little is known about the relationship between farm sustainability and NFC production. As such, the chapter uses the AS index to estimate the sustainability NFC production as well as the sustainability of production on marginal lands. Three key findings from the analysis are presented. First, farms producing NFCs are more sustainable on average than those that only produce food crops. Second, this association between higher AS levels and NFC production is nonlinear: The highest predicted sustainability levels occur when farms produce between 40% and 60% NFCs with respect to total output, but sustainability levels are predicted to decrease as specialization in NFCs increases. Third,

there appears to be no difference in farm sustainability when NFCs are produced on marginal lands. The chapter concludes that the production of NFCs can be beneficial for the sustainability of the bioeconomy, with the primary factor determining the relative level of AS being the ratio of NFC output to total output.

Chapter 4 presents a second application of the model by addressing differences in farm sustainability between organic and conventional production methods. A common perception is that organic farming is more sustainable mostly because of reduced environmental pressures, while also socioeconomic benefits from price premiums and greater employment opportunities. This perception is shared by both consumers and producers alike, and is often an underlying assumption behind funding and policies directed at promoting organic farming. While there are numerous studies comparing individual components of the two methods, such as differences in productivity, land quality, and chemical use, there is a lack of attention directed at comparing the sustainability of these systems from a holistic perspective. Chapter 4 addresses this gap in the literature by using the AS index to identify differences in farm sustainability between conventional and organic farming. The assessment is conducted first as a descriptive analysis comparing differences in conventional farms, farms that are converting to organic, and fully organic farms. The comparison is further subdivided across different farm sizes and types to provide a more nuanced understanding of the differences the different groups. A difference-in-difference (DiD) regression is then employed to estimate the potential for a causal relationship between AS and the conversion from conventional to organic production. The chapter finds that organic farms are more sustainable on average than conventional farms in every farm size and type classification, and the results of the DiD model suggest that there may be a causal relationship between AS and the conversion to organic. However, the estimated effect is quite small and large confidence intervals prevent the possibility of a definitive conclusion on causality.

The dissertation concludes by first providing an overview of the findings, then discussing the current limitations and suggesting further research into the development

of the index. The general directions for further research should include taking stock of agricultural databases internationally to compile a list of common items that can be used as a basis for scale linking, testing the model internationally, and developing a multidimensional model comprised of the economic, environmental, and social dimensions of sustainability.

The code used for all statistical analyses in the dissertation is provided in the Github repository brianbeadle/sustainability_index.

2

An application of item response theory for agricultural sustainability measurement

Co-authors: Stephan Brosig and Christoph Wunder

2.1 Introduction

Agricultural sustainability (AS) is a concept that has been evolving since the middle of the 20$^{\text{th}}$ century. Interest in the sustainability of agricultural systems was focused primarily on environmental concerns in 1950's and 60's (Pretty 2008), which expanded into a perspective of ecological interaction in the 1980's (Edwards 2020). It has since grown further to recognize the main principles of sustainable development (economic, environmental, and social) and is now a component in the United Nation's Sustainable Development Goals (specifically, SDG indicator 2.4.1: FAO 2018*b*).

Despite its increasing importance in both policy and practice, there is still not a universally accepted method for measuring AS. A broad range of variable-based tools have been developed for the task (Marchand et al. 2014; Zhen and Routray 2003), which are generally a collection of variables used to assess farm sustainability. As an example, the Monitoring Tool for Integrated Farm Sustainability (MOTIFS) framework (Meul, Nevens and Reheul 2009) is comprised of several ecological variables to monitor the performance of Flemish dairy farms. However, a significant drawback to existing methods is that they require a large set of variables that are rarely readily available (Frater and Franks 2013; Kelly et al. 2018; Zhen and Routray 2003). This is problematic because it often leads to the omission of important variables (Terres et al. 2015) and misinterpretations of the phenomena being measured (see OECD 2008).

Our research contributes to the literature by providing proof of concept for using item response theory (IRT) for AS measurement. Generally, IRT models can be used to empirically analyze the relationships between observed items and a latent variable. In contrast to existing AS indices, IRT model results are independent of the individual variables selected for the estimations (assuming the variables have certain properties) (see Lord 1953). This variable independence means that all items reflect the latent variable in a comparable way, and the latent variable values of subjects

(farms) do not change (except for random error) when some items are replaced by others or missing entirely. Variable independence is an important advantage over existing AS assessment methods because it (1) facilitates comparing AS scores across (samples from) different surveys comprising different variable sets[1], e.g. from different countries, and (2) it means that reliable AS indices can still be estimated for observations with some missing values.

The use of an IRT model to construct an AS index has another important advantage: it allows us to quantify the uncertainty of the AS measurement by identifying credible intervals that contain the true AS value with high probability. This way, we can take into account, for example, that the precision of the measurement may vary across different types of farms or regions.

There are two core challenges to creating a new index for AS measurement, the first of which involves a proper definition of the concept of AS. As a result of its definitional flexibility (Frater and Franks 2013; Franks 2010; White 2013), there is significant variation in how AS is defined and thus pursued in policy-making (Waltner-Toews 1996; Binder, Feola and Steinberger 2010). The working definition we use for AS follows that of the Brundtland Report as a farm that is "meeting the needs and aspirations of the present generation without compromising the ability of future generations to meet their needs" (Brundtland 1987, p. 292). From this definition, we can derive a set of variables that show the extent to which a farm is sustainable. For example, suggested goals to achieve this definition include the promotion of biodiversity on farms (Roy and Chan 2012), enabling stable productivity over time to cope with natural or economic shocks (Conway and Barbie 1988), and maintaining the well-being of the farmers and their families while meeting society's demands, values, and concerns (Lebacq, Baret and Stilmant 2013; Diazabakana et al. 2014).

[1]The basic requirement for these comparisons is that the data sets share a minimum set of common items, which can then be used to align the scales of the different estimations in a process called scale linking. For an introduction to the procedure in the context of education assessments, see Meyer and Zhu (2013).

The second challenge involves validation of the model. We conduct three main forms of model validation: First, we use leave-one-out cross-validation (LOO-CV) to find the best parameterization of our IRT model. We then use the results of the model to compare the distributions of farm level AS scores with existing knowledge on associations between AS and farm size and type. Finally, we conduct a series of sensitivity analyses to (1) test the robustness of the model results in the presence of missing data, and (2) simulate a concurrent scale linking procedure to evaluate the potential for expanding the AS index to larger geographic scales with different data sets.

Results of the model estimations suggest that the IRT model represents a feasible framework for estimating a farm level AS index. Our AS index is characterized by a positive relationship between farm size and AS, higher than average sustainability performances for crop and mixed farms, and below average performance for livestock farms (milk, grazing livestock, etc.). These trends are also reflected geographically, as the regions in the eastern part of Germany have the highest average sustainability scores, and areas with high livestock production are among the lowest. We additionally find that the results are robust even for farms with missing items, as well as in the case of scale linking simulations.

The remainder of the paper is organized as follows: After a brief literature review in Section 2.2, Section 2.3 provides a short background on IRT and the assumptions used in creating the model. Section 2.4 then discusses the data, items, and IRT model estimations used to generate the index. The model results are discussed in Section 2.5, and Section 2.6 concludes.

2.2 Review of existing methods

This section summarizes the four main methods of AS assessment outlined in Frater and Franks (2013), the first three of which belong to a larger group of environmental accounting methods and include life cycle analysis (LCA), green accounting, and

ecological footprinting. Conceptually, these approaches are designed to quantify both the direct and indirect effects of human activity (Patterson, McDonald and Hardy 2017), generally in the creation of a specific product, process, or service. Haas, Wetterich and Geier (2000) provides an example of a farm-level LCA that was developed to assess the environmental impact of farms in southern Germany by looking at factors such as resource use and consumption (including energy and chemicals), and impacts to biodiversity and animal welfare. A detailed overview of green accounting methods currently used to analyze European farming is provided by Halberg, Verschuur and Goodlass (2005), a majority of which focus specifically on energy, chemical, and nutrient usage. Finally, Blasi et al. (2016) present a farm-level application in ecological footprinting by comparing the inputs (e.g. energy, chemicals, and labor) and outputs for durum wheat production for a single farm in Italy.

There are a number of challenges presented by these approaches. First, because the methods are framed in environmental accounting, there is little attention given to the socioeconomic aspects of farming. Ecological footprinting is perhaps the weakest in this aspect because it only considers a narrow view of environmental impacts by measuring land use, emissions, and chemical use (Frater and Franks 2013). In contrast, LCAs are more flexible to a broader range of sustainability aspects through the development of social LCAs (SLCAs) (e.g. Prasara-A et al. 2019) and life cycle costing (LCC) (e.g. Baquero et al. 2011) as an economic approach to complement the existing LCA framework. However, data availability is still a central issue to these expansion efforts, as e.g. Chen and Holden (2017) and Martínez-Blanco et al. (2014) both cite a lack of key social variables in constructing agricultural SLCAs.

Methodologically, another drawback to these methods is the monetization of environmental impacts. The design of the frameworks are focused on quantifying and expressing the environmental impacts of agriculture as a monetary value, which is then used as an avoidance mechanism (Beckenbach, Hampicke and Schulz 1988; Krieg, Albrecht and Jäger 2013). Such an approach is highly subjective because

it is based on willingness to pay (Reap et al. 2008), which can vary widely based on location. Further, the monetization of environmental impacts imply that any consequences to farming can be compensated through financial means.

Composite indices are the fourth method identified by Frater and Franks (2013). This approach is used most often in AS research due to its flexibility in terms of methodological choices, as well as its ability to incorporate all sustainability dimensions. The basic framework for a composite index involves a collection of variables that are aggregated to explain a complex issue (OECD 2008). Index results are often reported as a simple sum-score aggregation (e.g. Vitunskiene and Dabkiene 2016), though more complex aggregation methods exist such as geometric mean aggregation (e.g. Talukder et al. 2017) or multi-criteria decision making (MCDA) (e.g. Gómez-Limón and Riesgo 2009). Despite significant debate in the literature over the respective strengths and weaknesses of each method, they generally produce similar statistical results (for methods comparisons in AS assessment, see Gómez-Limón and Riesgo 2009; Gómez-Limón and Sanchez-Fernandez 2010; Gómez-Limón, Arriaza and Guerrero-Baena 2020). It is noted that some studies choose instead to report the results as a "dashboard" of the individual variables. Examples of this approach include a group of studies in Ireland (e.g. Hennessy et al. 2013; Ryan et al. 2016; Lynch et al. 2016; Dillon et al. 2015) and the SDG 2.4.1 indicator (FAO 2018*b*), which reports both as a dashboard and an aggregated value. While useful, a dashboard can lead to more complex interpretations of the results, particularly when a large number of variables are used.

Regardless of the aggregation technique, data dependency is still a critical drawback of composite indices. The Farm Accountancy Data Network (FADN) often serves as a foundation for the assessments, but sustainability-oriented variables are limited since the data set is designed specifically as a financial database and excludes many environmental and social components needed for a robust sustainability index (see Kelly et al. 2018). As a result, a variety of alternative data sources have been merged with FADN to improve the design of the index. Such sources include na-

tional databases for Ireland (Buckley et al. 2015; 2016) and the UK (Westbury et al. 2011), the Land Parcel Identification System (LPIS) (Latruffe and Piet 2014), tax records (Latruffe and Mann 2015), and the International Farm Comparisons Network (IFCN) (Thorne and Fingleton 2006). Alternatively, studies such as Batalla, Pinto and Del Hierro (2014) and Sulewski, Kłoczko-Gajewska and Sroka (2018) use primary data to strengthen their respective analyses. The integration of both primary and secondary data is beneficial for the robustness of the composite indices; however, this can be time consuming, costly, and difficult to expand beyond the current sample groups.

2.3 IRT overview and assumptions

In this study, we propose measuring farm-level AS using an IRT model. This approach addresses three important drawbacks of existing approaches, as the IRT model allows: (1) calculation of an AS index from data sets with some missing values, (2) comparisons across different surveys with different variable sets, and (3) quantification of uncertainty. With origins in educational and psychological testing (for an overview, see van der Linden and Hambleton 1997; Cai et al. 2016), the general idea of an IRT model is that a set of categorical or dichotomous observed indicators can be used to measure a continuous unobserved latent construct on a common scale. In educational testing, for example, IRT models are used to model the relationship between a set of observed categorical responses to exam questions and the unobserved ability of the examinee. Other applications of IRT models are in medical research to assess health outcomes (Hays, Morales and Reise 2000) or to identify clinically meaningful subgroups of high-risk patients (Prenovost et al. 2018). In terms of socioeconomic applications, IRT has been used in poverty research to construct a deprivation scale (Cappellari and Jenkins 2007), and as a means to estimate wealth (Vandemoortele 2014).

The use of IRT for AS measurement has several advantages over existing methods. The first, and most critical in our view, is the aforementioned variable independence. In exploiting this advantage, we can design an index using local secondary sources and the model results can still be comparable on a larger scale (i.e. internationally). For example, multiple AS indices could be generated and compared using FADN data in Europe and USDA Census of Agriculture data in the United States. Second, the IRT model makes explicit the assumptions underlying the construction of the AS index. This is of particular importance when compared to composite indices, which leaves the assumptions underlying the method implicit (for a similar argument in the context of the measurement of deprivation, see Cappellari and Jenkins 2007). Finally, IRT allows - and accounts for - varying degrees of difficulty (Yount et al. 2019) and discrimination for each item (see Section 2.4.3 for an explanation of the item easiness and discrimination parameters) (Nguyen et al. 2014). This allows for more flexibility in the choice of items, as the IRT model does not assume that all the items are equally difficult for a farm to achieve, or that the items are equal with respect to their ability to explain the latent trait.

The IRT model requires two assumptions. The first assumption is the reflectivity of the model, where the items are considered as manifestations of the latent construct we are attempting to measure and cannot directly influence the underlying latent construct (Peterson, Gischlar and Peterson 2017). While this requirement has faced criticism in more recent years (see, e.g. Bollen and Diamantopoulos 2017), we follow this idea by assuming that (1) the level of AS for any given farm represents the farmer's mindset along with the structural conditions of the farm, and (2) this property impacts all of the observed items used in the model. We illustrate these assumptions in Figure 2.1, where the farmer's mindset refers to factors such as the farmer's knowledge or beliefs with regard to sustainability, their willingness to invest in technology, etc. Structural conditions refer to aspects of the farm itself that are constant or difficult to change within the farm, but vary across e.g. regions or farm types. Such conditions include, for example, local soil or weather conditions,

or differences in machinery requirements for different farm types. Together, the farmer's mindset and the structural conditions are what determines the items we include into the IRT model, meaning that the item values reflect farm AS.

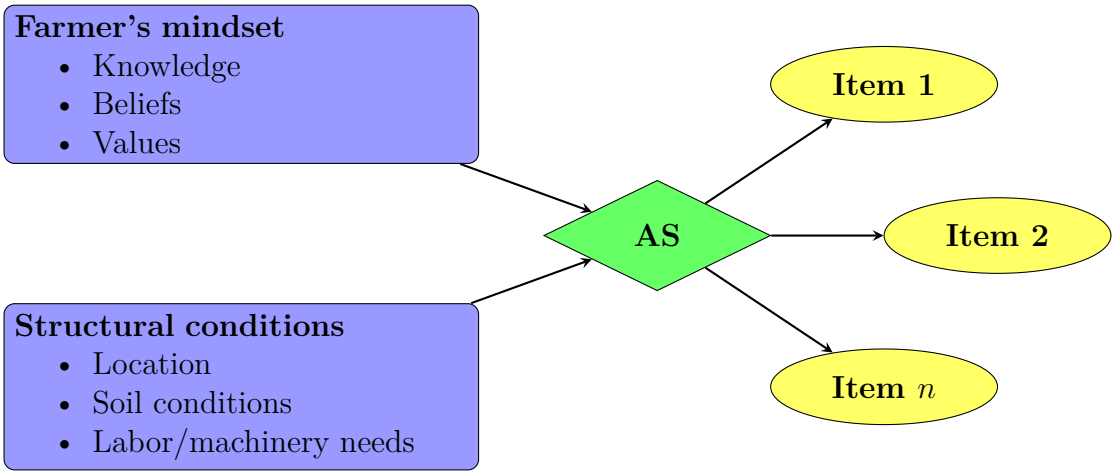

Figure 2.1: Flowchart depicting the reflectivity of the AS index with the decision to switch to organic as an observed component of the farmers' mindset

The second assumption we adhere to refers to unidimensionality in the model construct. IRT models are considered to be unidimensional when only one dominant component or factor is being measured by the model (Hambleton, Swaminathan and Rogers 1991). This term, however, can be misleading because a unidimensional trait often requires multiple processes or skills (Ziegler and Hagemann 2015). Further, testing for unidimensionality is not a straightforward process (Ziegler and Hagemann 2015), and it is unclear as to the severity of the problem if the assumption is violated. Zhang (2008), for instance, finds that models are generally robust against such a violation when secondary latent traits are present in the model. Since the definition, testing, and consequences of the assumption are rather ambiguous, we choose to approach the topic conceptually by assuming that AS is a unidimensional latent trait comprised of sub-processes following the economic, environmental, and social aspects of sustainability. However, this does not preclude the possibility of extending our approach so that the IRT model reflects multiple latent constructs. While less common, multidimensional IRT models are possible. Reckase (2009) provides and example of such a model.

2.4 Methods

In this section, we present the steps in producing our AS index using item response theory. After introducing the data used for the model, Subsection 2.4.2 provides an overview of the items used in the index. Further descriptions and calculations for each item are provided in Appendix A.1 and A.2. The final subsection provides the model statement for the item response model used to generate the index.

2.4.1 Data sources

Similar to other studies discussed in Section 2.2, we use farm-level FADN data in conjunction with other secondary sources. We use a sample of farms reporting to the system in Germany for the accounting year of 2013, as data for this year are the most recent available to us. The final sample size of 8,928 farms. We then supplement the information from FADN with data from Destatis for fuel and electricity prices, UC-Berkeley (2020) for energy unit conversions, Dämmgen (2009) for Germany-specific animal weight data, and the Bundesarbeitsagentur (Antoni, Ganzer and von Berge 2019) for regional median wages.

2.4.2 Item selection

In this subsection, we provide an overview of the nine items generated for the model. Table A.3 in Appendix A.1 provides a list of the items with general descriptions. Further information on the item selection choices and calculations can be found in Appendix A.1, and descriptions of the category thresholds with relative frequencies can be found in Appendix A.2.

Profitability is included in SDG 2.4.1 (FAO 2018*b*) to assess the economic viability of the farm and is considered by Schaller (1993) to be a requirement for AS. The item can be indicative of the quality of life for the farmers and their families (defined as

"livability" by Spicka et al. 2019), and is one of the drivers behind inter-generational succession of the farm (see Glauben, Tietje and Weiss 2005). The item for solvency combines both short- and long-term aspects of AS (see Slavickiene and Savickiene 2014) and signals the ability for farmers to continue their operations (Whitehead et al. 2016) by repaying their debts through the sale of assets if financial difficulties occur (Zwilling and Raab 2019). Similar to profitability, economic diversity is another key component in SDG 2.4.1 (FAO 2018*b*) and indicates the ability for the farm to recover from external shocks (e.g. weather disruptions, price fluctuations, etc.). Finally, multi-factor productivity (MFP), also referred to as total factor productivity (Goodridge 2007), is defined by Eurostat (2022*b*) as a means of measuring economic performance by comparing the amount of output relative to the amount of combined inputs used to create said output.

We use three items to account for the environmental soundness of the farm. Despite a variety of different estimation methods[2], some form of pesticide expenditure is a central component in most AS assessments. It is estimated that less than 0.1% of total pesticides used actually reach the target pests, with the vast majority causing environmental contamination and adverse risks to public health (Pimentel 1995). Thus, the reduction of plant protection products is considered as a high priority goal in the path to AS (Lechenet et al. 2014*a*). Estimations for GHG emissions are also common in AS indices and frameworks (e.g. van der Meulen et al. 2014; Ryan et al. 2016; Vitunskiene and Dabkiene 2016; Dillon et al. 2015; Lynch et al. 2016). As of 2019, almost 8% of emissions in Germany are from agriculture, with methane and nitrous oxide from animal husbandry and soils (respectively) making up a majority of total emissions (Rösemann et al. 2021). Thus, a reduction in agricultural emissions is critical, as Germany aims to reduce emissions by at least 80% by 2050 (compared to 1990 levels) (Weingarten et al. 2016). The final environmental item is land ecosystem quality. Approximately half of all land in Germany is used for agricultural purposes

[2] As examples, Longhitano et al. (2012) calculates the indicator as expenditure per hectare, van der Meulen et al. (2014) and Westbury et al. (2011) use the expenditure variable provided by FADN to estimate physical quantities

(Destatis 2020), and intensive production impacts the quality of air, water, soils, and biodiversity (Schiefer, Lair and Blum 2015).

To reflect the social desirability of the farm, we chose items that reflect the farms' contribution to rural economic development. Sullivan (2003) contends that a farm's ability to support agricultural workers and other businesses in the community is a key aspect of agricultural social sustainability. The item for the provision of employment reflects this by measuring the ratio of expenditures on wages and contract work, with total output of the farm in the denominator as a means to control for farm size and reflect the employment intensity of the farm. We then include an item for wage ratio, which compares the relative differences between the average wages paid on the farm relative to the median wage in the region (NUTS 3 level). The indicator is included as a proxy to estimate the extent to which agricultural workers in the region are able to cope with economic shocks and maintain a sustainable standard of living.

2.4.3 IRT application

In this section, we introduce the IRT model as an alternative approach to existing AS assessment methods. We use the graded response model (GRM) to summarize the AS items (for an overview of item response modeling and the GRM, see Bürkner 2019; Samejima 1997*a*). The outcome of the ith sustainability item for farm j, y_{ij}, is measured with C categories representing the ratings of sustainability on an ordered scale. For that purpose, the continuous variables are transformed into ordinal scales with $C = 4$ categories taking on the labels "very unsustainable", "unsustainable", "sustainable", and "very sustainable"[3]. Using ordered categories and the GRM allows us to combine the information from different continuous variables with quite different empirical distributions. Table A.3 in Appendix A.2 provides details about the definition of the categories.

[3]In a sensitivity analysis, we use $C = 3$ categories with labels "unsustainable", "neutral", and "sustainable".

In the GRM, the probability of a particular category is

$$P(y_{ij} = c|\boldsymbol{\tau}, \psi_{ij}) = F(\tau_c - \psi_{ij}) - F(\tau_{c-1} - \psi_{ij}), \tag{2.1}$$

where F denotes the CDF of the standard logistic distribution and $\boldsymbol{\tau}$ is a vector of $C-1$ unknown thresholds.[4] The distributional parameter ψ_{ij} can be expressed as a function of farm parameters, θ_j, and item parameters, ξ_i:

$$\psi_{ij} = \theta_j + \xi_i. \tag{2.2}$$

The farm parameter θ_j represents the latent construct of agricultural sustainability (i.e. the AS index). The larger the value of the AS index is, the larger the probability that the farm is classified as "very sustainable" for each of the items and the more likely the farm is to fit the definition of a farm that is "meeting the needs and aspirations of the present generation without compromising the ability of future generations to meet their needs" (Brundtland 1987, p. 292).

The specification of the IRT model in equation 2.2 relies on the unrealistic assumptions that the effect of farm-specific agricultural sustainability on each item probability is constant. To relax this assumption, we introduce an item-specific discrimination parameter, α_i, that reflects that some items can better differentiate among farms with different degrees of agricultural sustainability than others:

$$\psi_{ij} = \alpha_i(\theta_j + \xi_i) = \alpha_i\theta_j + \delta_i. \tag{2.3}$$

We fit the model in a Bayesian framework using the `brms` package (Bürkner 2017) in R (R Core Team 2021), which allows to interface with the probabilistic programming language Stan (Carpenter et al. 2017).[5] We use weakly informative prior distributions that help to improve convergence of the sampling algorithm while they do not

[4] The threshold parameters τ_1, τ_2, and τ_3 are freely estimated whereas τ_0 and τ_4 are set to $-\infty$ and $+\infty$, respectively.

[5] The model was fit using R version 3.6.3, `brms` version 2.16.3, and Stan version 2.21.0.

have a strong influence on the posterior distribution because of the large amount of sample data available from the FADN. Following Bürkner (2019), we impose two constraints to ensure identification. First, we restrict the discrimination parameters α_i to be positive because a change of the sign of α_i can be offset by a change of the sign of $\theta_j + \xi_i$. This constraint is not overly restrictive because higher categories of the items represent always a higher degree of sustainability.[6] Second, we fix the standard deviations of the farm-specific parameters to 1, as the multiplicative relationship of α_i and θ_j does not allow to freely estimate the scale of the farm-specific parameters. That is, the scale of the farm-specific parameters is determined by the scale of the discrimination parameters.

2.5 Results

The following section presents the results generated by the IRT model. Subsection 2.5.1 uses leave-one-out cross-validation to compare the two IRT models introduced in Subsection 2.4.3. Subsection 2.5.2 provides an overview of the item parameter estimations, where we explain the interpretation and use of the parameters and provide an example of how they affect the index calculations. Subsection 2.5.3 then compares patterns in the predicted probabilities of the sustainability categories with respect to (1) farm type, (2) farm size, and (3) geographic location. Finally, in Subsection 2.5.4 we conduct two sensitivity analyses of the index to test the robustness of the index results in the presence of missing items. The first analysis randomly drops items for portions of the sample and compares the results with the full sample estimations. In the second analysis, we simulate concurrent scale linking procedures (see Meyer and Zhu 2013) by splitting the data set into two samples and dropping different items from each group.

[6] A negative sign of α_i would imply that a higher degree of AS is associated with a decrease in the probability for the category "sustainable" and an increase in the probability for the category "unsustainable".

2.5.1 Model comparison

We compare the restricted IRT model without a discrimination parameter (equation 2.2) and the unrestricted IRT model with a discrimination parameter (equation 3.4). The purpose of the comparison is to determine which model has a better fit. We use approximate leave-one-out cross-validation (LOO-CV) to measure the predictive accuracy for the purpose of model comparison (Vehtari, Gelman and Gabry 2017). LOO-CV works by leaving out one data point from the training set, fitting the model to the remaining data points, and then scoring the model on the left-out data point. This process is then repeated for each data point in the training set.

LOO-CV helps us identify potential issues with overfitting or underfitting our data while also providing an objective measure for choosing between models of different complexities. This measure is the expected log pointwise predictive density (ELPD), which is an estimate of the out-of-sample predictive accuracy for each model. The ELPD can be used to assess how well a probabilistic model can explain an observed set of data points. The model with the highest ELPD best explains our observations while minimizing overfitting or underfitting issues.

Table A.4 in Appendix A.3 presents the results of a comparison between the two models, one with and one without the discrimination parameter. The ELPD values for both models are shown along with their respective standard errors. Furthermore, the difference in ELPDs and its corresponding standard error is also reported. From these findings, we can conclude that the model including the discrimination parameter fits the data substantially better than that without the discrimination parameter; this is indicated by the substantially larger ELPD of the unrestricted model. Therefore, due to its superior performance on predictive accuracy, we proceed with the model that includes the discrimination parameter.

2.5.2 Item parameters

The interpretation of the item parameters is based on a visual representation of the easiness and discrimination parameters, with the aim of assessing the model design and fit of the items. Figure 2.2 presents the means and 95% credible intervals of the parameter estimations, with the item numbers corresponding to those provided in Table A.3 in Appendix A.3.

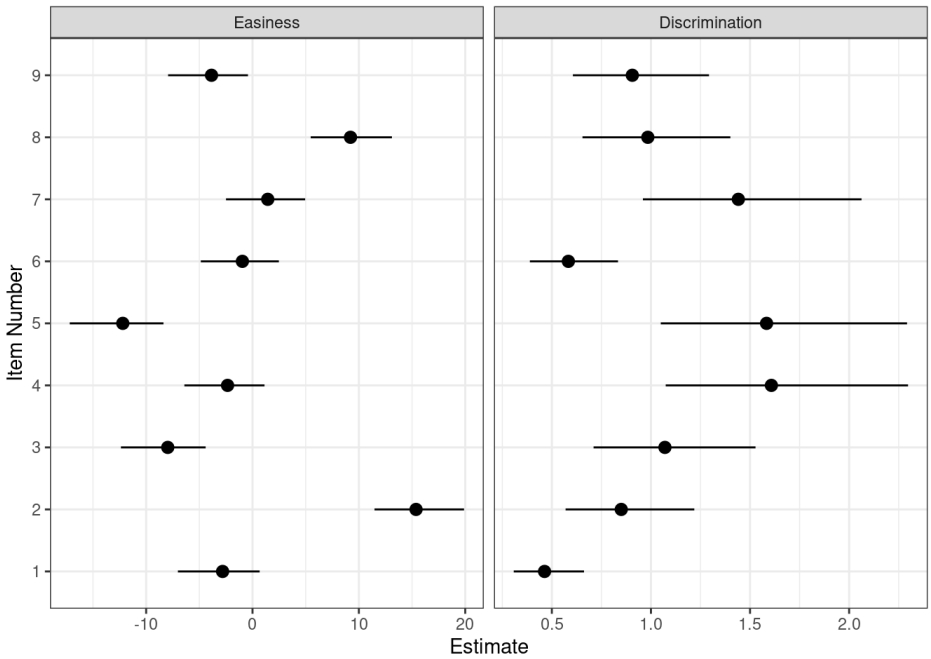

Figure 2.2: Easiness ξ_i and discrimination α_i parameter estimates with 95% credible intervals

The discrimination parameter is useful for evaluating the choice of (1) the model design and (2) the items used for the index. We first visually inspect the item-specific discrimination parameters to check the fit of the model. The observed variation in α with largest values more than threefold the size of the smallest (Figure 2.2) validates the use of the item-specific parameters introduced in equation 3.4 for our model. This finding confirms the results from the model comparison using LOO-CV carried out in the Subsection 2.5.1.

With respect to item selection, α can be used as a guide to include or omit certain items from the model based on their respective discriminatory power. This is especially useful if future research included a larger bank of test items to choose from.

In this regard, the model suggests that the provision of employment and economic diversity are the most important items to include in the model because they are the best at differentiating between the farms with posterior means of approximately $\alpha = 1.6$. In contrast, the decision could be made to either omit or replace profitability and the expenditure on pesticides from future models since they have the weakest discrimination values ($\alpha_1 = 0.4$ and $\alpha_6 = 0.6$, respectively). Items could be replaced with substitutes that measure a similar phenomenon (e.g. farms' profit margins) but have a higher discrimination value.

Similarly, item easiness ξ_i (left panel of Figure 2.2) can be used to determine the appropriateness and accuracy of the items in the model. For reference, the values produced by the model can be confirmed by comparing the relative frequencies of each category c found in Table A.3. The results suggest that (in Germany) it is easy for farms to remain highly solvent, with a vast majority of those in the sample (almost 70%) maintaining a debt to asset ratio of less than 0.3, and thus resulting in a very high easiness value of approximately $\xi_2 = 15$. In contrast, we find that very few farms score well in the provision of employment item, with an easiness value of $\xi_5 = -12$. Similar to the discrimination parameter, future research could make the decision to exclude both solvency and the provision of employment from a larger test bank because there are categories in these items with almost no information. As Table A.3 shows, there are only 207 of 8,928 farms that are classified as insolvent, and only 35 farms with the highest level of employment relative to output. However, we contend that these items may be beneficial to the model by capturing unique characteristics of the latent trait that are not common in the sample.

As an example of how the extreme values in the easiness parameters affect the predicted probabilities of a farm achieving a particular category of a given item, we provide an interpretation of the AS index using three sample farms. We show the probability of the sample farm j achieving the highest category ("very sustainable") for the easiest item (solvency, $i = 2$) and the most difficult item (provision of employment, $i = 5$). As shown in Table A.5 in Appendix A.3, a farm with an AS

score of $\theta = 1.57$ has a 77% chance of being classified as very sustainable for solvency, decreasing to a 69% chance when the farm has a AS score of $\theta = -0.58$. The same decreasing trend is also found with respect to the provision of employment item; however, due to its exceedingly low value for the easiness parameter, the probability of a random farm achieving the highest sustainability category is virtually zero. The results imply that these items may be the most useful in the model for farms in the far sides of the distribution, as e.g. only the least sustainable farms are likely to insolvent, and only the most sustainable farms are likely to have a high provision of employment.

2.5.3 Modeling farm covariates

In this subsection, we generate average predicted probabilities with respect to farm size, farm type, and region (NUTS 2 level). We additionally review the literature to identify mechanisms that might contribute to the identified patterns of the predicted probabilities across groups of farms with different characteristics.

Farm type comparisons

Figure 2.3 presents the posterior means of the predicted probabilities for all sustainability categories by the TF8 farm type classifications, and Table A.6 in Appendix A.3 presents the results numerically with standard errors of the estimates. TF8 classifications refer to a group of codes denoted the type of agricultural specialization for each farm (see European Commission 2000). Our findings suggest that field-crop farms and mixed farms are the most sustainable on average, with the predicted probability of a random farm of these types achieving a sustainable category of "very sustainable" approximately equal to 0.17 ($s.e. = 0.06$) and 0.16 ($s.e. = 0.04$), respectively. Correspondingly, these farms have the lowest likelihood of being classified as "very unsustainable" with predicted probabilities of, respectively, 0.22 ($s.e. = 0.06$) and 0.24 ($s.e. = 0.07$). In contrast, all of the livestock-based farm types (i.e. milk,

other grazing livestock, and granivores) have relatively low probabilities of achieving a "very sustainable" classification, and wine farms rank the lowest with the predicted probability of a "very sustainable" classification approximately equal to 0.09.

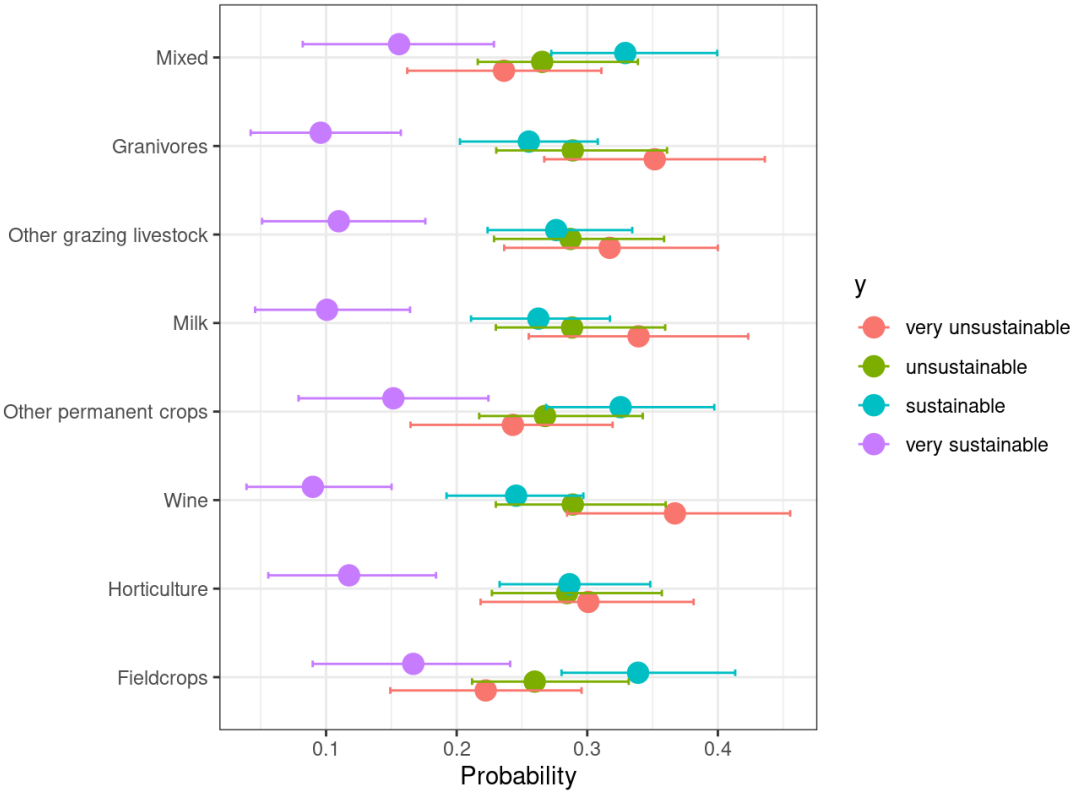

Figure 2.3: Predicted probabilities with 80% credible intervals, by farm type.

In general, we find that the results of the analysis are consistent with known associations between the type of farm and its relative sustainability. It is likely that the key mechanisms contributing to the gap between fieldcrops and the livestock categories are a result of differences in productivity (see Woods 2019) and much higher levels of emissions in livestock production (i.e. methane and nitrous) (see e.g. Haenel et al. 2020). Consequently, authors have more recently been promoting mixed farming as a way to improve nitrogen balances and increase productivity (Mosnier et al. 2022), and as a general improvement to ecosystem services (Martin et al. 2016). Our model is consistent with these earlier findings about the benefits mixed farming systems on AS. While the exact mechanism behind the low values for vineyards is unclear, viticulture is commonly recognized as an area for GHG emissions reductions (see e.g. Marras et al. 2015; Vázquez-Rowe, Rugani and Benetto 2013).

Economic size class comparisons

We next compare the predicted probabilities by economic size class, which is defined by FADN as the annual standard output[7] of the farm (in €) and is calculated as a categorical variable (see European Commission 2000). As shown by the predicted probabilities displayed in Figure 2.4 (Table A.6 in Appendix A.3 provides numerical values of the predictions with standard errors), we find a positive relationship between farm size and AS. Farms with a standard output of 50,000-100,000€ per year are the least likely to be classified as "very sustainable", with a predicted probability of 0.11, which gradually increases to 0.27 for farms with a standard output of more than 3,000,000€ per year. Correspondingly, the probabilities of a farm in the same size classes being classified as "very unsustainable" are 0.30 and 0.13, respectively. There is also a divergence within the middle sustainability categories as farm size increases: the predicted probabilities of the "unsustainable" and "sustainable" categories are very similar in the lower size classes, but the likelihood of a "sustainable" classification increases with farm size while the likelihood of an "unsustainable" classification decreases.

Similar to the results of the farm type analysis, these findings generally correspond with known associations between farm size and its relative sustainability in existing literature. Authors often cite economic benefits to larger-scale farming that include, as examples: improved productivity, profits, and solvency (van der Meulen et al. 2014). Additionally, large farms can comparatively have environmental advantages (Kirner and Kratochvil 2006) such as lower chemical usage per hectare (Ren et al. 2019).

[7]Eurostat (2023*b*) defines standard output as "the average monetary value of the agricultural output at farm-gate price, in euro per hectare or per head of livestock".

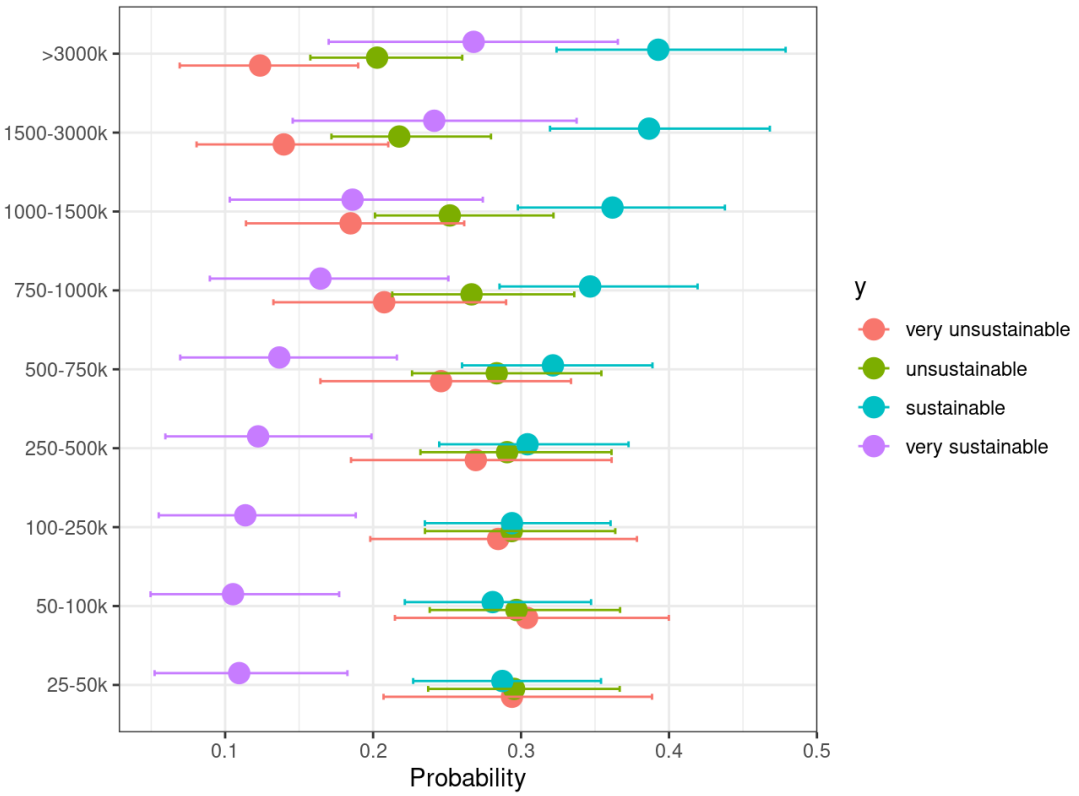

Figure 2.4: Predicted probabilities with 80% credibility intervals, by economic size class.

Regional analysis

To conclude this subsection, we look at the average predicted probabilities by NUTS 2 region and find that the regional averages generally reflect the differences in farm size and type. Figure 2.5 shows the average predicted probability of a random farm achieving the "very sustainable" category, and Table A.8 in Appendix A.3 additionally includes standard errors of the probabilities and 80% credible intervals. The results show that the highest average probabilities are in the eastern part of Germany (formerly GDR), an area primarily comprised of large-scale fieldcrop farms. In the southern regions of the country (i.e. Bavaria and Baden-Württemberg), small family farms are the most prevalent type of agricultural holding, which is assumed to be the reason for lower predicted probabilities compared to the eastern regions. Finally, the regions with the lowest predicted probabilities in the north-western regions of the country are likely due to a higher proportion of livestock farming.

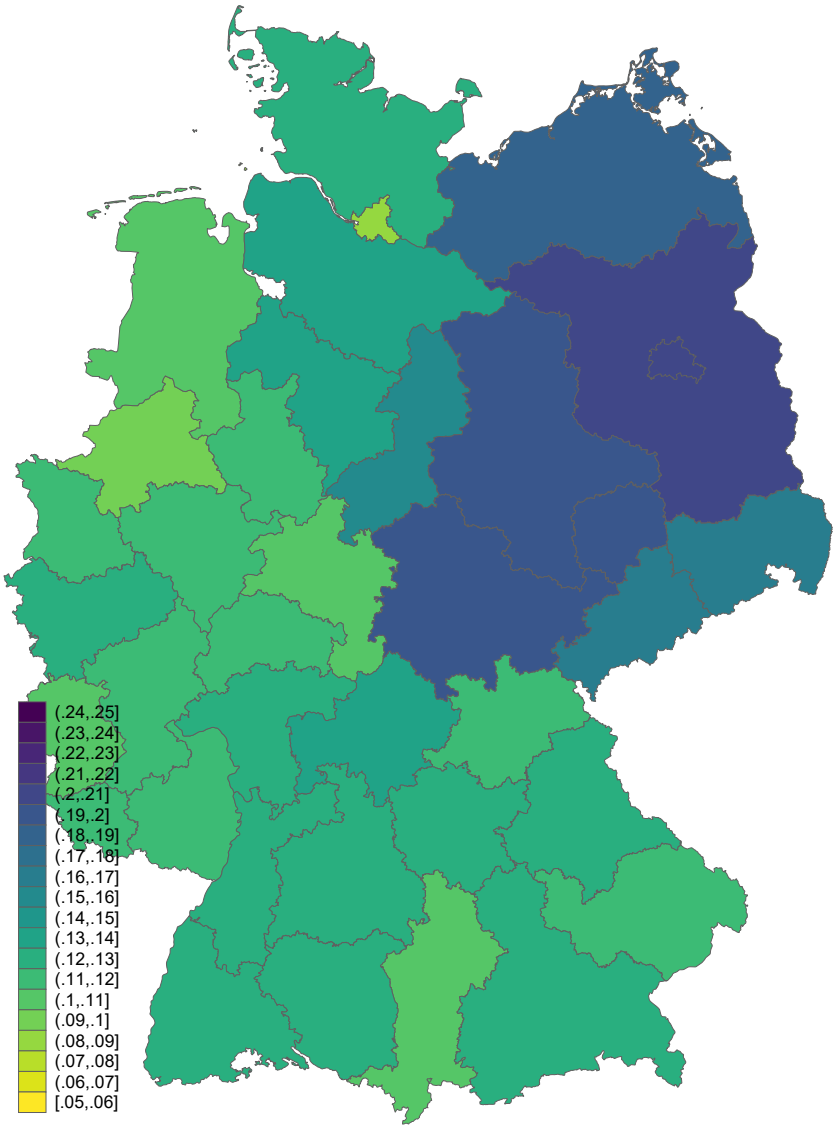

Figure 2.5: Regional averages of the predicted probability for a random farm achieving the "very sustainable" category

2.5.4 Sensitivity analysis

The following subsection presents the results of two sensitivity analyses. In the first analysis, we test the robustness of the index in the presence of missing items by randomly dropping items from different proportions of the sample and comparing the index estimations with the complete index (i.e. control group). The second analysis then simulates a concurrent scale linking procedure by splitting the sample into two groups and systematically removing one item from each group.

Missing data tests

For the first analysis, we conduct nine tests where we randomly omit up to three items for 10%, 30%, and 50% of the farms in the sample. The purpose of the exercise is to test the accuracy of the index results in the case where a sample contains randomly missing observations. An example of this situation would be in developing an index using survey data with incomplete responses. We compare the latent sustainability scores (θ) of two samples, where one sample contains missing items and the other is the complete sample we use for all previous analyses (i.e. control group). Comparisons between the groups are performed using correlation coefficients and scatter plots (see Table A.9 and Figure A.2 in Appendix A.3).

We find that in all nine tests, the results of the index are robust against randomly missing items. When items are randomly removed from 10% of the sample, the comparisons with the control group are nearly identical. The correlation coefficients are equal to 0.9918 when one item is removed, 0.9868 when two items are removed, and 0.9823 when three items are removed. When items are removed from 30% of the sample, the correlation coefficients are still strong with values of 0.9806 and 0.9646 when one and two items are removed (respectively), and 0.9500 when three items are removed. Some accuracy is lost when we remove items from 50% of the sample, where the coefficients range from 0.9691 when one item is removed, 0.9449 when two items are removed, and 0.9178 when three items are removed. Despite some information loss in the most extreme examples (i.e. 3 missing items from half of the sample), our results strongly support the argument that an AS index using IRT can handle data sets with missing items.

Scale linking simulations

The second sensitivity analysis simulates concurrent scale linking. This concept refers to the process of combining more than one set of data (and item list) into a single data set and estimating all parameters and results simultaneously (Meyer

and Zhu 2013). The ability to combine multiple data sets is particularly useful in the context of AS assessment, as there are no international standards for farm data collection. As such, the implementation of scale linking for an AS index would allow researchers to compare farm sustainability using data sets from FADN as well as national data bases (e.g. the Censuses of Agricultural for the United States and Canada), data collected for SDG indicator 2.4.1, and other large agricultural surveys.

We simulate the linking procedure using the common item nonequivalent groups design (Kolen and Brennan 2014), where groups of farms share a set of common items along with a subset of items that are unique to the group. The method is selected because the heterogeneity of information available across agricultural data sets means that most data sources will have a set of very similar (or identical) variables that can be used to build common items. Examples of common variables in agricultural data sets include land size, number of livestock units, and basic financial components such as expenditures and a monetary value of output. The advantage of nonequivalent group linking is that after the set of common items are developed, researchers can then include items specific to the needs of the region or group they are studying. For example, land tenure rights are a critical topic for farm sustainability in many developing countries (see e.g. Xu et al. 2018), but such an item would be less practical for an AS index in Germany.

The simulations are done by splitting our sample into two groups (subsamples), which are divided between former East and West Germany. The East/West division was chosen because significant structural differences persist between the regions since the reunification (see e.g. Beckmann and Hagedorn 2018), so the exercise can more closely simulate scale linking on an international level. We then systematically drop one distinct item from each subsample and compare the model estimations with the results produced by the full sample (i.e. control group). The experiment is simplified in that we (1) maintain a total of eight items for each subsample, and (2) only consider one combination of dropped items. That is, we examine, for example,

only the situation in which we drop profitability in the West subsample and solvency in Wast subsample, but we do not drop profitability in the East and solvency in the West.

We first compare latent sustainability scores between the subsamples and control group using correlation coefficients and scatter plots. Referring to Table A.10 in Appendix A.3, we interpret the correlation coefficients as a measure of information lost in the subsamples versus the control group. The results suggest that most of the scale linking tests are robust to concurrent scale linking. A majority of the correlation coefficients remain above 0.94 and reach as high as 0.9742 in the case of the test missing items 1 (profitability) and 2 (solvency). However, the tests where items 4 (economic diversity) or 7 (GHG emissions) in the west group show a higher degree of information loss. The coefficients for these tests range from 0.84 to 0.89.

Turning next to the scatter plots (Figures A.4 through A.12 in Appendix A.3), we can see how the omission of different combinations of items impacts the subsample estimates. Items with low discrimination parameters (see Figure 2.2) have a low impact on the estimations, as evidenced by scatter plot results fitting tighter to the linear fit line (as well as higher correlation coefficients). For instance, profitability and solvency (top left panel of Figure A.4) are both relatively low in terms of discrimination, and scatter plot results from their omission show points that are clustered tightly to the fit line. In contrast, economic diversity and GHG emissions both have high discrimination parameters, and the omission of these items from the subsamples results in scatter plots that show (1) points that are spread farther away from the fit line, and (2) evidence of linear bias in some of the farm estimates. Examples of the latter effect are the top right panel of Figure A.6, where the west subsample (in blue) appears to have θ patterns across three parallel fit lines, and in the bottom left panel if Figure A.10 where the omission of item 5 (provision of employment) causes two distinct parallel patterns in the East subsample.

We further look into the East/West differences by plotting the mean predicted probabilities of each NUTS 2 region (similar to Subsection 2.5.3) using the same scale

linking tests. Despite clear differences in the absolute values of the predicted probabilities (see Figures A.13 through A.18 in Appendix A.3), the results show that the relative differences between the regions are similar to the control group in every exercise (see Figure 2.5). We do find some variation among the regions, with an example being the bottom right panel of Figure A.17 where Saxony Anhalt shows a higher probability than Thuringia of achieving a "very sustainable" classification, yet Thuringia is a stronger performer in the bottom middle panel of Figure A.18. However, the broader trends, particularly with respect to the East/West differences found in the control group, remain for every simulation.

In general, we conclude that scale linking is a plausible approach to expanding the AS index internationally. However, care should be exercised in selecting and testing the items, as the results of the simulations show varying levels of sensitivity to different item combinations.

2.6 Conclusion

The sustainability of agricultural production is increasingly becoming a focus within large-scale policies and goals (e.g. CAP, SDGs, etc.), yet there is still not a universally accepted method for measuring it. In this paper, we provided proof of concept for the novel use of an IRT model for AS measurement. We used nine ordinal items and estimated the index scores with the graded response model developed by Samejima (1969). The results of the model are generally consistent with the literature, finding a positive relationship between farm size and sustainability, as well as higher sustainability performance on average for field crop and mixed production systems.

The IRT model has several advantages over existing AS methods. We have demonstrated that a reliable AS index can be developed using readily-available data that (1) can handle randomly missing data, which would be useful in the case of e.g. incomplete survey data, and (2) can be expanded to larger geographic scales containing different data sets using scale linking procedures (see Meyer and Zhu 2013).

We view the latter advantage as the most important aspect of the proposed AS index given the heterogeneity among agricultural data sets around the world. Other advantages of the proposed index include the ability to design an index with explicit assumptions of the underlying construction, and the ability to estimate item-specific easiness and discrimination parameters.

A drawback to using a unidimensional construct for the model is that it does not distinguish between the economic, environmental, and social considerations of sustainability. As such, there may be trade-offs among different aspects of sustainability (e.g. a decrease in profits from investing in emission-reducing technologies) that could be overlooked in process of making targeted policies. We therefore suggest that future research should be focused on increasing the level of detail by constructing a multidimensional IRT model with a larger bank of items to choose from. This would enable researchers to distinguish between the different aspects of sustainability and develop a better understanding of the trade-offs one might face when interpreting the results for policies. Consciously taking such trade-offs into account can lead the way towards the best policies possible.

To further exploit the benefits of IRT models in AS assessment, we also suggest that future research be directed toward expanding the model internationally. Section 2.5.4 demonstrated the potential for scale linking the index across data sets with (slightly) different sets of items, so we suggest that the next step should involve the development of a set of common items that can be found in most international agricultural data sets. After a common item set is established, researchers will have the ability to customize their own AS indices and compare the results at an international level. Such an exercise would have benefits for both domestic and international policies targeting the sustainability of agricultural systems.

3

Agricultural sustainability, non-food crops, and marginal land production: Implications for the bioeconomy

Co-author: Stephan Brosig

3.1 Introduction

In response to the growing need to reduce dependence on fossil fuels and mitigate the consequences of greenhouse gas emissions, governments have been developing strategies and initiatives focused on the promotion of the bioeconomy (e.g. White House 2012; European Commission 2012). Defined by McCormick and Kautto (2013) (p. 2589) as "an economy where the basic building blocks for materials, chemicals and energy are derived from renewable biological resources", such initiatives aim to transition towards an economic structure centered around renewable resources while improving rural economies, employment, growth, and the environment (Gawel, Pannicke and Hagemann 2019). In this context, bioeconomy models are designed to accommodate the three dimensions of sustainability: economic, environmental, and social (D'Amato et al. 2017). Germany's *National Bioeconomy Strategy* (BMBF 2020), for example, aims to meet objectives of the 2030 Agenda for Sustainable Development through the creation of a bioeconomy model focused on innovation that is within ecological boundaries and incorporates society into its development.

At the heart of the bioeconomy is the production and use of non-food crops (NFCs). Current bioeconomy strategies view the agricultural sector as the primary producer in the value chain (Efken et al. 2016) and expand the agro-industry to include a wide variety of non-food products previously made with fossil fuels (Bastos Lima 2018). Approximately 16% of all agriculturally used land in Germany is currently dedicated to the production of energy and industrial crops (about 2.67 million hectares in total) (BMBF 2015), with significant growth in the non-food sector expected in the coming decades (OECD 2009; Sheppard et al. 2011). However, unintended side effects to the bioeconomy have begun to emerge in recent years (Egenolf and Bringezu 2019), many of which are a consequence of NFC production. Such side effects include, for example, threats to biodiversity and concerns over increased chemical usage (Pfau et al. 2014).

Perhaps more importantly, the continuing increase in the demand for NFCs has led the issue of land availability (and land use competition) to be one of the primary limiting factors in the development of the bioeconomy (Pfau et al. 2014). Known as the "food versus fuel" debate (see Kretschmer, Bowyer and Buckwell 2012), this issue places the increasing demand for NFCs at the forefront of problems such as land competition (Johansson and Azar 2007) and rising food prices (Zilberman et al. 2013). As a potential solution to this issue, authors such as Fu et al. (2022) and Mitchell et al. (2016) suggest the use of marginal lands for NFC production. While these lands can ease some of pressure exerted on the food supply, the conversion of marginal lands may still have negative impacts to sustainability. Raghu et al. (2011), for example, cites impacts to biodiversity as a potential consequence to marginal land production.

Given the aforementioned sustainability concerns with NFCs and marginal land use, we ask the question: Are there observable relationships between the sustainability of farms, NFC production, and the use of marginal land to produce NFCs? In response to calls for a socio-ecological approach to sustainability assessments (see Jiang, Jacobson and Langholtz 2019; Wohlfahrt et al. 2019), we propose using item response theory (IRT) for the development of a farm level agricultural sustainability index (ASI). Originally used in educational and psychological testing, IRT models were developed to measure an unobserved latent trait using a set of observed characteristics (items) (for an overview of item response theory, see van der Linden and Hambleton 1997). The derivation of the ASI using an IRT model has the important advantages over other methods in that it (a) allows for the inclusion of economic, environmental, and social model inputs; and (b) offers more flexibility in terms of data requirements.

We generate the ASI using an IRT framework for a sample of farms in Germany. We estimate a parametric and a nonparametric regression using splines to explore the nexus between agricultural sustainability and NFC production in a flexible way, taking into account potential differences between farms producing on marginal lands and those not producing on marginal lands.

3.2 Data and methods

3.2.1 Data

Data from the Farm Accountancy Data Network (FADN) form the foundation of our sample. We use all farms in the German FADN data set for the year 2013, giving us a sample size of 8,928 farms. A caveat to using FADN is that in Germany, it is common for farms to register biogas production as separate legal entities even though the business is physically located on the farm and has technical, organizational, and economic interlinks with the farming activity. This legal separation introduces bias in the ASI scores, which could be significant depending on the success of the biogas plant relative to the rest of the farm. For example, a farm that is operating a highly profitable and productive biogas plant with many employees is likely to be very sustainable in economic terms. However, if that farm has very little activity outside of biogas production and reports low levels of profit, employment, etc., there will be a negative bias in the sustainability index because the farm will appear to be stagnant.

3.2.2 Independent variables

This subsection presents an overview of the calculations used for the independent variables in the reminder of the analysis. The raw data consists of each farm's NFC output (in €), total output (in €), marginal land classification, economic size, and farm type. FADN provides two NFC variables for energy crops and industrial crops, both of which are reported as the total value produced in the accounting year. From these, we generated two variables reflecting the proportion that each NFC type contributes to total output:

$$energy_j = \frac{E_j}{O_j} \times 100, \tag{3.1}$$

and

$$industrial_j = \frac{I_j}{O_j} \times 100, \tag{3.2}$$

where the variables for energy crop production *energy* and industrial crop production *industrial* are ratios of, respectively, energy crop output E and industrial crop output I to total output O for farm j. A small proportion of the farms (n=16, or about 0.18% of total observations in the sample) reported producing both energy and industrial crops in the same accounting year. An inspection of these farms found that in all but one case, the relative proportions of each crop type were highly unequal, so we categorize these farms as either *energy* or *industrial* based on which ratio is larger.

Marginal lands are defined as falling under three general definitions: (1) land unsuitable for food production, which are areas where food production *is not possible* and can be a result of degradation, poor soil, or bad weather; (2) ambiguous lower quality land, which is similar to definition (1) except that food production is *unsuitable* rather than impossible; and (3) economically marginal land, which simply refers to land that is not cost effective for food production (Shortall 2013). We use the farm's less favored area (LFA) classification as a proxy for marginal lands. LFAs are generally defined as land areas that are either mountainous or suffering from issues such as low productivity or a dwindling agriculturally-dependent population (Cooper et al. 2006). In the FADN data set, LFA classifications are reported under three categories: (1) not in an LFA, (2) in an LFA - other than mountains, and (3) in an LFA - mountains.

We generate marginal land classification as a dummy variable, ml, where $ml = 0$ if a farm is not located in a LFA and $ml = 1$ if a farm is located in an LFA. We do not differentiate between mountainous and non-mountainous LFA classifications, as

the third classification for farms located in the mountains makes up a very small proportion of the sample (approximately 1.8%), none of which produce NFCs. Table B.1 in Appendix B provides the descriptive statistics of the sample with respect to both marginal land classification and the type of NFC that the farm produces.

3.2.3 Items for assessing agricultural sustainability

We define nine continuous variables to be used as indicators of agricultural sustainability. Because IRT requires the use of binary or categorical variables in the model, we then transform the indicators into 4-category ordinal items for the generation of the ASI. The categories convey the extent to which the respective aspect of sustainability is achieved on the respective farm, expressed as 0="low", 1="mid", 2="high", and 3="very high". The items correspond to those used in Chapter 2, with Table A.3 in Appendix A providing the descriptions of the items and frequencies of each category in the sample.

3.2.4 Estimation of the AS index

To capture agricultural sustainability as a single comprehensive measure for each farm, we construct the AS index using a graded response model (GRM) (see Bürkner 2017; Samejima 1997*b*). The GRM is an extension of basic item response models that use binary items (e.g. yes/no, true/false, etc.) by allowing multiple ordered responses for each item. We make two assumptions when applying item response theory to agricultural sustainability, where (a) the sustainability of a farm is a continuous latent trait that cannot be directly observed or measured, and (b) this trait impacts the observed items in the index, meaning that the latter are regarded as reflective indicators of agricultural sustainability.

The general GRM operates on the conditional probability that a subject (farm) will be categorized – based on the values of nine sustainability indicators, in one of the $C = 4$ categories of agricultural sustainability from Table A.3 in Appendix A. The

probability of a farm j being classified in a particular category based on item i is

$$P(y_{ij} = c|\tau, \psi_{ij}) = F(\tau_c - \psi_{ij}) - F(\tau_{c-1} - \psi_{ij}), \tag{3.3}$$

where F denotes the CDF of the standard logistic distribution and τ is a vector of $C - 1$ unknown thresholds.[1] We specify a two-parameter IRT model for the distributional parameter ψ_{ij}:

$$\psi_{ij} = \alpha_i(\theta_j + \xi_i). \tag{3.4}$$

The farm parameter θ_j represents the latent construct of agricultural sustainability (i.e. the ASI score) for farm j. The larger the value of the ASI score is, the larger is the probability that the farm is classified as "very high" for each of the items. The item-specific parameter, ξ_i, can be interpreted as the easiness of item i. The larger ξ_i is, the higher is the probability that the farm is classified as "very high" for sustainability, regardless of the farm. Furthermore, equation 3.4 includes an item-specific discrimination parameter, α_i, that reflects that some items can better differentiate among farms with different degrees of agricultural sustainability than others.

Next, we extend our analysis and model the determinants of the latent agricultural sustainability. For that purpose, we specify two regression models for θ_j. First, we include a set of dummy variables, z_{kj}, that represent the type of the farm defined by three categories of NFC production (i.e. energy, industrial, none) and whether or not the farm produces on marginal lands:

[1] For $C = 4$ categories, the threshold parameters τ_1, τ_2, and τ_3 are freely estimated whereas τ_0 and τ_4 are set to $-\infty$ and $+\infty$, respectively.

$$\theta_j = \sum_{k=2}^{6} \zeta_k z_{kj} + \tilde{\theta}_j, \tag{3.5}$$

where ζ_k is the effect of the farm type and $\tilde{\theta}_j$ is the unexplained part of the latent agricultural sustainability, assuming that $\tilde{\theta}_j \sim N(0, \sigma^2_{\tilde{\theta}})$. The second specification for θ_j further refines the analysis by modeling the shares of energy crop in total output and industrial crop in total output as determinants of latent agricultural sustainability using flexible, non-parametric functions:

$$\theta_j = f_1(energy_j) + f_2(energy_j, ml_j) + g_1(industrial_j) + g_2(industrial_j, ml_j) + \tilde{\theta}_j. \tag{3.6}$$

More precisely, the functions f_1 and g_1 represent thin-plate splines, and the functions f_2 and g_2 are interaction splines that allow for different relationships between farms that produce on marginal lands and those that do not produce on marginal lands (for another application of splines in an IRT model framework, see, Kolczynska et al. 2020). We refer to Wood (2017) for a detailed technical explanation of splines. We fit the model in a Bayesian framework using the `brms` package (Bürkner 2017) and the probabilistic programming language Stan (Carpenter et al. 2017) in R (R Core Team 2021). We checked the convergence of the eight Markov Chain Monte Carlo (MCMC) chains for each parameter using the scale reduction factor $\hat{R}$, which is close to one ($\hat{R} < 1.05$) for all parameters (Gelman et al. 2013). Also, we inspected the MCMC trace plots and calculated the effective sample sizes.

3.3 Results and discussion

There are four parts to this section. In the first part, we validate the model by presenting the easiness and discrimination parameters. The second and third sections present how predicted probabilities of agricultural sustainability levels vary with

NFC type and marginal land classification, and with the proportion of NFC output to total output. We conclude the section with a discussion on what the results may imply for current and future bioeconomy strategies and policies.

3.3.1 Easiness and discrimination parameters

Figure B.1 in the appendix presents the posterior means along with 95% credible intervals of the easiness and discrimination parameters of each item, with the item numbers corresponding to those found in Table A.3. The results show large positive easiness parameters for items 2 (solvency) and 8 (multi-factor productivity), indicating that a relatively large proportion of farms were classified as "very sustainable" by these items. On the contrary, we find large negative easiness parameters for items 3 (wage ratio) and 5 (provision of employment), indicating that many farms were classified as very unsustainable by these items.

The two-parameter model allows us to further estimate a discrimination parameter for each item, which is assumed to be positive for identification. The discrimination parameters refer to an item's ability to differentiate between farms with different levels of agricultural sustainability. Items with large discrimination are more sensitive to small differences in agricultural sustainability than items with small discrimination. Here, we find that the items with the lowest ability to differentiate are profitability and pesticide expenditure, while the items with the highest discrimination are economic diversity, the provision of the employment, and GHG emissions.

3.3.2 Results for farm types

In this subsection, we present results from the parametric model that includes a set of indicator variables for the type of farm. The model uses six categories that are derived from combining the type of NFC production and the type of marginal land production, which are lettered as letters A-F in Tables B.1 and B.2 (Appendix B). The reference category are non-NFC producing farms that are not on marginal

lands (i.e. category A). Table B.2 presents the regression coefficients of the indicator variables of the farm type (see equation 3.5). The estimates reflect differences between farm types in the likelihood of higher agricultural sustainability levels but they can be interpreted in relative terms only, as the latent agricultural sustainability is measured on an arbitrary scale.

The results show that it is practically irrelevant for agricultural sustainability whether a farm produces on marginal land or not. That is, for any given type of NFC production, there is no evidence that farms producing on marginal lands are less sustainable than farms not producing on marginal lands. The 95% credible intervals for the two marginal lands categories overlap for given types of NFC production. However, the results document that NFC production matters for agricultural sustainability. Both types of NFC production, energy crops and industrial crops, show a clear positive association with agricultural sustainability.

3.3.3 Predicted probabilities

We next estimate the effect sizes of NFCs on agricultural sustainability by looking at predicted probabilities. Figures 3.1 and 3.2 show the predicted probabilities of a random farm achieving each sustainability category C for any given proportion of NFC output to total output. The left and right side panel show, respectively, the results for farms either on or not on marginal land. Using the baseline of no NFC production (i.e. $energy = 0$ and $industrial = 0$) as an example, we can see that the likelihood is nearly equal for a farm achieving one of the bottom three categories, with the probabilities all close to 0.3 (or 30%). Reaching the top category of "very high" is more difficult, however, with a predicted probability of about 0.12 (12%).

Looking first to the results for energy crops (Figure 3.1), we find that a farm's probability of reaching the "high" or "very high" sustainability category increases with the share of energy crop output in total output up to a share of about 60%. At that proportion, the probabilities of a farm being classified in those categories

are about 60% cumulatively (38% and 22%, respectively). Further increases of the energy crop share beyond 60% are associated with declines of sustainability. In other words, the nonlinear relationship can be described as inverted u-shaped. It is interesting to note that farms at the extreme that produce 100% energy crops have predicted probabilities that are still slightly better than the baseline. This finding suggests that producing energy crops at any proportion corresponds with higher agricultural sustainability.

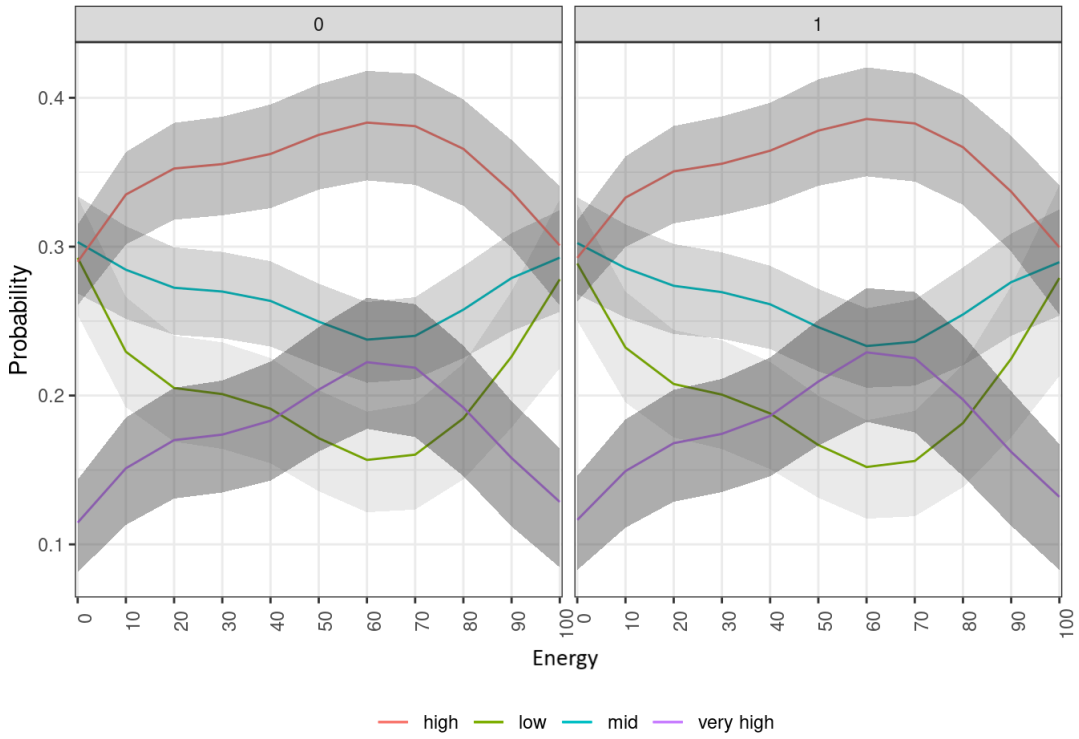

Figure 3.1: Predicted probabilities of achieving each category C by marginal land ml classification and the proportion of energy crop output to total output.

With respect to industrial crops (Figure 3.2), we again find an inverted u-shaped relationship/association between the share in total output and predicted probabilities of sustainability categories. However, there are two key differences from energy crops. First, predicted sustainability peaks much lower in the range on the x-axis: whereas predicted sustainability began to decrease at $energy = 60$ for energy crops, it begins to decrease at $industrial = 40$ for industrial crops. Perhaps as a consequence, the second key finding is that we find negative association with sustainability

when farms become specialized in industrial crops. As *industrial* exceeds approximately 90%, the categories with the highest probabilities are "low" and "mid".

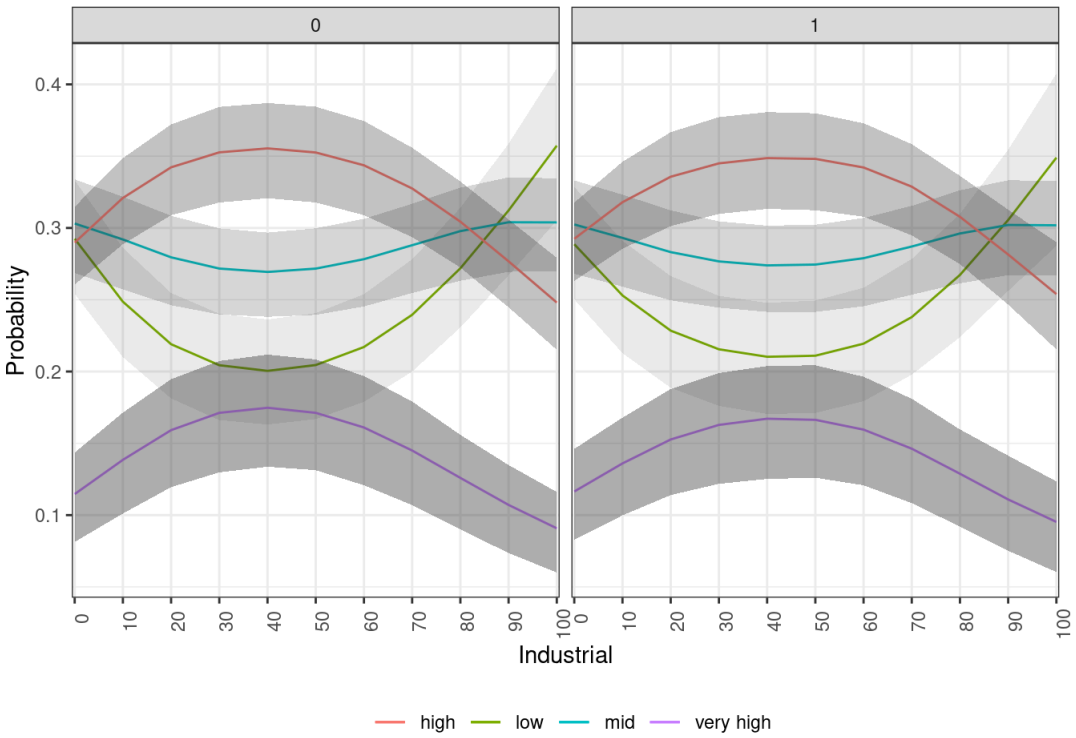

Figure 3.2: Predicted probabilities of achieving each category C by marginal land ml classification and the proportion of industrial crop output to total output.

In line with the findings in Section 3.3.2, the strong similarities between the left and right panels in Figures 3.1 and 3.2 indicates that operating on marginal lands does not make any meaningful difference with regard to predicted sustainability probabilities.

3.3.4 Implications for the bioeconomy

Germany's bioeconomy strategy (BMBF 2020) recognizes the need to improve sustainability in agriculture and suggests innovation-based solutions such as locally-specific smart farming or vertical farming. While such innovations are certainly beneficial for improving agricultural sustainability, our findings suggest that there may be simpler options as well. Given the nonlinear relationships we found between sustainability and the proportion of NFC output to total output, it may be ad-

vantageous for bioeconomy strategies to discourage farms from specializing in NFC production. Instead, farmers should produce a combination of food crops and NFCs. Policymakers and farmers should consider both the type and level of production intensity, favoring higher proportions of energy crops over industrial crops.

Regarding the production of NFCs on marginal lands, another key area of focus in Germany's national strategy (see section 4.1 in BMBF 2020), we conclude that there appears to be no significant difference in farm sustainability with respect to marginal land classification. Differences across these classifications in the estimated likelihood of higher sustainability levels are not large enough to suggest any substantial conclusions, as evidenced by virtually identical predicted probabilities in the right- and left-hand panels of figures 3.1 and 3.2. As such, we find that the production of NFCs on marginal lands is a viable option to ease pressures on the food supply without negatively impacting the sustainability of the farms.

3.4 Conclusion

The primary objectives of this paper were to (a) study the relationships between the sustainability of farms in Germany, NFC production, and producing on marginal land (as measured by LFA classification), and (b) evaluate the potential impacts of these relationships on the sustainability goals of the bioeconomy. Through the application of an agricultural sustainability index (ASI), we found that (1) there are positive relationships between farm sustainability and NFC production, (2) these relationships are nonlinear and can even lead to negative associations between NFC production and agricultural sustainability in the case of industrial crops, and (3) it appears not to make a difference to sustainability whether NFCs are produced on marginal or non-marginal lands. Thus, we conclude that the production of NFCs may be beneficial for sustainability efforts in the bioeconomy depending on the type of NFC and the intensity of production, and that production on marginal lands is unlikely to have a significant impact on these efforts.

However, these findings should serve as a starting point for future research into sustainable NFC use in the bioeconomy. Since we are currently only viewing sustainability through the lens of the farms themselves, our findings do not represent an overall assessment of the NFC value chain. Further research should continue to parse out the individual components of the value chain to gain a more comprehensive view of what is required for the bioeconomy to maintain sustainable resource production and manufacturing.

4

A comparative analysis on the farm-level sustainability of conventional versus organic production in Germany

Co-author: Stephan Brosig

4.1 Introduction

Recent decades have experienced a significant rise in organic agricultural production. Occurring as a result of rapidly increasing demand for organic foods, the area of land under organic production in Europe alone has expanded to almost 15 million hectares in 2017, up from only about 100,000 hectares in 1985 (see figure 72 in Willer, Schaack and Lernoud 2019). Today, this translates to approximately 9% of all agriculturally used land in Europe (Eurostat 2022*c*).

One of the reasons for the rise in demand is the perception that organic production is a more sustainable alternative to conventional methods. Many consumers are concerned about the impacts of conventional production (Meemken and Qaim 2018) and view organic foods as more environmentally friendly (Seufert and Ramankutty 2017). Similarly, farmers commonly cite environmental benefits in the decision to convert to organic production methods, rather than for economic gains (Cranfield, Henson and Holliday 2010). This perception is also reflected in sustainability-oriented policy development. For example, the European Commission's *Farm to Fork* strategy (European Commission 2020) claims organics to be a key sector in the coming years, and the new Common Agricultural Policy (CAP) mentions organic food as part of its eco-scheme, with at least 25% of the direct payments budget dedicated to the scheme in the 2023-2027 CAP plan (European Commission 2022*b*).

Despite this rise in organic production, the consequences of using these methods on the sustainability of the farms is not fully understood yet. As such, this research asks two questions: First, are there differences in agricultural sustainability (AS) between conventional and organic farms in Germany? Second, is there a causal effect on AS when farms transition from conventional to organic production? Identifying a causal link between organic farming and AS would have relevance to both policy making and for general audiences in making sustainable consumption and production choices.

The current scientific literature on the sustainability of organic farming identifies four key sustainability themes where conventional and organic farms often differ: productivity, economic viability, environmental soundness, and (contributions to) social well-being (Reganold and Wachter 2016). While there is a breadth of literature discussing components of each theme individually, there is little attention given to addressing farm level AS differences between conventional and organic farms. Such an approach would have the distinct advantage of accounting for trade-offs between the individual components.

Our research contributes to the existing literature by conducting a two-step estimation process to test differences in AS between conventional and organic production methods at the farm level. The first step modifies the AS index proposed in Chapter 2 as a time-series model estimating a single sustainability score for each farm over a 10-year observation period. In the second step, we use descriptive statistics and a difference-in-differences (DiD) model to identify changes in AS before and after the transition to organic production.

Our analysis faces two important challenges, however, the first of which involves the use and interpretation of the statistics derived from the AS index. Our two-step approach requires that the index scores are treated as observed data for the the descriptive and causal analyses. These scores, however, are predicted within a Bayesian framework and contain a certain degree of uncertainty. This uncertainty in the measurement of AS is ignored in the second step of the analysis.

The second challenge to the analysis is selection bias. There are two primary mechanisms to explain differences in AS with respect to conventional versus organic production: self-selection, and learning by doing (for a similar argument in the context of firm productivity, see De Loecker 2007). The former refers to the idea that farmers are not randomly selected for conventional or organic production. Rather, the decision to produce organically is based on the farmers' mindset and includes a variety of factors such as personal values and beliefs. As such, farmers who choose to produce organically may already be more concerned about sustainability and

taking steps to improve their operations, while those who decide to continue using conventional methods may naturally make different choices[1]. In contrast, learning by doing suggests that once the farmers make the transition to organic, they gain new knowledge and skills that enable them to become more sustainable. It is this latter mechanism that we are attempting to capture in the analysis. We take account of selection bias by (1) creating subgroups to compare differences between farms that remain conventional during the observation period versus farms that have not yet made the conversion to organic, (2) using a DiD model with propensity score matching, and (3) running multiple iterations of the DiD model with different control groups and sample sizes.

By using data from the Farm Accountancy Data Network (FADN) and applying our descriptive and DiD regression analyses, our research reveals several key findings about the sustainability of organic versus conventional farms. The descriptive analysis finds that farms observed to be organic throughout the observation period are more sustainable on average than farms that remain conventional. This finding holds regardless of the size or type of the farm. When looking specifically at farms that make the transition to organic during the observation period, we find evidence of a non-linear effect to AS from the conversion to AS. During the conversion process there is an decrease in AS followed by a slow long-term increase after farms complete the conversion to organic. While the findings suggest a causal relationship between AS and organic farming, the size of the confidence intervals prevent a definitive conclusion.

The remainder of the paper is organized as follows: Section 2 briefly summarizes key differences in conventional versus organic agricultural production, Section 3 outlines the two-step estimation process, the results are presented in Section 4, and Section 5 concludes.

[1]For example, a survey study by Läpple (2013) concluded that conventional farmers are less environmentally aware than farmers who are (or have been) producing organically.

4.2 Literature review

In this section, we review the literature to discuss the structural and operational differences between conventional and organic farms that may contribute to differences in AS. The review is structured around the four sustainability themes described by Reganold and Wachter (2016).

The productivity theme generally refers to differences in the intensity of production on the farm. By design, organic farming is less intensive than conventional methods to promote healthy soils without the need for chemical additives. As such, this strand of literature generally expresses productivity in terms of yield (by weight) over a given land area and finds that organic farms are less productive than conventional (see e.g. De Ponti, Rijk and Van Ittersum 2012; Seufert, Ramankutty and Foley 2012). This method, however, is limited in terms of explanatory power with regard to AS. A better alternative employed by others is a multi-factor productivity estimation (also known as total factor productivity) to incorporate measures of resource use efficiency. Baležentis (2015), for instance, uses an input/output ratio with measurements for land, labor, and assets relative to monetary outputs of crops and livestock. Using this approach, the authors find no notable differences in productivity between conventional and organic farming. When looking at a disaggregated view of productivity, Coomes et al. (2019) emphasizes that many of the new technologies embraced under organic farming, such as no-tilling or the use of precision inputs, are beneficial for both productivity and sustainability.

The strand of literature on economic viability points out that while yields may be lower, organic farms are generally more profitable. Despite a finding that conventional farmers are typically more profit-oriented (Läpple 2013), a meta-analysis by Crowder and Reganold (2015) concluded that the price premiums found in organic products resulted in organic farms being up to 35% more profitable than conventional. The study also noted a benefit/cost ratio up to 24% higher in organic farms. What is more, organic practices such as maintaining soil health and water conser-

vation can increase profits over time (Kleemann and Abdulai 2013). However, an important consideration for economic viability is debt. Canavari et al. (2007) and Krause and Machek (2018), for example, both find that organic farms typically have higher debt levels compared to conventional practices. This can have an impact on the economic sustainability of a farm, as it can signal the ability to continue operations in the event of financial difficulties (Whitehead et al. 2016; Zwilling and Raab 2019).

Two important topics dealing with the environmental soundness of farms refer to chemical use and greenhouse gas (GHG) emissions. Reducing pesticides is considered as a high priority for AS (Lechenet et al. 2014*b*), and the use of these chemicals is already either banned or heavily restricted in organic farming (Niggli 2015). The regulations have had a clear impact, as Geissen et al. (2021) reported 70-90% less pesticide residues in organic soils. Pesticide reductions can have economic benefits as well, as a study by Klonsky (2012) concludes that spending on pesticides (measured in USD/acre) is considerable lower in organic production systems.

The distinction between conventional and organic farming with respect to emissions is less clear. Organic practices typically favor modes of production with potential for lower emissions. As examples, Küstermann, Kainz and Hülsbergen (2008) finds higher potential for carbon sequestration as a result of growing perennial legumes and grasses, and West and Marland (2002) finds significantly lower emissions in practicing zero tillage, a practice more common in organic production. However, Venkat (2012) finds that while farms that recently converted to organic do in fact show significant reductions in emissions (almost 18% compared to conventional), steady-state organic (i.e. systems that have not changed in a long time) show higher levels of emissions. Further, a meta-analysis by Mondelaers, Aertsens and Van Huylenbroeck (2009) concludes that despite lower emissions by organic farms over a given land area, lower yields per unit of land result in a reduction or complete elimination of the effect.

A key component to the social well-being of organic farming is the effect it has on farm employment. A large-scale analysis by (Finley et al. 2018) found that the intensity of employment on organic farms (measured as a ratio of hired labor per acre) is considerably higher than that of conventional farms. Further, the authors found that the employment is typically more stable, as the percentage of workers being employed more than 150 work days per year is much higher on organic farms. This has a noticeable overall impact on the rural labor market, as Mon and Holland (2006) finds improvements to regional total employment near organic farms as a result of higher labor requirements.

4.3 Data and variable construction

In this section, we provide an overview of the data and variable calculations we use in all analyses. After introducing the data in Subsection 4.3.1, Subsection 4.3.2 then describes the control variables used in all regressions. Subsection 4.3.3 then describes the derivation of the organic groupings we use to differentiate between conventional farms, farms switching to organic, and organic farms.

4.3.1 Data

We use an unbalanced farm-level panel data set from the Farm Accountancy Data Network (FADN) with a 10-year observation period of 2004-2013. Because we are interested in analyzing sustainability over multiple time periods, we drop all farms with fewer than five[2] observations in the sample. This minimum threshold was selected because farms are required to undergo a transition period of up to three years[3] prior to receiving official organic certification, so a five-year minimum should

[2]Sensitivity analyses were conducted by running multiple iterations of the model while varying the minimum number of observations per farm. Results of the analyses are presented in Subsection 4.5.3.

[3]The duration of the conversion varies by farm type, with minimum conversion periods of one year for pig and poultry, two years for land ruminant grazing annual crops, and three years for fruit orchards European Commission (2022*a*)

allow us to observe at least two time periods outside of the transition period to aid in identifying the point in which the farm begins the transition.

4.3.2 Descriptive variables

Given the large differences in AS across farm sizes and types (see Chapter 2), we include these characteristics as variables in a descriptive analysis (Subsection 4.5.1) to provide a more nuanced understanding on the differences between conventional and organic farms. Farm size is determined by its standard output, which is defined by Eurostat (2023*b*) as "the average monetary value of the agricultural output at farm-gate price, in euro per hectare or per head of livestock". Farm type defines the specialization of each farm, and we use a set of broadly defined classifications[4] such as fieldcrops, milk farms, or vineyards. Both variables are categorical and are supplied by FADN, the definitions and descriptions of which can be found in European Commission (2000).

4.3.3 Organic status and grouping variable

The main explanatory variable for our analysis is the organic status of the farm. The variable is self-reported yearly by the farmers in the FADN data set following the EU's official organic certification (Regulation (EEC) Number 2092/1991: Council of European Union 1991). The German FADN data contains three categories of organic classification: fully conventional (which we denote as category C), partially organic or transitioning to organic (P/T), and fully organic (O).

Next, we categorize the farms into groups according to their reported organic classifications over the duration they are observed in the data set. We create five groups with labels "starters" (S), "always conventional" (AC), "always organic" (AO), "quitters" (Q), and "disqualified" (DQ). Table C.1 in the appendix provides a set of nine example farms with five generic time periods (denoted as year Y) to show how the farms are assigned to each group.

[4]We use the TF8 classifications supplied by FADN.

The benchmark we set for farms in the starters group is that they must have entered the observation period as fully conventional (C) and converted to fully organic (O) with *at least* one year of transition (P/T). Example farm 1 presents a complete three-year transition period, where the farm is first registered as C in year 1, reports the three-year mandatory period as P/T in years 2-4, then reports as O in year 5. We also allow farms to the starters group that report periods of P/T after the initial conversion to O is complete. Farm 2 provides an example of this, where it reports with a shorter conversion period in year 3 prior to moving to full organic in year 4, but later reports a period of P/T in year 5. We also create subcategories for farms in the starters group that differentiate between pre-treatment[5] (labeled as pre-S) and post-treatment (post-S). In Table C.1, example farm 2 is pre-S in years 1-2, and post-S in years 3-5. The subcategories are used for checking the representativeness of our sample and assessing the potential for selection bias in the model.

Farms categorized as always conventional (AC) or always organic (AO) are those that do not report in all three organic classifications during the observation period. The key criteria are that AC farms can never report as O within the observation period, and AO farms can never report as C. However, we do allow for periods of P/T in both groups. AC farms are shown by example farm 3, which remains fully conventional for all observed years; and by example farm 4, which is fully conventional but reports some years as P/T. Example farms 5 and 6 show the same patterns for always organic farms, with the exception that they primarily report as fully organic rather than conventional.

The fourth group of "quitters" (Q) are those that exit organic production during the observation period. Farm 7 is an example of a fully organic farm that reverted back to conventional, and farm 8 shows a starter that entered the organic market and exited shortly after. We drop these observations because (1) the sample size for these types is comparatively quite small so the results may be less reliable than the other groups, and (2) keeping them in the analysis would violate the assumption in

[5]See Subsection 4.4.1 for the definition of the treatment.

the model of treatment irreversability. The DiD model assumes that once a subject (farm) is treated (transitions to organic), it cannot "forget" about the treatment experience (see Callaway and Sant'Anna 2021).

Finally, the fifth group labeled "disqualified" (DQ) are farms that do not report a partial/transition period prior to switching to fully organic from fully conventional. Since EU regulations require a transition period prior to being certified as an organic farm, it is unclear what the abrupt transition represents. We regard this as an accounting error and exclude the farms without a transition period.

Table C.2 in the appendix presents the frequencies of the farms in the sample relative to the organic groupings, and the farm size and type covariates. The final sample size over the entire observation period is 64,102 observations from 8,170 individual farms. Of these, a vast majority of the farms (>95%) are in the always conventional group, 4% are in the always organic group, and less than 1% are in the starters group.

To assess the representativeness of our sample, we rely on the statistics provided by Eurostat (2023*a*) measuring organic production as a percentage of total utilized agricultural area (UAA) in Germany. We replicate this statistic within our sample by calculating annual percentages representing the share of UAA under organic production relative to total UAA. The estimation includes AO and post-S farms, with Figure C.1 showing the comparison between our estimations and those from Eurostat. The results suggest that organic farms are slightly underrepresented in our sample. In 2004, approximately 4.5% of total UAA in Germany was under organic production, while only about 3.8% of the UAA in our sample is produced organically. The largest difference between these values is in 2012, where the percentages are 5.8% in the Eurostat figures and 3.5% in our sample. The cause of this under-representation is likely a consequence of the FADN data set, as farms with less than 25,000€ per year of standard output are not reported in the data. As such, we do not make inferences to the total population of farms in Germany when reporting the results.

4.3.4 AS index

We measure the sustainability of each farm in the sample using the AS index developed in Chapter 2. The advantage to using the AS index is that it can build on the literature discussed in Section 4.2 by estimating a single farm level sustainability score that allows us to take into account trade-offs between different aspects of farm sustainability. Examples of potential trade-offs in the conversion from conventional to organic include reduced yields in exchange for higher prices, or increased labor requirements in exchange for savings on inputs (e.g. chemicals).

The index is estimated using the nine sustainability items described in Table C.3 in the appendix. The items loosely follow the four sustainability themes identified by Reganold and Wachter (2016) and cover all of the topics addressed in Section 4.2. Economic viability is accounted for using items for the profitability, solvency, and economic diversity of the farm. The items for farms' expenditure on pesticides, estimated GHG emissions, and estimated land ecosystem quality account for the environmental soundness theme. With respect to the social well-being theme, we view a narrow scope of the theme by inspecting the farms' wage ratio (i.e. average wages paid on the farm relative to regional median wages) and the intensity of employment generated from production (i.e. expenditures on wages relative to total output). Finally, the productivity theme is estimated using a multi-factor measure similar to that of Baležentis (2015), which estimates the farms' total value added relative to factor inputs for land, labor, and capital.

In contrast to the cross-sectional AS index estimated in Chapters 2 and 3, our analysis requires the use of time-series data to estimate changes in AS over time as farms convert from conventional to organic production methods. Section C.2 in the appendix provides a model statement for the development of a time-series AS index spanning the same 10-year observation period specified in Subsection 4.3.1. A key aspect of the time-series version of the index is that it incorporates a set of time dummies shown in equation C.3 that allow for the index to vary over time.

The variable from the AS index estimations that we use for the remainder of the analysis is the latent trait of sustainability for each farm, θ (see equation C.4). To help with the interpretation of the magnitude of the results presented in Section 4.5, we standardize θ to have a mean of zero and standard deviation of one, which we denote as θ'.

4.4 Empirical strategy

In this section, we discuss the approach for estimating a difference-in-differences (DiD) regression model for our analysis. Subsection 4.4.1 outlines the identification strategy we use for the model, then Subsection 4.4.2 discusses the model assumptions. Subsection 4.4.3 then provides an overview of a series of robustness tests used to validate the model results.

4.4.1 Identification strategy

To address the question of a causal relationship between AS and the conversion from conventional to organic production, we rely on a DiD estimation strategy. Examples of other DiD models applied to agriculture include an analysis of employment effects in Eastern Germany as a result of the Common Agricultural Policy (CAP) (Petrick and Zier 2011), and impacts of immigration enforcement on the agricultural sector in the US (Kostandini, Mykerezi and Escalante 2014). The general idea is that a DiD model compares a change over time between a treated group and a control group before and after a treatment occurs. In the context of our analysis, we define the treatment as the conversion from conventional to organic production, the treated group as the farms that are observed to convert to organic during the observation period (group S), and the control group as the group of farms that are never treated (group AC)[6].

[6]Alternatively, in a robustness test in Subsection 4.4.3 we substitute the AC control group with the "not yet treated" (pre-S) farms.

A significant challenge to our analysis comes from the fact that the treatment can take place at any time during the observation period (i.e. staggered treatment). DiD models are generally designed to measure a treatment occurring over a single time period (e.g. from the introduction of a new policy), and the effects are estimated using a two-way fixed effects (TWFE) approach (e.g. Hackenberger et al. 2021). The TWFE model takes into account both time and group fixed effects, where the group refers to the units of analysis (i.e. farms). The group fixed effects would then control for any group-level confounding factors that might affect the outcome variable, such as differences in baseline characteristics, unobserved heterogeneity, or selection bias. Time fixed effects control for any time-varying confounding factors that might affect the outcome variable, such as changes in the economy, technology, or weather (for an overview of TWFE models, see de Chaisemartin and D'Haultfoeuille 2022).

While the TWFE model produces reliable estimations in the case of a treatment occurring in a single time period, the use of TWFE in models with staggered treatment times has come under scrutiny lately. Authors such as Goodman-Bacon (2021) show that TWFE models introduce bias into staggered treatment estimates as a result of "forbidden" comparisons. This occurs when the TWFE makes comparisons between units that are both already treated in addition to comparisons between treated and "not yet treated" (i.e. "clean" comparisons). In this context, TWFE will suffer from heterogeneity bias if the treatment effect is not homogeneous. The consequence of heterogeneity bias is that the direction of the estimated coefficient may be negatively weighted, meaning that the TWFE coefficients may have an opposite sign compared to the individual treatment effects (Roth et al. 2022).

For that reason, we use the DiD model developed by Callaway and Sant'Anna (2021) in our analysis (henceforth, CSDiD). The model overcomes the problem of "forbidden" comparisons by considering a group-time average treatment effect (ATT) that differentiates between groups of farms that are not yet treated versus those that already treated. In simple terms, the group-time ATT means that all ATTs are estimated relative to the year a particular group of farms converted to organic during

the observation period. This approach uses a new time variable t to represent the number of time periods before and after the conversion to organic. For example, the ATT value reported in time period $t = 4$ represents the treatment effect in the year 2008 for farms treated in 2004, the treatment effect in the year 2011 for farms treated in 2007, and so on.

However, a challenge to our approach is in properly identifying the treatment time $t = 0$, i.e. the exact point at which a farm converts to organic. The conversion to organic essentially has two clearly defined time periods: the first time it reports as P/T after being fully conventional (C), and the first time it reports as fully organic (O) after one to three years (minimum) of P/T. We chose to define $t = 0$ as the first year a farm reports as P/T under the rationale that it should be the first demonstrable actions taken by the farmer to produce organically. In contrast, if we had selected the first time period that a farm reports as fully organic, the prior P/T time periods may distort the ATT estimates because the assumption of no anticipation effects could be violated (see Subsection 4.4.2). Referring to the last column in Table C.1 in the appendix as an example, $t = 0$ occurs in year 2 for example farm 1, and in year 3 for example farm 2. Note that the time variable is not required for the AC group, and the AO group is excluded from the CSDiD model.

4.4.2 Model assumptions

In this subsection, we outline three key assumptions for the CSDiD estimations. First, one of the basic ideas of any DiD model is the assumption that in the absence of a treatment, both the treatment and the control group would follow a similar trend over time (parallel trends assumption). There are several methods to test this assumption, such as the introduction of an interaction term (see, e.g. Tang et al. 2022). However, parallel trends tests have been shown to have low power and can introduce additional bias into the model estimates (Roth et al. 2022). The CSDiD model controls for this assumption in the group-time comparisons.

The second assumption for the model is that there are no anticipation effects. If the assumption is violated, it would mean that farmers begin making changes to their operations that affect AS prior to the actual transition into organic production. While the CSDiD model does allow to control for this assumption, we estimate the model assuming that the assumption holds and inspect the pre-treatment time periods in the CSDiD model estimates for trends in the ATT that might suggest anticipation effects. This inspection is discussed in detail in Subsection 4.5.2.

Finally, as highlighted in Section 4.1, selection bias is a significant challenge to our study. If present in the model, it would violate the assumption of random sampling. The CSDiD model accounts for selection bias by incorporating a generalized propensity score that is uniformly bounded away from one to rule out irregular identification. Propensity scores function as a counterfactual by matching treated subjects (farms that switch to organic) with observationally similar untreated subjects (farms that do not switch) (see Heinrich et al. 2010), thereby correcting for the aforementioned issue with TWFE models in staggered treatment applications. As an additional step, we inspect patterns between the AC and pre-S groups in our descriptive analysis to look for evidence of selection bias prior to the CSDiD analysis (see Section 4.5.1).

4.4.3 Robustness tests

We conduct two robustness tests to verify the main CSDiD model estimations. For the first test, we run a second CSDiD regression using the "not yet treated" farms, i.e. the pre-S group. The rationale behind running a second regression is so that we can account for the strengths in both estimations. On the one hand, the large sample size of the AC group (see Table C.2) is favorable for use as the control group; however, a requirement of using a "never treated" group as the control is that the farms should be similar to the "eventually treated" group (i.e. pre-S). Because of the difficulty in determining how similar the AC and pre-S groups are, we add the

second regression so that all farms in the regression are either "not yet treated" (pre-S) or have been treated (post-S).

In the second test, we run 12 iterations of the CSDiD model using both the AC and pre-S control groups (i.e. six iterations of each control group). In each iteration we incrementally increase the minimum number of observations per farm in the sample from five up to ten observations (i.e. balanced panel data set). The purpose of the test is to assess whether any change in the minimum number of observations would have an effect on the magnitude or direction of the ATTs, as well as how the confidence intervals might be affected.

4.5 Results

In this section, we provide a descriptive analysis of AS among the organic groupings and discuss the findings of the CSDiD model estimations. In Subsection 4.5.1, we present an overview of the descriptive analysis using the variables introduced in Subsection 4.3.2. Subsection 4.5.2 then discusses the CSDiD model results produced using the AC control group and a minimum of five observations per farm. We conclude with Subsection 4.5.3 by discussing findings from the robustness tests outlined in Subsection 4.4.3.

4.5.1 Descriptive analysis

In this section, we present a descriptive analysis on the differences in θ' among the farm groups. Group differences are tested using a series of ANOVA and pairwise comparison estimations that focus on the contrast (i.e. the relative differences in the means of θ') between the groups.

For this subsection, we reduce the data set to only include data from 2004. We select a single year for the analysis to avoid distorting the tests from repeated observations of the same farm, and the first year of the observation period is chosen specifically

so that all farms in the starters group are categorized as pre-S. We are particularly interested in the pre-S farms because the rationale is that since both the AC and pre-S farms are fully conventional in the time period they are being observed, any differences in θ' between the groups may be a sign of selection bias. We first run the ANOVA and pairwise tests for the entire 2004 sample, then subdivide the sample in groups for the eight farm types and nine economic size classes provided in the data set.

We summarize two important findings about the full 2004 sample. First, the kernel density plot shown in Figure 4.1 suggests that selection bias is unlikely in the sample. This is evidenced by the similarity in the density curves between the AC and pre-S groups.

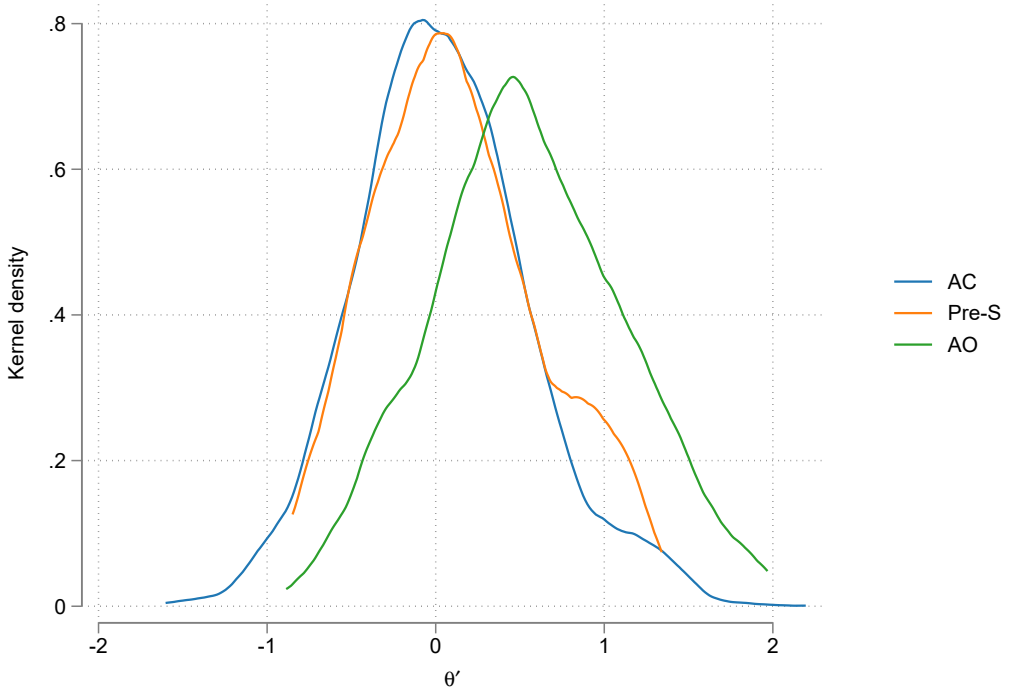

Figure 4.1: Kernel densities of θ' for the AC, pre-S, and AO groups in 2004

The second finding from Figure 4.1 is that AO farms are clearly more sustainable on average than the AC or pre-S farms. As shown in Table C.5 in the appendix, the mean θ' for AO farms is 0.54 while the pre-S and and AC means are 0.12 and 0.05, respectively. When the groups are tested using the pairwise mean comparison

in Table C.7, we find no meaningful differences between the pre-S and AC groups (contrast = 0.08), while the contrast values are 0.50 between the AO and AC groups and 0.42 between the AO and pre-S groups.

Next, we subdivide the sample using the farm size and type descriptive variables discussed in Subsection 4.3.2. We conduct a separate set of ANOVA and pairwise comparison tests for each farm size and type classification. Looking first at the farm type classifications, Figure 4.2 presents box plots of the θ' distributions and Tables C.8 through C.31 present the results of the ANOVA and pairwise comparison tests.

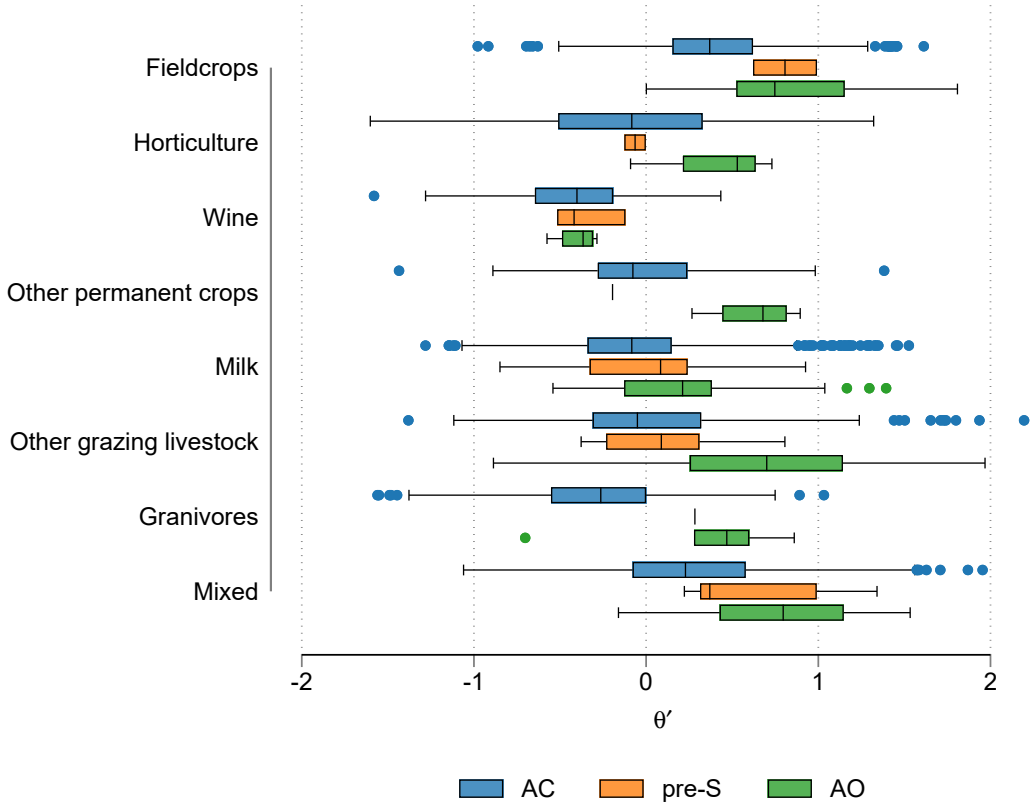

Figure 4.2: θ' distributions by TF8 farm type, 2004

We find that the mean θ' values for AO farms are significantly higher than for the AC farms in most of the farm type classifications. The largest contrasts between AO and AC farms are found in grazing livestock (0.70, Table C.25) and permanent crop farms (0.67, C.19). These differences do not hold for all farm types, however, as there appears to be no significant differences in θ' between the AO and AC

groups in horticulture farms and vineyards (Tables C.13 and C.15, respectively). With respect to evidence for selection bias, we do not find statistically significant differences between the pre-S and AC farms in any of the farm types. The finding suggests that selection bias is not an issue.

With respect to the economic size classes, Figure 4.3 presents box plots of the distributions for each organic grouping, and Tables C.32 through C.58 show the results of the ANOVA and pairwise comparison tests. Similar to the farm type comparisons, we find no meaningful differences in θ' between the AC and pre-S groups in any of the size classes. This again suggests a lack of selection bias.

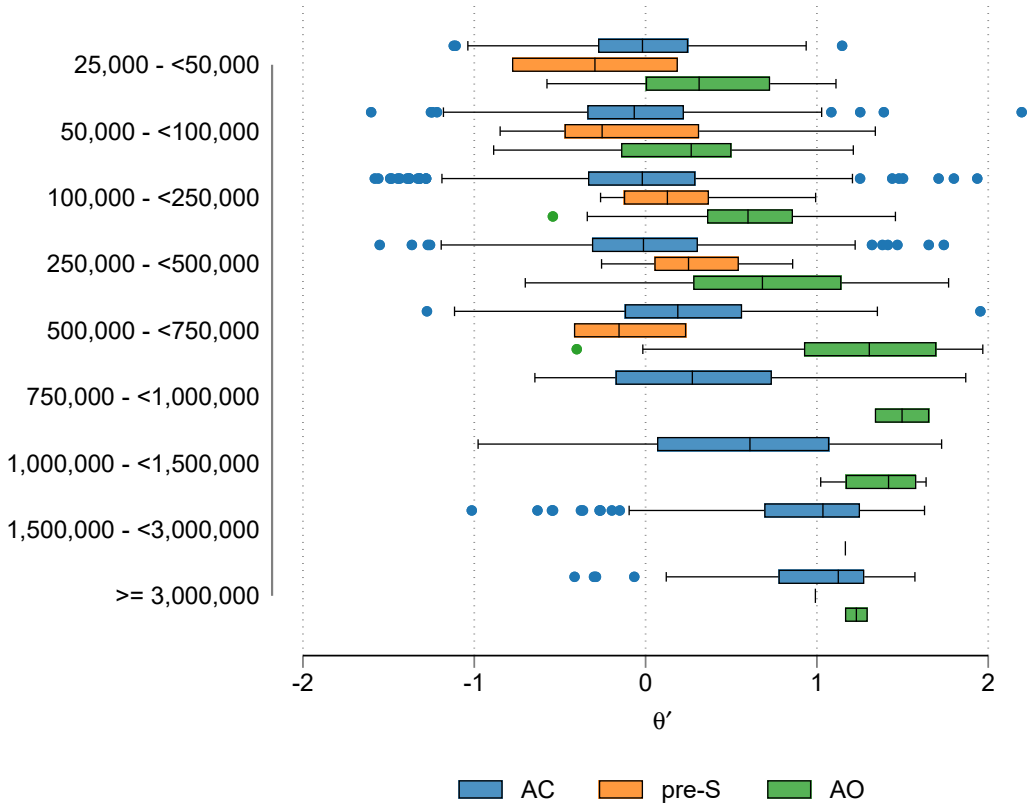

Figure 4.3: θ' distributions by economic size class, 2004

We also find that the differences in θ' between the AO and AC groups vary across the economic size classes. The contrast between the two groups steadily increases relative to farm size up to 1,000,000€ of standard output per year. As shown in Table C.34, the contrast between AO and AC farms is 0.37 for farms in the

25,000 -<50,000€ and the largest contrast between these groups is 1.17 for farms in the 750,000 -<1,000,000€ range (Table C.49). This gap diminishes for farms exceeding 1,000,000€ of standard output (Tables C.52, C.55, and C.58), though the small sample sizes introduce a high level of uncertainty in the findings for these size classes.

4.5.2 CSDiD model analysis

The following subsection presents the results of the CSDiD model specified in Subsection 4.4. Results of the model estimations are provided by Figure 4.4 showing the ATT on θ' based on length of exposure with 95% confidence intervals. The length of exposure is defined as the duration (measured in time periods t, where $t = 1$ year) before and after a farm begins the conversion to organic production. Table C.59 in the appendix provides numerical results of the estimations shown in Figure 4.4.

We discuss two main findings from the results. First, an examination of the pre-treatment results (i.e. $t < 0$) is inconclusive regarding the potential for anticipation effects in the model. The volatility in the pre-treatment trend is suggestive of anticipation effects, as we would expect AS to become unstable if farmers begin making structural changes to their operations. Literature discussing the steps taken by farmers in the years prior to treatment is sparse; however, it seems likely that this volatility could be a result of factors such as investments into new equipment, preliminary changes to cropping systems, etc. Despite this volatility, it is unclear how large the anticipation effects, if present, have on the outcome of the model. The overall pre-treatment mean ATT is 0.018 with 95% confidence intervals of [-0.122, 0.157] (Table C.59), which does not suggest an overall positive or negative trend away from zero.

Second, the post-treatment trend suggests a non-linear causal effect to AS from the transition to organic. We find a slight increase in the ATT for the first two post-treatment time periods, a decrease in time period $t = 3$, and a relatively steady

increase for the remaining time periods. The findings suggest that there may be a long-term positive effect on AS from the organic transition, but that the transition period when farms are still in the process of conversion (up to three years depending on the type of farm) remains volatile. The European commission offers financial support for transitioning farms in recognition of potential hardships during this period (see European Commission 2022*a*), which may contribute to the downswing in the ATT during the conversion. However, overall the large confidence intervals preclude definitive conclusions on the effect of the organic conversion to AS. Only in last time period does the 95% confidence interval begin to suggest a significant effect in the lower boundary, but we do not regard this finding as conclusive evidence of a causal relationship.

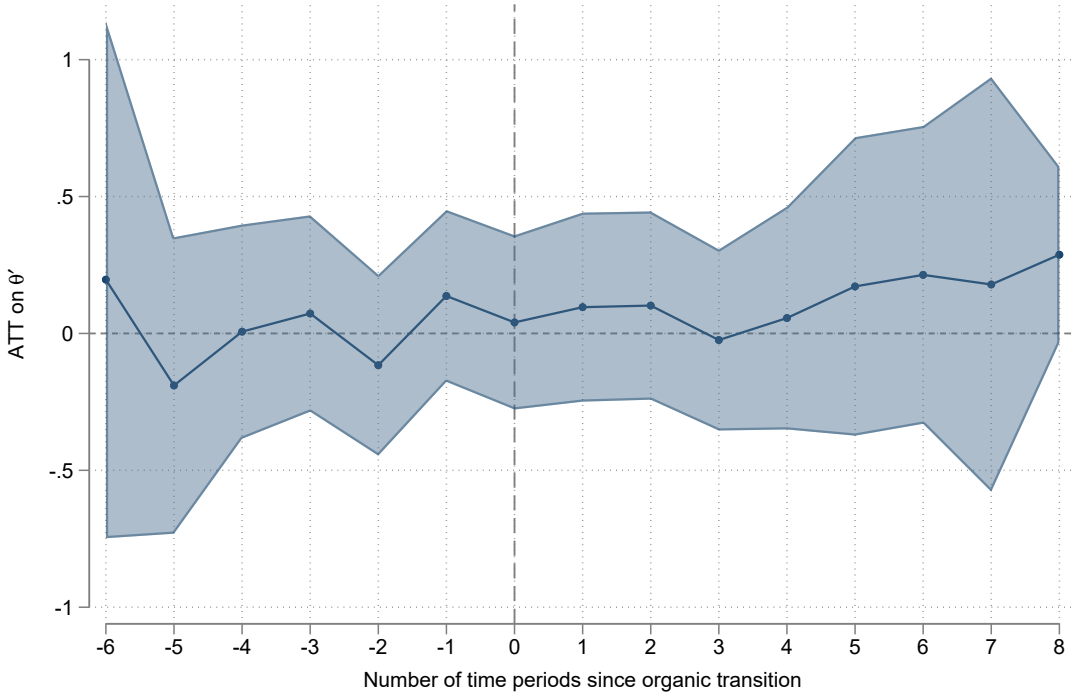

Figure 4.4: ATT on θ' with AC control group and a minimum of five observations per farm

4.5.3 Robustness test results

In this subsection, we present the results of the robustness tests discussed in Subsection 4.4.3. We first inspect the test that substitutes the AC control group with the "not yet treated" i.e. pre-S group. As shown in Figure C.3 and Table C.60, there are virtually no differences in the CSDiD estimates between the two control groups when a minimum of five observations per farm are included in the sample.

Similarly, we find only minor differences in the tests where we vary the minimum number of observations for each farm in the sample. As shown in Figures C.4 and C.5 in the appendix, there are slight variations in the ATT depending on the minimum number of observations. With respect to the pre-treatment trends, we notice less volatility in the time period $t = -5$ with the tests that have minimums of 6, 7, and 8 observations. Overall, we find that the confidence intervals widen as the minimum number of observations increase. This is expected since a higher restriction in the minimum observations results in a decreased sample size. Interestingly, the test with a minimum of 10 observations per farm (i.e. balanced panel) shows a smoother post-treatment trend with a steady increase in the ATT. The reason for this finding is unclear. Finally, similar to the previous test, we find that there are no noticeable differences between the tests substituting the AC control group with the pre-S farms.

4.6 Conclusion

The goal of this paper was to identify differences in farm sustainability between conventional and organic production methods. We used a 10-year FADN data set in Germany and separated farms into groups based on whether the farm is always conventional during the observation period, converts to organic production while it is being observed, or is fully organic during this period. We then use a time-series farm level AS index to compare relative sustainability of the groups. The comparison is first done descriptively using ANOVA and pairwise comparison tests

for a single year of the data, and for each farm size and type classification. We then used a difference-in-differences model developed by Callaway and Sant'Anna (2021) that allows for staggered treatment times (CSDiD) to look for a potential causal relationship between farm sustainability and the transition from conventional to organic farming.

We highlight the following findings from the analyses. First, we found that the farms in the always organic (AO) group are more sustainable on average than farms that are always conventional (AC) regardless of farm size or type. The descriptive analysis also shows that the pre-S farms are nearly identical to the AC farms in almost all farm types and sizes, suggesting that selection bias is not an issue in our sample. Second, the results of the CSDiD model show evidence suggestive of a non-linear effect to AS from the conversion to organic, where the years that the farm is transitioning are relatively volatile, but a steady long-term effect emerges in subsequent years. However, the size of the confidence intervals preclude a definitive conclusion of causality, and an inspection of the pre-treatment trends suggests the possibility of anticipation effects.

Based on the results, we hypothesize that there might be effects to AS from the transition to organic production that occur over a longer duration than what is captured in our analysis. This hypothesis is based on the findings of (1) a lack of evidence for selection bias, (2) an upward trend in the post-treatment ATT values after the conversion period is complete (Figure 4.4), and (3) clear differences in farm sustainability when comparing the AC and AO groups (Figure 4.1). However, we recognize that an alternative hypothesis to explain the differences between the AC and AO groups is that there may be time-sensitive components of the organic transition that are unobserved in our analysis. This hypothesis would suggest that farms making the transition to organic during the observation period of 2004-2013 could somehow be systematically different from the farms that made the transition in decades prior. An example of potential differences between these farms is discussed by Padel (2001), who uses an innovation diffusion perspective to highlight differences

in early versus late adopters of organic methods.

Considering both of the aforementioned hypotheses, we regard the observation period as the most critical limitation to the analysis. We suggest that future research should replicate the study with a longer set of FADN data. If the duration of the study were expanded to include all available years in the FADN data set, we could test both hypotheses and make a more definitive conclusion on potential causal effects to AS from the transition to organic.

5

Conclusion

5.1 Summary of results

The goal of the dissertation was to present the design and application of a new novel approach to estimating the sustainability of farms using IRT. Overall, the findings throughout the dissertation suggest that the proposed AS measurement method is a viable substitution for existing methods. The research further demonstrated the usefulness of the index in addressing important issues related to the sustainability of farms under different methods, agricultural products, and scales of production.

Chapter 2 provided proof of concept for using IRT by (1) generating nine sustainability items on a 4-category ordinal scale and estimating an AS index using the graded response model developed by Samejima (1969; 1997*b*), (2) comparing the index estimations with known associations between AS and farm type and size, and (3) conducting a series of simulations to test the ability for the index to handle missing data and accommodate scale linking procedures.

Testing of the proposed index found that the patterns in the results with respect to sustainability differences across farm types and sizes were generally consistent with the literature. Further testing of the index found that (with proper care) the index is generally robust against missing data, and can allow for substitutions in

some of the items by using scale linking. The latter finding is particularly useful for expanding the model internationally in the future.

Chapter 3 then demonstrated an application of the index by using a descriptive analysis of the AS scores to identify correlations between farm sustainability and non-food crop (NFC) production, as well as producing on marginal lands. The topic has relevance for current and future bioeconomy strategies and policies. The analysis was carried out by estimating the proportion of NFC output to total output for each farm, then using the ratios as explanatory variables along with control variables for farm size and type to look for changes in AS relative to NFC output ratios.

The index found that the most sustainable farms on average were those that produced a combination of NFCs and food crops. Depending on the type of NFC produced (either energy crops or industrial crops), farms with the highest predicted probabilities of being very sustainable were those that produced between 40% and 60% NFC output to total output. There were no noticeable differences if the farm was producing on marginal land or not. As such, Chapter 3 concluded that NFC production may be a positive effort for sustainability in the context of the bioeconomy, and that producing on marginal lands may aid in meeting the demand for NFCs without negative consequences to AS.

Finally, Chapter 4 presented another application of the index by addressing the sustainability of organic production. The chapter compared differences between conventional and organic farms descriptively by farm type and size, then used a difference-in-differences (DiD) model to examine the potential for a causal relationship between AS and the conversion to organic production.

Results of Chapter 4 did not find conclusive evidence of a causal relationship between AS and converting to organic production; however, the analysis did reveal that organic farms are more sustainable on average than conventional farms across all farm sizes and types.

5.2 Limitations and future research

While the dissertation has demonstrated potential for the proposed AS index, there are three main limitations to consider. The following subsections discuss these limitations and suggest directions for future research to improve the proposed index.

5.2.1 Large-scale feasibility

The use of a single country in the development of the index presents limited opportunities for testing the large-scale flexibility of the model. While the simulations between the former East and West Germany in Chapter 2 have shown that scale linking is a plausible option, the differences between these regions are minimal relative to the differences across the Global North and South divide, for example. Comparisons between developed countries such as Germany and developing countries differ significantly in terms of both the farms themselves, as well as the availability and quality of the data that can be used to build an AS index.

An important first step towards an international AS index using IRT would be to take inventory of available agricultural data sets (and the variables included in them) in different regions around the world. While this would be an arduous task, the advantage to starting here is that a full inventory of agricultural variables would enable the development of a set of common items applicable to all countries that could be used as a "base" model. As mentioned in Chapter 2, these items would likely be limited to basic factors of operation such as the land size of the farm, monetary values for output and expenditures, and input quantities (e.g. labor).

5.2.2 Holistic approach versus multidimensional

While Chapter 4 shows an advantage of the holistic approach to estimating AS as a single measurement, there may be important trade-offs between the sustainability

dimensions that are not accounted for. Examples may include hiring more employees (which is beneficial for social sustainability by increasing rural employment opportunities) or investing in new technologies that reduce environmental impacts, both of which would come at the expense of the farms' economic sustainability by lowering profits.

Overcoming this limitation would involve the development and testing of a multi-dimensional version of the AS index. Under this approach, each item should be categorized according to the sustainability dimension it belongs in (i.e. economic, environmental, or social) to build the basic framework of the index. Estimating the index could then be done compartmentally by estimating three unidimensional models individually, or by using a more complex multi-dimensional item response theory model (MIRT). Reckase (1997) and McDonald (2000) provide overviews of the latter approach, and statistical packages such as `mirt` in R (Chalmers 2012) are designed for this application.

5.2.3 Limited observation period

Lastly, the observation period currently used in the dissertation limits the potential for examining long-term effects. This limitation was demonstrated in Chapter 4, where findings show clear differences between organic and conventional farms, yet any measurable effects to sustainability were not becoming realized until the later stages of the analysis. While other unobserved factors may account for the lack of conclusive evidence, it is likely that a longer duration of study may have provided a more nuanced understanding to the effects to AS over time.

The obvious solution to this limitation is to simply extend the observation period; however, one must consider the limitations of the data available as well. This is not a problem for the case of FADN or other large statistical databases found in e.g. the United States or Canada, but many developing countries may offer less freedom in extending the observation period.

Bibliography

Antoni, Manfred, A Ganzer, and P von Berge. 2019. "Stichprobe der Integrierten Arbeitsmarktbiografien Regionalfile (SIAB-R) 1975 - 2017." *FDZ-Datenreport*, 04/2019 (de).

Baležentis, Tomas. 2015. "The sources of the total factor productivity growth in Lithuanian family farms: A Färe-Primont index approach." *Prague Economic Papers*, 24(2): 225–241.

Baquero, Grau, Bernat Esteban, Jordi-Roger Riba, Antoni Rius, and Rita Puig. 2011. "An evaluation of the life cycle cost of rapeseed oil as a straight vegetable oil fuel to replace petroleum diesel in agriculture." *Biomass and Bioenergy*, 35(8): 3687–3697.

Bastos Lima, Mairon G. 2018. "Toward multipurpose agriculture: Food, fuels, flex crops, and prospects for a bioeconomy." *Global Environmental Politics*, 18(2): 143–150.

Batalla, MI, M Pinto, and O Del Hierro. 2014. "Environmental, social and economic aptitudes for sustainable viability of sheep farming systems in northern Spain." In *11th European IFSA Symposium 'Farming systems facing global challenges: Capacities and strategies'*. 1493–1502. International Farming Systems Association (IFSA) Europe.

Bau, David B., Gary A. Hachfeld, C. Robert Holcomb, Nathan J. Hulinsky, and Megan L. Roberts. 2018. "Ratios and measurements in farm finance." *https://extension.umn.edu/farm-finance/ratios-and-measurements#solvency-796061*.

Beckenbach, Frank, Ulrich Hampicke, and Werner Schulz. 1988. "Möglichkeiten und Grenzen der Monetarisierung von Natur und Umwelt." *Schriftenreihe des IÖW*, 20(88): 3–18.

Beckmann, Volker, and Konrad Hagedorn. 2018. "Decollectivisation and privatisation policies and resulting structural changes of agriculture in Eastern Germany." In *Agricultural privatisation, land reform and farm restructuring in Central and Eastern Europe*. 105–155. Routledge.

Binder, Claudia R, Giuseppe Feola, and Julia K Steinberger. 2010. "Considering the normative, systemic and procedural dimensions in indicator-based sustainability assessments in agriculture." *Environmental Impact Assessment Review*, 30(2): 71–81.

Blasi, Emanuele, N Passeri, Silvio Franco, and A Galli. 2016. "An ecological footprint approach to environmental–economic evaluation of farm results." *Agricultural Systems*, 145: 76–82.

BMBF. 2015. "Bioeconomy in Germany: Opportunities for a bio-based and sustainable future." *https://www.bmbf.de/SharedDocs/Publikationen/de/bmbf/FS/31106_Bioeconomy_in_Germany_en.pdf*.

BMBF. 2020. "National bioeconomy strategy." *https://www.bmbf.de/bmbf/shareddocs/downloads/files/bmbf_bioeconomy-strategy_summary_en.pdf?__blob=publicationFile&v=1*.

Bollen, Kenneth A, and Adamantios Diamantopoulos. 2017. "In defense of causal-formative indicators: A minority report." *Psychological Methods*, 22(3): 581.

Bonny, Sylvie. 2011. "Herbicide-tolerant transgenic soybean over 15 years of cultivation: pesticide use, weed resistance, and some economic issues. The case of the USA." *Sustainability*, 3(9): 1302–1322.

Brundtland, Gro Harlem. 1987. "Our common future—Call for action." *Environmental Conservation*, 14(4): 291–294.

Buckley, Cathal, David P Wall, Brian Moran, and Paul NC Murphy. 2015. "Developing the EU Farm Accountancy Data Network to derive indicators around the sustainable use of nitrogen and phosphorus at farm level." *Nutrient Cycling in Agroecosystems*, 102(3): 319–333.

Buckley, Cathal, David P Wall, Brian Moran, Stephen O'Neill, and Paul NC Murphy. 2016. "Farm gate level nitrogen balance and use efficiency changes post implementation of the EU Nitrates Directive." *Nutrient Cycling in Agroecosystems*, 104(1): 1–13.

Bürkner, Paul-Christian. 2017. "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1): 1–28.

Bürkner, Paul-Christian. 2019. "Bayesian Item Response Modeling in R with brms and Stan." Aalto University, Department of Computer Science arXiv preprint arXiv:1905.09501.

Cai, Li, Kilchan Choi, Mark Hansen, and Lauren Harrell. 2016. "Item Response Theory." *Annual Review of Statistics and Its Application*, 3(1): 297–321.

Callaway, Brantly, and Pedro HC Sant'Anna. 2021. "Difference-in-differences with multiple time periods." *Journal of Econometrics*, 225(2): 200–230.

Canavari, Maurizio, Rino Ghelfi, Kent D Olson, and Sergio Rivaroli. 2007. "A comparative profitability analysis of organic and conventional farms in Emilia-Romagna and in Minnesota." 31–45, Springer.

Cappellari, Lorenzo, and Stephen P. Jenkins. 2007. "Summarizing multiple deprivation indicators." In *Inequality and Poverty Re-Examined.* , ed. Stephen P. Jenkins and John Micklewright, 166–184. Oxford:Oxford University Press.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1): 1–32.

Chalmers, R Philip. 2012. "mirt: A multidimensional item response theory package for the R environment." *Journal of statistical Software*, 48: 1–29.

Chen, Wenhao, and Nicholas M Holden. 2017. "Social life cycle assessment of average Irish dairy farm." *The International Journal of Life Cycle Assessment*, 22(9): 1459–1472.

Coderoni, S, G Bonati, L D'Angelo, D Longhitano, M Mambella, A Papaleo, and S Vanino. 2013. "Using FADN data to estimate agricultural greenhouse gases emissions at farm level." *Pacioli*, 20: 13–054.

Coderoni, Silvia, and Roberto Esposti. 2018. "CAP payments and agricultural GHG emissions in Italy. A farm-level assessment." *Science of the Total Environment*, 627: 427–437.

Conway, Gordon R, and Edward B Barbie. 1988. "After the green revolution: sustainable and equitable agricultural development." *Futures*, 20(6): 651–670.

Coomes, Oliver T, Bradford L Barham, Graham K MacDonald, Navin Ramankutty, and Jean-Paul Chavas. 2019. "Leveraging total factor productivity growth for sustainable and resilient farming." *Nature Sustainability*, 2(1): 22–28.

Cooper, T, D Baldock, M Rayment, T Kuhmonen, I Terluin, V Swales, X Poux, D Zakeossian, and M Farmer. 2006. "An evaluation of the less favoured area measure in the 25 member states of the European Union." *London: Institute for European Environmental Policy*.

Council of European Union. 1991. "Council regulation (EU) no 2092/91." *https://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=CONSLEG:1991R2092:20060506:EN:PDF*.

Cranfield, John, Spencer Henson, and James Holliday. 2010. "The motives, benefits, and problems of conversion to organic production." *Agriculture and Human Values*, 27(3): 291–306.

Crowder, David W, and John P Reganold. 2015. "Financial competitiveness of organic agriculture on a global scale." *Proceedings of the National Academy of Sciences*, 112(24): 7611–7616.

D'Amato, Dalia, Nils Droste, Ben Allen, Marianne Kettunen, Katja Lähtinen, Jaana Korhonen, Pekka Leskinen, Brent D Matthies, and Anne Toppinen. 2017. "Green, circular, bio economy: A comparative analysis of sustainability avenues." *Journal of Cleaner Production*, 168: 716–734.

Dämmgen, Ulrich. 2009. *Calculations of Emission from German Agriculture-National Emission Inventory Report (NIR) 2009 for 2007: Tables.* VTI.

Dantsis, Theodoros, Caterina Douma, Christina Giourga, Aggeliki Loumou, and Eleni A Polychronaki. 2010. "A methodological approach to assess and compare the sustainability level of agricultural plant production systems." *Ecological Indicators*, 10(2): 256–263.

de Chaisemartin, Clément, and Xavier D'Haultfoeuille. 2022. "Two-way fixed effects and differences-in-differences with heterogeneous treatment effects: A survey." National Bureau of Economic Research, Cambridge, MA:National Bureau of Economic Research.

De Loecker, Jan. 2007. "Do exports generate higher productivity? Evidence from Slovenia." *Journal of International Economics*, 73(1): 69–98.

De Ponti, Tomek, Bert Rijk, and Martin K Van Ittersum. 2012. "The crop yield gap between organic and conventional agriculture." *Agricultural Systems*, 108: 1–9.

Destatis. 2020. "Floor area total according to types of use in Germany." *https://www.destatis.de/EN/Themes/Economic-Sectors-Enterprises/Agriculture-Forestry-Fisheries/Land-Use/Tables/areas-new.html;jsessionid=98219EB06F11567128CF53389DC92B4D.live712*.

Diazabakana, Ambre, Laure Latruffe, Christian Bockstaller, Yann Desjeux, John Finn, Edel Kelly, Mary Ryan, and Sandra Uthes. 2014. "A review of farm level indicators of sustainability with a focus on CAP and FADN." PhD diss. auto-saisine.

Dillon, Emma Jane, Thia Hennessy, Cathal Buckley, Trevor Donnellan, Kevin Hanrahan, Brian Moran, and Mary Ryan. 2015. "Measuring progress in agricultural sustainability to support policy-making." *International Journal of Agricultural Sustainability*, 14(1): 31–44.

DLG. 2016. "DLG-Nachhaltigkeitsbericht 2016." Frankfurt: DLG e.V.

Edwards, Clive A. 2020. *Sustainable agricultural systems*. CRC Press.

Efken, Josef, Walter Dirksmeyer, Peter Kreins, and Marius Knecht. 2016. "Measuring the importance of the bioeconomy in Germany: Concept and illustration." *NJAS-Wageningen Journal of Life Sciences*, 77: 9–17.

Egenolf, Vincent, and Stefan Bringezu. 2019. "Conceptualization of an Indicator System for Assessing the Sustainability of the Bioeconomy." *Sustainability*, 11(2): 443.

Ehrmann, Markus. 2010. "Assessing ecological and economic impacts of policy scenarios on farm level."

European Commission. 2000. "Community Committee for the Farm Accountancy Data Network (FADN): Definitions and Variables used in FADN."

European Commission. 2012. "Innovating for Sustainable Growth: A Bioeconomy for Europe."

European Commission. 2020. *Farm to fork strategy: For a fair, healthy and environmentally-friendly food system.* European Union.

European Commission. 2022*a*. "Becoming an organic farmer." *https://agriculture.ec.europa.eu/farming/organic-farming/becoming-organic-farmer_en*, Accessed: 2022-11-09.

European Commission. 2022*b*. "The new common agricultural policy: 2023-27."

Eurostat. 2017. "Glossary: Carbon dioxide equivalent." *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Carbon_dioxide_equivalent*.

Eurostat. 2022*a*. "Gross value added of the agricultural industry - basic and producer prices." *https://data.europa.eu/data/datasets/bwzxcrbwhsymehubjmz6mw?locale=en*.

Eurostat. 2022*b*. "Multifactor productivity." *https://ec.europa.eu/eurostat/web/experimental-statistics/multifactor-productivity*.

Eurostat. 2022*c*. "Organic farming statistics." *https://ec.europa.eu/eurostat/statistics-explained/index.php?*

Eurostat. 2023*a*. "Area under organic farming." *https://ec.europa.eu/eurostat/databrowser/view/SDG_02_40__custom_1957798/bookmark/table?lang=en&bookmarkId=b3ff6964-0783-49b9-aaa7-3ab7dc858ab5*.

Eurostat. 2023*b*. "Glossary: Standard output." *https://ec.europa.eu/eurostat/statistics-explained/index.php?title=Glossary:Standard_output_(SO)#:~:text=The%20standard%20output%20of%20an,or%20per%20head%20of%20livestock.*

Ewert, Frank, Martin K van Ittersum, Irina Bezlepkina, Olivier Therond, Erling Andersen, Hatem Belhouchette, Christian Bockstaller, Floor Brouwer, Thomas Heckelei, Sander Janssen, et al. 2009. "A methodology for enhanced flexibility of integrated assessment in agriculture." *Environmental Science & Policy*, 12(5): 546–561.

FAO. 2018*a*. *Guidelines for the measurement of productivity and efficiency in agriculture.* Food and Agriculture Organization of the United Nations, Rome.

FAO. 2018*b*. *SDG Indicator 2.4.1: Proportion of agricultural area under productive and sustainable agriculture. Methodological note.* Food and Agriculture Organization of the United Nations, Rome.

Finley, Lynn, M Jahi Chappell, Paul Thiers, and James Roy Moore. 2018. "Does organic farming present greater opportunities for employment and community development than conventional farming? A survey-based investigation in California and Washington." *Agroecology and Sustainable Food Systems*, 42(5): 552–572.

Franks, Jeremy. 2010. "Boundary organizations for sustainable land management: The example of Dutch Environmental Co-operatives." *Ecological Economics*, 70(2): 283–295.

Frater, Poppy, and Jeremy Franks. 2013. "Measuring agricultural sustainability at the farm-level: A pragmatic approach." *International Journal of Agricultural Management*, 2(4): 207–225.

Fu, Tongcheng, Yi Xu, Meng Li, Shuai Xue, Zengqiang Duan, and Guang Hui Xie. 2022. "Bioenergy Production Potential of 16 Energy Crops on Marginal Land in China." *BioEnergy Research*, 1–19.

Gawel, Erik, Nadine Pannicke, and Nina Hagemann. 2019. "A path transition towards a bioeconomy—The crucial role of sustainability." *Sustainability*, 11(11): 3005.

Geissen, Violette, Vera Silva, Esperanza Huerta Lwanga, Nicolas Beriot, Klaas Oostindie, Zhaoqi Bin, Erin Pyne, Sjors Busink, Paul Zomer, Hans Mol, et al. 2021. "Cocktails of pesticide residues in conventional and organic farming systems in Europe–Legacy of the past and turning point for the future." *Environmental Pollution*, 278: 116827.

Geist, Helmut J, and Eric F Lambin. 2001. "What drives tropical deforestation." *LUCC Report series*, 4: 116.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. 2013. *Bayesian Data Analysis.* Boca Raton:CRC Press.

Gerland, Patrick, Adrian E Raftery, Hana Ševčíková, Nan Li, Danan Gu, Thomas Spoorenberg, Leontine Alkema, Bailey K Fosdick, Jennifer Chunn, Nevena Lalic, et al. 2014. "World population stabilization unlikely this century." *Science*, 346(6206): 234–237.

Gerrard, Catherine L, Susanne Padel, and Simon Moakes. 2012. "The use of Farm Business Survey data to compare the environmental performance of organic and conventional farms." *International Journal of Agricultural Management*, 2(1): 5–16.

Glauben, Thomas, Hendrik Tietje, and Christoph R Weiss. 2005. "Analysing family farm succession: a probit and a competing risk approach." Department of Food Economics and Consumption Studies, University of Kiel. No. 724-2016-49080.

Gómez-Limón, José A, and Gabriela Sanchez-Fernandez. 2010. "Empirical evaluation of agricultural sustainability using composite indicators." *Ecological economics*, 69(5): 1062–1075.

Gómez-Limón, José A, and Laura Riesgo. 2009. "Alternative approaches to the construction of a composite indicator of agricultural sustainability: An application to irrigated agriculture in the Duero basin in Spain." *Journal of Environmental Management*, 90(11): 3345–3362.

Gómez-Limón, José A, Manuel Arriaza, and M Dolores Guerrero-Baena. 2020. "Building a composite indicator to measure environmental sustainability using alternative weighting methods." *Sustainability*, 12(11): 4398.

Goodman-Bacon, Andrew. 2021. "Difference-in-differences with variation in treatment timing." *Journal of Econometrics*, 225(2): 254–277.

Goodridge, Peter. 2007. "Multi-factor productivity analysis." *Economic and Labour Market Review*, 1(7): 32.

Haas, Guido, Frank Wetterich, and Uwe Geier. 2000. "Life cycle assessment framework in agriculture on the farm level." *The International Journal of Life Cycle Assessment*, 5(6): 345–348.

Hackenberger, Armin, Marian Rümmele, Jakob Schwerter, and Miriam Sturm. 2021. "Elections and unemployment benefits for families: Did the Family Benefit Dispute affect election outcomes in Germany?" *European Journal of Political Economy*, 66(101955): 101955.

Haenel, Hans-Dieter, Claus Rösemann, Ulrich Dämmgen, Ulrike Döring, Sebastian Wulf, Brigitte Eurich-Menden, Annette Freibauer, Helmut Döhler, Carsten Schreiner, Bernhard Osterburg, et al. 2020. *Calculations of gaseous and particulate emissions from German agriculture 1990-2018: report on methods and data (RMD) Submission 2020.* Thünen Report.

Hajian, Mohammadhadi, and Somayeh Jangchi Kashani. 2021. "Evolution of the concept of sustainability. From Brundtland Report to sustainable development goals." In *Sustainable Resource Management*. 1–24. Elsevier.

Halberg, Niels, Gerwin Verschuur, and Gillian Goodlass. 2005. "Farm level environmental indicators; are they useful?: an overview of green accounting systems for European farms." *Agriculture, ecosystems & environment*, 105(1-2): 195–212.

Hambleton, Ronald K, Hariharan Swaminathan, and H Jane Rogers. 1991. *Fundamentals of item response theory.* Vol. 2, Sage.

Hays, Ron D, Leo S Morales, and Steve P Reise. 2000. "Item response theory and health outcomes measurement in the 21st century." *Medical Care*, 38(9 Suppl): II28–II42.

Haywood, Lorren Kirsty, Nikki Funke, Michelle Audouin, Constansia Musvoto, and Anton Nahman. 2019. "The Sustainable Development Goals in South Africa: Investigating the need for multi-stakeholder partnerships." *Development Southern Africa*, 36(5): 555–569.

Heinrich, Carolyn, Alessandro Maffioli, Gonzalo Vazquez, et al. 2010. "A primer for applying propensity-score matching." *Inter-American Development Bank*.

Hennessy, Thia, Cathal Buckley, Emma Dillon, Trevor Donnellan, Kevin Hanrahan, Brian Moran, and Mary Ryan. 2013. *Measuring Farm Level*

Sustainability with the Teagasc National Farm Survey. Agricultural Economics & Farm Surveys Department, Rural Economy and Development Programme, Teagasc.

IPCC. 2006. *IPCC guidelines for national greenhouse gas inventories.* Prepared by the National Greenhouse Gas Inventories Programme.

Jiang, Wei, Michael G Jacobson, and Matthew H Langholtz. 2019. "A sustainability framework for assessing studies about marginal lands for planting perennial energy crops." *Biofuels, Bioproducts and Biorefining*, 13(1): 228–240.

Johansson, Daniel JA, and Christian Azar. 2007. "A scenario based analysis of land competition between food and bioenergy production in the US." *Climatic Change*, 82(3): 267–291.

Kehlenbeck, Hella, Jovanka Saltzmann, Jürgen Schwarz, Peter Zwerger, and Henning Nordmeyer. 2016. "Economic assessment of alternatives for glyphosate application in arable farming." *Julius-Kühn-Archiv*, 2016(452): 279.

Kelly, Edel, Laure Latruffe, Yann Desjeux, Mary Ryan, Sandra Uthes, Ambre Diazabakana, Emma Dillon, and John Finn. 2018. "Sustainability indicators for improved assessment of the effects of agricultural policy across the EU: Is FADN the answer?" *Ecological Indicators*, 89: 903–911.

Kirner, Leopold, and Ruth Kratochvil. 2006. "The role of farm size in the sustainability of dairy farming in Austria: An empirical approach based on farm accounting data." *Journal of Sustainable Agriculture*, 28(4): 105–124.

Kleemann, Linda, and Awudu Abdulai. 2013. "Organic certification, agro-ecological practices and return on investment: Evidence from pineapple producers in Ghana." *Ecological Economics*, 93: 330–341.

Klonsky, Karen. 2012. "Comparison of production costs and resource use for organic and conventional production systems." *American Journal of Agricultural Economics*, 94(2): 314–321.

Kolczynska, Marta, Paul Christian Bürkner, Lauren Kennedy, and Aki Vehtari. 2020. "Modeling Public Opinion over Time and Space: Trust in State Institutions in Europe, 1989-2019." SocArXiv.

Kolen, Michael J., and Robert L. Brennan. 2014. *Test equating, scaling, and linking: Methods and practices.* Springer.

Kostandini, Genti, Elton Mykerezi, and Cesar Escalante. 2014. "The impact of immigration enforcement on the US farming sector." *American Journal of Agricultural Economics*, 96(1): 172–192.

Krause, Josef, and Ondřej Machek. 2018. "A comparative analysis of organic and conventional farmers in the Czech Republic." *Agricultural Economics*, 64(1): 1–8.

Kretschmer, Bettina, Catherine Bowyer, and Allan Buckwell. 2012. "EU Biofuel Use and Agricultural Commodity Prices: A review of the evidence base." *Institute for European Environmental Policy (IEEP), London.*

Krieg, H, S Albrecht, and M Jäger. 2013. "Systematic monetisation of environmental impacts." *WIT Transactions on Ecology and the Environment*, 173: 513–524.

Küstermann, Björn, Maximilian Kainz, and Kurt-Jürgen Hülsbergen. 2008. "Modeling carbon cycles and estimation of greenhouse gas emissions from organic and conventional farming systems." *Renewable Agriculture and Food Systems*, 23(1): 38–52.

Läpple, Doris. 2013. "Comparing attitudes and characteristics of organic, former organic and conventional farmers: Evidence from Ireland." *Renewable Agriculture and Food Systems*, 28(4): 329–337.

Latruffe, Laure, and Laurent Piet. 2014. "Does land fragmentation affect farm performance? A case study from Brittany, France." *Agricultural Systems*, 129: 68–80.

Latruffe, Laure, and Stefan Mann. 2015. "Is part-time farming less subsidised? The example of direct payments in France and Switzerland." *Cahiers Agricultures*, 24(1): 20–27.

Lebacq, Thérésa, Philippe V Baret, and Didier Stilmant. 2013. "Sustainability indicators for livestock farming. A review." *Agronomy for Sustainable Development*, 33(2): 311–327.

Lechenet, Martin, Vincent Bretagnolle, Christian Bockstaller, François Boissinot, Marie-Sophie Petit, Sandrine Petit, and Nicolas M Munier-Jolain. 2014*a*. "Reconciling pesticide reduction with economic and environmental sustainability in arable farming." *PloS one*, 9(6): e97922.

Lechenet, Martin, Vincent Bretagnolle, Christian Bockstaller, François Boissinot, Marie-Sophie Petit, Sandrine Petit, and Nicolas M Munier-Jolain. 2014*b*. "Reconciling pesticide reduction with economic and environmental sustainability in arable farming." *PloS one*, 9(6): e97922.

Longhitano, Davide, A Bodini, A Povellato, and A Scardera. 2012. "Assessing farm sustainability. An application with the Italian FADN sample." In *1st AIEAA Conference 'Towards a Sustainable Bio-economy: Economic Issues and Policy Challenges.*

Lord, Frederic M. 1953. "The relation of test score to the trait underlying the test." *Educational and Psychological Measurement*, 13(4): 517–549.

Lynch, John, Thia Hennessy, Cathal Buckley, Emma Dillon, Trevor Donnellan, Kevin Hanrahan, Brian Moran, and Mary Ryan. 2016. "Teagasc National Farm Survey 2015 sustainability report." *Athenry, Co. Galway: Teagasc.*

MacLeod, Matthew, Hans Peter H Arp, Mine B Tekman, and Annika Jahnke. 2021. "The global threat from plastic pollution." *Science*, 373(6550): 61–65.

Marchand, Fleur, Lies Debruyne, Laure Triste, Catherine Gerrard, Susanne Padel, and Ludwig Lauwers. 2014. "Key characteristics for tool choice in indicator-based sustainability assessment at farm level." *Ecology and Society*, 19(3).

Marras, Serena, Sara Masia, Pierpaolo Duce, Donatella Spano, and Costantino Sirca. 2015. "Carbon footprint assessment on a mature vineyard." *Agricultural and Forest Meteorology*, 214: 350–356.

Martínez-Blanco, Julia, Annekatrin Lehmann, Pere Muñoz, Assumpció Antón, Marzia Traverso, Joan Rieradevall, and Matthias Finkbeiner. 2014. "Application challenges for the social Life Cycle Assessment of fertilizers within life cycle sustainability assessment." *Journal of Cleaner Production*, 69: 34–48.

Martin, Guillaume, Marc Moraine, Julie Ryschawy, Marie-Angélina Magne, Masayasu Asai, Jean-Pierre Sarthou, Michel Duru, and Olivier Therond. 2016. "Crop–livestock integration beyond the farm level: A review." *Agronomy for Sustainable Development*, 36(3): 1–21.

McCormick, Kes, and Niina Kautto. 2013. "The bioeconomy in Europe: An overview." *Sustainability*, 5(6): 2589–2608.

McDonald, Roderick P. 2000. "A basis for multidimensional item response theory." *Applied Psychological Measurement*, 24(2): 99–114.

Meemken, Eva-Marie, and Matin Qaim. 2018. "Organic agriculture, food security, and the environment." *Annual Review of Resource Economics*, 10: 39–63.

Meul, Marijke, Frank Nevens, and Dirk Reheul. 2009. "Validating sustainability indicators: Focus on ecological aspects of Flemish dairy farms." *Ecological Indicators*, 9(2): 284–295.

Meyer, J Patrick, and Shi Zhu. 2013. "Fair and equitable measurement of student learning in MOOCs: An introduction to item response theory, scale linking, and score equating." *Research & Practice in Assessment*, 8: 26–39.

Mitchell, RB, MR Schmer, WF Anderson, Virginia Jin, KS Balkcom, James Kiniry, Alisa Coffin, and Paul White. 2016. "Dedicated energy crops and crop residues for bioenergy feedstocks in the central and eastern USA." *Bioenergy research*, 9(2): 384–398.

Mondelaers, Koen, Joris Aertsens, and Guido Van Huylenbroeck. 2009. "A meta-analysis of the differences in environmental impacts between organic and conventional farming." *British Food Journal*, 111(10): 1098–1119.

Mon, Pon Nya, and David W Holland. 2006. "Organic apple production in Washington State: An input–output analysis." *Renewable Agriculture and Food Systems*, 21(2): 134–141.

Moser, Susanne C. 2010. "Communicating climate change: history, challenges, process and future directions." *Wiley Interdisciplinary Reviews: Climate Change*, 1(1): 31–53.

Mosnier, Claire, Marc Benoit, Jean Joseph Minviel, and Patrick Veysset. 2022. "Does mixing livestock farming enterprises improve farm and product sustainability?" *International Journal of Agricultural Sustainability*, 1–15.

Murawski, Steven A. 2000. "Definitions of overfishing from an ecosystem perspective." *ICES Journal of Marine Science*, 57(3): 649–658.

Nguyen, Tam H, Hae-Ra Han, Miyong T Kim, and Kitty S Chan. 2014. "An introduction to item response theory for patient-reported outcome measurement." *The Patient-Patient-Centered Outcomes Research*, 7(1): 23–35.

Niggli, Urs. 2015. "Sustainability of organic food production: challenges and innovations." *Proceedings of the Nutrition Society*, 74(1): 83–88.

OECD. 2008. *Handbook on constructing composite indicators: Methodology and user guide.* OECD publishing, Paris.

OECD. 2009. *The bioeconomy to 2030: Designing a policy agenda.* OECD Publishing, Paris.

OECD. 2019. "Under pressure: The squeezed middle class." *OECD Publishing*.

Padel, Susanne. 2001. "Conversion to organic farming: A typical example of the diffusion of an innovation?" *Sociologia Ruralis*, 41(1): 40–61.

Patterson, Murray, Garry McDonald, and Derrylea Hardy. 2017. "Is there more in common than we think? Convergence of ecological footprinting, emergy analysis, life cycle assessment and other methods of environmental accounting." *Ecological Modelling*, 362: 19–36.

Peterson, Christina Hamme, Karen L Gischlar, and N Andrew Peterson. 2017. "Item construction using reflective, formative, or Rasch measurement models: Implications for group work." *The Journal for Specialists in Group Work*, 42(1): 17–32.

Petrick, Martin, and Patrick Zier. 2011. "Regional employment impacts of Common Agricultural Policy measures in Eastern Germany: A difference-in-differences approach." *Agricultural Economics*, 42(2): 183–193.

Pfau, Swinda F, Janneke E Hagens, Ben Dankbaar, and Antoine JM Smits. 2014. "Visions of sustainability in bioeconomy research." *Sustainability*, 6(3): 1222–1249.

Pimentel, David. 1995. "Amounts of pesticides reaching target pests: Environmental impacts and ethics." *Journal of Agricultural and Environmental Ethics*, 8(1): 17–29.

Prasara-A, Jittima, Shabbir H Gheewala, Thapat Silalertruksa, Patcharaporn Pongpat, and Wanchat Sawaengsak. 2019. "Environmental and social life cycle assessment to enhance sustainability of sugarcane-based products in Thailand." *Clean Technologies and Environmental Policy*, 21(7): 1447–1458.

Prenovost, KM, SD Fihn, ML Maciejewski, K Nelson, S Vijan, and AM Rosland. 2018. "Using item response theory with health system data to identify latent groups of patients with multiple health conditions." *PLoS One*, 13(11): e0206915.

Pretty, Jules. 2008. "Agricultural sustainability: concepts, principles and evidence." *Philosophical Transactions of the Royal Society B: Biological Sciences*, 363(1491): 447–465.

Raghu, S, JL Spencer, AS Davis, and RN Wiedenmann. 2011. "Ecological considerations in the sustainable development of terrestrial biofuel crops." *Current Opinion in Environmental Sustainability*, 3(1-2): 15–23.

R Core Team. 2021. "R: A Language and Environment for Statistical Computing." Vienna, Austria, R Foundation for Statistical Computing.

Reap, John, Felipe Roman, Scott Duncan, and Bert Bras. 2008. "A survey of unresolved problems in life cycle assessment." *The International Journal of Life Cycle Assessment*, 13(5): 374–388.

Reckase, Mark D. 1997. "The past and future of multidimensional item response theory." *Applied Psychological Measurement*, 21(1): 25–36.

Reckase, Mark D. 2009. "Multidimensional item response theory models." In *Multidimensional item response theory*. 79–112. Springer.

Reganold, John P, and Jonathan M Wachter. 2016. "Organic agriculture in the twenty-first century." *Nature Plants*, 2(2): 1–8.

Reidsma, Pytrik, Tonnie Tekelenburg, Maurits Van den Berg, and Rob Alkemade. 2006. "Impacts of land-use change on biodiversity: An assessment of agricultural biodiversity in the European Union." *Agriculture, Ecosystems & Environment*, 114(1): 86–102.

Ren, Chenchen, Shen Liu, Hans Van Grinsven, Stefan Reis, Shuqin Jin, Hongbin Liu, and Baojing Gu. 2019. "The impact of farm size on agricultural sustainability." *Journal of Cleaner Production*, 220: 357–367.

Rösemann, Claus, Hans-Dieter Haenel, Cora Vos, Ulrich Dämmgen, Ulrike Döring, Sebastian Wulf, Brigitte Eurich-Menden, Annette Freibauer, Helmut Döhler, Carsten Schreiner, et al. 2021. *Calculations of gaseous and particulate emissions from German agriculture 1990-2019: Report on methods and data (RMD) Submission 2021.* Thünen Report.

Roth, Jonathan, Pedro HC Sant'Anna, Alyssa Bilinski, and John Poe. 2022. "What's trending in difference-in-differences? A synthesis of the recent econometrics literature." *arXiv preprint arXiv:2201.01194*.

Roy, Ranjan, and Ngai Weng Chan. 2012. "An assessment of agricultural sustainability indicators in Bangladesh: Review and synthesis." *The Environmentalist*, 32(1): 99–110.

Ryan, Mary, Thia Hennessy, Cathal Buckley, Emma J Dillon, Trevor Donnellan, Kevin Hanrahan, and Brian Moran. 2016. "Developing farm-level sustainability indicators for Ireland using the Teagasc National Farm Survey." *Irish Journal of Agricultural and Food Research*, 55(2): 112–125.

Sachs, Jeffrey D. 2012. "From millennium development goals to sustainable development goals." *The Lancet*, 379(9832): 2206–2211.

Samejima, Fumiko. 1969. "Estimation of latent ability using a response pattern of graded scores." *Psychometrika monograph supplement*.

Samejima, Fumiko. 1997*a*. "Graded response model." In *Handbook of Modern Item Response Theory*. , ed. Wim J. van der Linden and Ronald K. Hambleton, 85–100. New York:Springer.

Samejima, Fumiko. 1997*b*. "Graded response model." In *Handbook of modern item response theory*. 85–100. Springer.

Schaller, Neill. 1993. "The concept of agricultural sustainability." *Agriculture, Ecosystems & Environment*, 46(1-4): 89–97.

Schiefer, Jasmin, Georg J Lair, and Winfried EH Blum. 2015. "Indicators for the definition of land quality as a basis for the sustainable intensification of agricultural production." *International Soil and Water Conservation Research*, 3(1): 42–49.

Seufert, Verena, and Navin Ramankutty. 2017. "Many shades of gray—The context-dependent performance of organic agriculture." *Science Advances*, 3(3): e1602638.

Seufert, Verena, Navin Ramankutty, and Jonathan A Foley. 2012. "Comparing the yields of organic and conventional agriculture." *Nature*, 485(7397): 229–232.

Sheppard, Andy W, Iain Gillespie, Mikael Hirsch, and Cameron Begley. 2011. "Biosecurity and sustainability within the growing global bioeconomy." *Current Opinion in Environmental Sustainability*, 3(1-2): 4–10.

Shortall, OK. 2013. ""Marginal land" for energy crops: Exploring definitions and embedded assumptions." *Energy Policy*, 62: 19–27.

Sippel, Sebastian, Nicolai Meinshausen, Erich M Fischer, Enikő Székely, and Reto Knutti. 2020. "Climate change now detectable from any single day of weather at global scale." *Nature Climate Change*, 10(1): 35–41.

Slavickiene, Astrida, and Jurate Savickiene. 2014. "Comparative analysis of farm economic viability assessment methodologies." *European Scientific Journal*, 10(7).

Spicka, Jindřich, Tomas Hlavsa, Katerina Soukupova, and Marie Stolbova. 2019. "Approaches to estimation the farm-level economic viability and sustainability in agriculture: A literature review." *Agricultural Economics*, 65(6): 289–297.

Stott, Peter. 2016. "How climate change affects extreme weather events." *Science*, 352(6293): 1517–1518.

Sulewski, Piotr, Anna Kłoczko-Gajewska, and Wojciech Sroka. 2018. "Relations between agri-environmental, economic and social dimensions of farms' sustainability." *Sustainability*, 10(12): 4629.

Sullivan, Preston. 2003. "Applying the principles of sustainable farming." *National Center for Appropriate Technology*.

Talukder, Byomkesh, Keith W Hipel, Gary W vanLoon, et al. 2017. "Developing composite indicators for agricultural sustainability assessment: Effect of normalization and aggregation techniques." *Resources*, 6(4): 66.

Tang, Hengyun, Jianqing Zhang, Fei Fan, and Zhengwen Wang. 2022. "High-speed rail, urban form, and regional innovation: A time-varying difference-in-differences approach." *Technology Analysis & Strategic Management*, 1–15.

Terres, Jean-Michel, Luigi Nisini Scacchiafichi, Annett Wania, Margarida Ambar, Emeric Anguiano, Allan Buckwell, Adele Coppola, Alexander Gocht, Helena Nordström Källström, Philippe Pointereau, et al. 2015. "Farmland abandonment in Europe: Identification of drivers and indicators, and development of a composite indicator of risk." *Land Use Policy*, 49: 20–34.

Thorne, FS, and W Fingleton. 2006. "Examining the relative competitiveness of milk production: An Irish case study (1996–2004)." *Journal of International Farm Management*, 3(4): 49–61.

Tubiello, Francesco N, Mirella Salvatore, Simone Rossi, Alessandro Ferrara, Nuala Fitton, and Pete Smith. 2013. "The FAOSTAT database of greenhouse gas emissions from agriculture." *Environmental Research Letters*, 8(1): 015009.

UC-Berkeley. 2020. "List of common conversion factors (engineering conversion factors)." *http://w.astro.berkeley.edu/ wright/fuel_energy.html*.

Umweltbundesamt, Bundesanstalt für Geowissenschaften und Rohstoffe, Statistisches Bundesamt. 2007.

United Nations, Department of Economic, and Social Affairs Sustainable Development. 2015. "Transforming our world: the 2030 Agenda for Sustainable Development."

Vallance, Suzanne, Harvey C Perkins, and Jennifer E Dixon. 2011. "What is social sustainability? A clarification of concepts." *Geoforum*, 42(3): 342–348.

Vandemoortele, Milo. 2014. "Measuring household wealth with latent trait modelling: An application to Malawian DHS data." *Social Indicators Research*, 118(2): 877–891.

van der Linden, Wim J., and Ronald K. Hambleton. 1997. "Item Response Theory: Brief History, Common Models, and Extensions." In *Handbook of Modem Item Response Theory.* , ed. Wim J. van der Linden and Ronald K. Hambleton, 1–28. Springer.

van der Meulen, HAB, MA Dolman, JH Jager, and GS Venema. 2014. "The impact of farm size on sustainability of Dutch dairy farms." *International Journal of Agricultural Management*, 3(2): 119–123.

Vázquez-Rowe, Ian, Benedetto Rugani, and Enrico Benetto. 2013. "Tapping carbon footprint variations in the European wine sector." *Journal of Cleaner Production*, 43: 146–155.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432.

Venkat, Kumar. 2012. "Comparison of twelve organic and conventional farming systems: a life cycle greenhouse gas emissions perspective." *Journal of Sustainable Agriculture*, 36(6): 620–649.

Vitunskiene, Vlada, and Vida Dabkiene. 2016. "Framework for assessing the farm relative sustainability: a Lithuanian case study." *Agricultural Economics*, 62(3): 134–148.

Waltner-Toews, David. 1996. "Ecosystem health: A framework for implementing sustainability in agriculture." *Bioscience*, 46(9): 686–689.

Weingarten, Peter, Jürgen Bauhus, Ulrike Arens-Azevedo, and Alfons Balmann. 2016. "Climate change mitigation in agriculture and forestry and in the downstream sectors of food and timber use."

Westbury, DB, JR Park, AL Mauchline, RT Crane, and SR Mortimer. 2011. "Assessing the environmental performance of English arable and livestock holdings using data from the Farm Accountancy Data Network (FADN)." *Journal of Environmental Management*, 92(3): 902–909.

West, TO, and G Marland. 2002. "Net carbon flux from agricultural ecosystems: methodology for full carbon cycle analyses." *Environmental Pollution*, 116(3): 439–444.

Whitehead, Jay, Yuan Lu, Holly Still, Jonathan Wallis, Hannah Gentle, Henrik Moller, et al. 2016. "Target setting and burden sharing in sustainability assessment beyond the farm level." 12–15.

White House. 2012. "National Bioeconomy Blueprint." *https://obamawhitehouse.archives.gov/sites/default/files/microsites/ostp/national_bioeconomy_blueprint_april_2012.pdf*.

White, Mark A. 2013. "Sustainability: I know it when I see it." *Ecological Economics*, 86: 213–217.

Willer, Helga, Diana Schaack, and Julia Lernoud. 2019. "Organic farming and market development in Europe and the European Union." In *The World of Organic Agriculture. Statistics and Emerging Trends 2019*. 217–254. Research Institute of Organic Agriculture FiBL and IFOAM-Organics International.

Wohlfahrt, Julie, Fabien Ferchaud, Benoit Gabrielle, C Godard, Bernard Kurek, Chantal Loyce, and Olivier Therond. 2019. "Characteristics of bioeconomy systems and sustainability issues at the territorial scale. A review." *Journal of Cleaner Production*, 232: 898–909.

Wood, Simon N. 2017. *Generalized Additive Models: An Introduction with R.* Boca Raton:Chapman & Hall/CRC Press.

Woods, J. 2019. "Total Factor Productivity for England by Farm Type, based on the Farm Business Survey." Department for Environment, Food and Rural Affairs, United Kingdom Government.

Xu, Yuting, Xianjin Huang, Helen XH Bao, Xiang Ju, Taiyang Zhong, Zhigang Chen, and Yan Zhou. 2018. "Rural land rights reform and agro-environmental sustainability: Empirical evidence from China." *Land use policy*, 74: 73–87.

Yount, Kathryn M, Yuk Fai Cheong, Lauren Maxwell, Jessica Heckert, Elena M Martinez, and Gregory Seymour. 2019. "Measurement properties of the project-level Women's Empowerment in Agriculture Index." *World development*, 124: 104639.

Zhang, BO. 2008. "Application of unidimensional item response models to tests with items sensitive to secondary dimensions." *The Journal of Experimental Education*, 77(2): 147–166.

Zhen, Lin, and Jayant K Routray. 2003. "Operational indicators for measuring agricultural sustainability in developing countries." *Environmental management*, 32(1): 34–46.

Ziegler, Matthias, and Dirk Hagemann. 2015. "Testing the unidimensionality of items: Pitfalls and loopholes." *European Journal of Psychological Assessment.*

Zilberman, David, Gal Hochman, Deepak Rajagopal, Steve Sexton, and Govinda Timilsina. 2013. "The impact of biofuels on commodity food prices: Assessment of findings." *American Journal of Agricultural Economics*, 95(2): 275–281.

Zwilling, B., and D. Raab. 2019. "Solvency on the farm." *Farmdoc Daily*, (9):176.

Appendix A

A.1 Item selection considerations and definitions

In the literature, **profitability** (item 1 in table A.3) is measured in a variety of ways such as income relative to operation income (Ehrmann 2010) and gross margin per hectare (Ryan et al. 2016). Perhaps more appropriate in the context of sustainability, van der Meulen et al. (2014) calculates profitability as a ratio of farm net income to unpaid annual work units (AWU). The inclusion of unpaid (family) labor in the equation can (a) ensure adequate accounting for farmer(s) income sufficiency (Gómez-Limón, Arriaza and Guerrero-Baena 2020), (b) control for variations in farm size (van der Meulen et al. 2014), and (c) reflect the potential for contributing to social sustainability attributes (Sullivan 2003) such as inter-generational succession of the farm[1] and the ability for farmers to contribute to the local economy. This method, however, is unsuitable for corporate or cooperative farms that do not have unpaid labor. Since all labor costs are already compensated in these cases, the proposed solution to measure family and corporate farms equally is to subtract from family farms' income an allowance for unpaid (family) labor. We compute this allowance as family labor input quantity times the median wage $\tilde{w}$ for the federal state the farm is located in (denoted by subscript fs) FNI as:

[1] A study by Glauben, Tietje and Weiss (2005) found that profitability of the farm has a significant influence on the likelihood of inter-generational succession in Germany.

$$i_{profit} = FNI - \tilde{w}_{fs}, \tag{A.1}$$

where i_{π} is the farm's profit and FNI is calculated as total output plus the balance of current subsidies and taxes, less depreciation, intermediate costs (specific costs and farm overhead costs), and total external factors (wages, rent, and interest paid) (European Commission 2000). In this calculation, the variable is a absolute sum for the whole farm, which is then converted to a relative calculation when converted to an ordinal item.

Solvency (item 2) is calculated using the same equation as Vitunskiene and Dabkiene (2016), the indicator uses a common formulation as the farms' total debts to total assets:

$$i_{solvency} = \frac{D_T}{A_T}, \tag{A.2}$$

where D_T is the combined total T of short-, medium-, and long-term loans, and A_T refers to the total of both long-term fixed assets (land, buildings, machinery, and breeding livestock) and short-term current assets (non-breeding livestock, the stock of agricultural products, and other circulating capital) (European Commission 2000).

Wage ratio WR (item 3) is measured as the ratio of the average hourly wage paid on the farm (total wages paid to total paid labor hours) to the median wage of the region. In contrast to Vitunskiene and Dabkiene (2016) who calculates this indicator as the ratio of average annual wages for farm workers to average wages in the whole country, the calculation for this index is the ratio of average wages on the farm to the median income in each NUTS3 region to capture regional differences in purchasing power and the cost of living:

$$WR = \frac{\overline{w}_h}{\overline{w}_h^{n3}}, \tag{A.3}$$

where $\overline{w}_h$ is the mean hourly h wage on the farm and $\overline{w}_h^{n3}$ is the mean hourly wage in the NUTS3 $n3$ region the farm is located in.

Whereas the fourth item for **economic diversity** ED is calculated as a dummy variable used in SDG 2.4.1 to signal if a single agricultural product accounts for more than 66% of total output on the farm, we instead calculate the indicator as a continuous value:

$$i_{ED} = max\left(\frac{\phi_n}{O_T}\right), \tag{A.4}$$

where the value for economic diversity is the maximum max contribution of a single nth product ϕ to the farm's total output O_T, with n representing any combination of the 19 products specified in FADN such as grains, milk, wine, etc.

The **provision of employment** indicator PE (item 5) is calculated as the ratio of total expenditure on wages w_T and contracted work c to total output O_T to control for farm size and reflect the intensity to which farms are providing income opportunities for agricultural workers and businesses in the region:

$$PE = \frac{w_T + c}{O_T}, \tag{A.5}$$

with contracted work including the expenditures for both contracted services and hired machinery (European Commission 2000).

The **expenditure on pesticides** EP (item 6) is measured as the total pesticide expenditure EP_t (in €) over the farm's land area in UAA L_{UAA}:

$$EP_{UAA} = \frac{EP_T}{UAA_{ha,T}}. \tag{A.6}$$

The seventh item for **GHG emissions** measures the total emissions produced by the farm and includes both direct emissions (e.g. nitrogen from fertilizers) and indirect emissions (e.g. CO_2 from fuel consumption). The framework of emission sources used for estimating total emissions developed by Coderoni and Esposti (2018) and Coderoni et al. (2013) is modified to suit German agriculture[2]. All calculations are performed in accordance to IPCC (2006) tier-1 and tier-2 estimates and summed to produce a total level of emissions for the farm:

$$GHG_{CO_2-eq} = \frac{\Sigma_{s=1}^{11} GHG_s \times GWP_v}{VA_G}, \tag{A.7}$$

where the level of CO_2 equivalent greenhouse gas emissions GHG_{CO_2-eq} on the farm is the sum of greenhouse gases from source s converted to CO_2 equivalents using their global warming potential values GWP_v (see Eurostat 2017), as shown in Table A.1. To reflect the farm's $CO_{2\text{-eq}}$ intensity, total emissions are divided by the farm's gross value added VA_G (see Umweltbundesamt 2007), which is the total monetary value received by the producer including subsidies, less taxes and intermediate consumption (Eurostat 2022*a*).

We define **multi-factor productivity** MFP (item 8) as the quotient of total value added VA_T over the sum of quantities q of the three factor inputs I (land, labor, and capital) at given factor prices p_I:

$$i_{MFP} = \frac{VA_T}{\sum p_I q_I}, \tag{A.8}$$

[2]Variables for CH_4 from rice cultivation and biological N fixation are not included in the German FADN data set and are thus omitted from the calculation.

where VA_T is the sum of total output O_T and subsidies and taxes, less total intermediate input $I_{int,T}$ expenditure (i.e. specific costs and farming overheads):

$$VA_T = O_T + S_T - I_{int,T}. \tag{A.9}$$

Factor inputs I are estimated on an annual basis (i.e. one full agricultural season) (see FAO 2018*a*, p. 57) to reflect the opportunity costs of using the inputs for production. The price of land input I_l includes the total expenditure for rented land r_T plus an opportunity cost that is the sum of land owned o in hectares $UAA_{ha,o}$ times an estimated rental value of the land according to the average rent paid per hectare $\overline{r}_{ha,TF14}$ TF14 farm type:

$$I_l = r_T + \Sigma(UAA_{ha,o} \times \overline{r}_{ha,TF14}). \tag{A.10}$$

Similarly, labor input i_L is an aggregation of the farm's total wage expenditure for paid labor $w_{T,pl}$ plus an estimated opportunity cost equal to the total quantity of unpaid labor in annual work units $L_{awu,ul}$ times an expected return to labor based on the average wage per *awu* of paid labor $\overline{w}_{awu,pl}$ on the farm:

$$i_L = w_{T,pl} + \Sigma(L_{awu,ul} \times \overline{w}_{awu,pl}). \tag{A.11}$$

Finally, the value for capital input i_K is calculated as the total interest paid by the farm ι_T plus the net worth of the farm capital K (without land values) multiplied by an assumed interest rate ι_{ar} of 4%:

$$i_K = \iota_t + (K \times \iota_{ar}). \tag{A.12}$$

In contrast to other studies that use the Shannon Index (e.g. Westbury et al. 2011; Gerrard, Padel and Moakes 2012) measuring the quantity and evenness of different types of land use, **land ecosystem quality** LEQ (item 9) assigns different values to agricultural land based on the type and intensity of production on the farm. The total ecosystem quality value LEQ_T is calculated as:

$$LEQ_T = \frac{(L_{ip} \times EQ_{lt}) + UAA_{np}}{L_T}, \tag{A.13}$$

where L_{ip} is the total land used in production ip (e.g. cereals, vineyards, etc.), which is weighted by the ecological quality percentage EQ of land type lt (see table A.2). The product is then added to land with no production L_{np} (woods, agricultural fallows and land set aside, and natural grassland), and divided by the total land area of the farm: $L_T = L_{ip} + L_{np}$. All land values are reported in hectares and the EQ_{lt} values shown in table A.2 are derived from Reidsma et al. (2006).

GHG type	GWP_v	GHG emission source (GHG_s)
CH_4	25	Manure management
		Enteric fermentation
N_2O	298	Manure management
		Synthetic fertilizers
		Crop residue
		Atmospheric deposition
		Leeching and run-off
CO_2	1	Energy
		Forest land
		Cropland
		Grassland

Table A.1: GHG emission sources. Source: Adapted from Coderoni et al. (2013)

EQ_{lt}	Description of L_{ip}	O	Ir	G	LU	In
0.5	Irrigated		X			
	Highly intensive					>250
0.10	Intensive					80-250
0.15	Highly intensive organic	X	X			>250
	Intensive arable grazing livestock			<66%	>80	
0.20	Intensive organic	X				80-250
	Intensive pasture			>66%	<2	<250
	Highly intensive pasture			>66%	>2 OR	>250
0.25	Extensive					<80
0.325	Extensive arable grazing livestock			<66%	<1	<80
0.35	Extensive organic	X				<80
0.4	Extensive pasture			>66%	<1	<80
1	Natural grassland			>66%	<0.3	
	No production					

EQ_{lt} = Ecological quality of land

L_{ip} = land in production

O = Denotes if the farm is organic, partially organic, or transitioning to organic

Ir = Denotes if the farm has installed irrigation

G = Percentage of land for forage crops (% of total used land

LU = Number of livestock units per hectare

In = Value of direct inputs (fertilizers, pesticides, and feedstuffs for grazing) per hectare

Table A.2: Percentage values of land ecosystem quality. Source: Reidsma et al. (2006)

A.2 Mapping of AS items to discrete (ordinal) categories

Table A.3 outlines the thresholds defining four discrete classes for each continuous item used in the GRM model. Wherever possible, the threshold values are assigned using absolute values from external sources. Debt to asset ratio thresholds for solvency (item 2) are provided by the University of Minnesota Extension (Bau et al. 2018). Thresholds for pesticide expenditure (item 6) involve converting the expenditure into kilograms (kg) of active ingredient (glyphosate) per hectare of land. Kehlenbeck et al. (2016) suggest a maximum of 3.6kg/ha, with the cost per kg being estimated from US data in Bonny (2011). Finally, wage ratio (item 3) thresholds are determined from (OECD 2019) definitions based on the agricultural worker's income relative to the median income in Germany.

The remaining indicator thresholds are assigned using relative statistical ranges. Economic diversity (item 4) and the provision of employment (item 5) should theoretically be a value with a range of [0, 1], so the cutoff points in the scale are set to 0.25, 0.5, and 0.75. Similarly, the range for land ecosystem quality (item 9) is normalized to create the same scale and threshold values. Thresholds for MFP (item 8) are set to 0, 0.5, and 1 to reflect negative productivity, 50% returns to inputs, and 100% returns to inputs. For profitability (item 1), thresholds are defined according to percentiles of the distribution of paid wages. Finally, the threshold values for GHG emissions (item 7) are calculated based on the median with cutoff points at 0 for farms that are net zero or negative in emission output, the median value, and twice the median value.

Item #	Variable	Description	Cat.	Thresholds	Obs.
1	Profitability	Farm net income less an allowance for unpaid labor based on percentiles of regional agricultural wages	1	$x < 0$	4127
			2	$0 \leq x < WA_{p25}$	1099
			3	$WA_{p25} \leq x < WA_{med}$	1629
			4	$WA_{med} \leq x$	2073
2	Solvency	Ratio of total debts to total assets	1	$1 \leq x$	207
			2	$0.6 \leq x < 1$	660
			3	$0.3 \leq x < 0.6$	1651
			4	$x < 0.3$	6410
3	Wage ratio	Ratio of average wages paid on the farm to overall median wage in the region (NUTS 3)	1	$x < 0.5$	6038
			2	$0.5 \leq x < 0.75$	1550
			3	$0.75 \leq x < 2$	1337
			4	$2 \leq x$	3
4	Economic diversity	The maximum percentage of a single agricultural product to total output	1	$0.75 \leq x$	2993
			2	$0.5 \leq x < 0.75$	3736
			3	$0.25 \leq x < 0.5$	2173
			4	$x < 0.25$	26
5	Provision of employment	Ratio of total expenditure on wages and contract work to total output of the farm	1	$x < 0.25$	8137
			2	$0.25 \leq x < 0.5$	698
			3	$0.5 \leq x < 0.75$	58
			4	$0.75 \leq x$	35

Item #	Variable	Description	Cat.	Thresholds	Obs.
6	Expenditure on pesticides	Ratio of total pesticide expenditure to total utilized agricultural area (in €/ha)	1	$144 < x$	3318
			2	$72 \leq x < 144$	2268
			3	$36 \leq x < 72$	1196
			4	$x \leq 36$	2146
7	GHG emissions	GHG intensity on the farm as a ratio of annual CO_2 equivalent greenhouse gases $CO2e$ emitted/absorbed to gross value added ($GHG = CO2e/€$)	1	$GHG_{2\times med} < x$	1609
			2	$GHG_{med} \leq x < GHG_{2\times med}$	2550
			3	$0 \leq x < GHG_{med}$	4308
			4	$x < 0$	461
8	Multi-factor productivity	Ratio of total value added to factor inputs for land, labor, and capital	1	$x \leq 0$	403
			2	$0 \leq x < 0.5$	1427
			3	$0.5 \leq x < 1$	2858
			4	$1 \leq x$	4240
9	Land ecosystem quality	Estimated land quality (in percent) relative to pristine (untouched) natural landscape	1	$x < 0.25$	3934
			2	$0.25 \leq x < 0.5$	3008
			3	$0.5 \leq x < 0.75$	1058
			4	$0.75 \leq x$	928

Note: *WA = wage allowance*, *med* = median value, and *p25* = 25^{th} percentile.

Median and percentile values are calculated within the data set.

Table A.3: Agricultural sustainability items with category definitions and frequencies

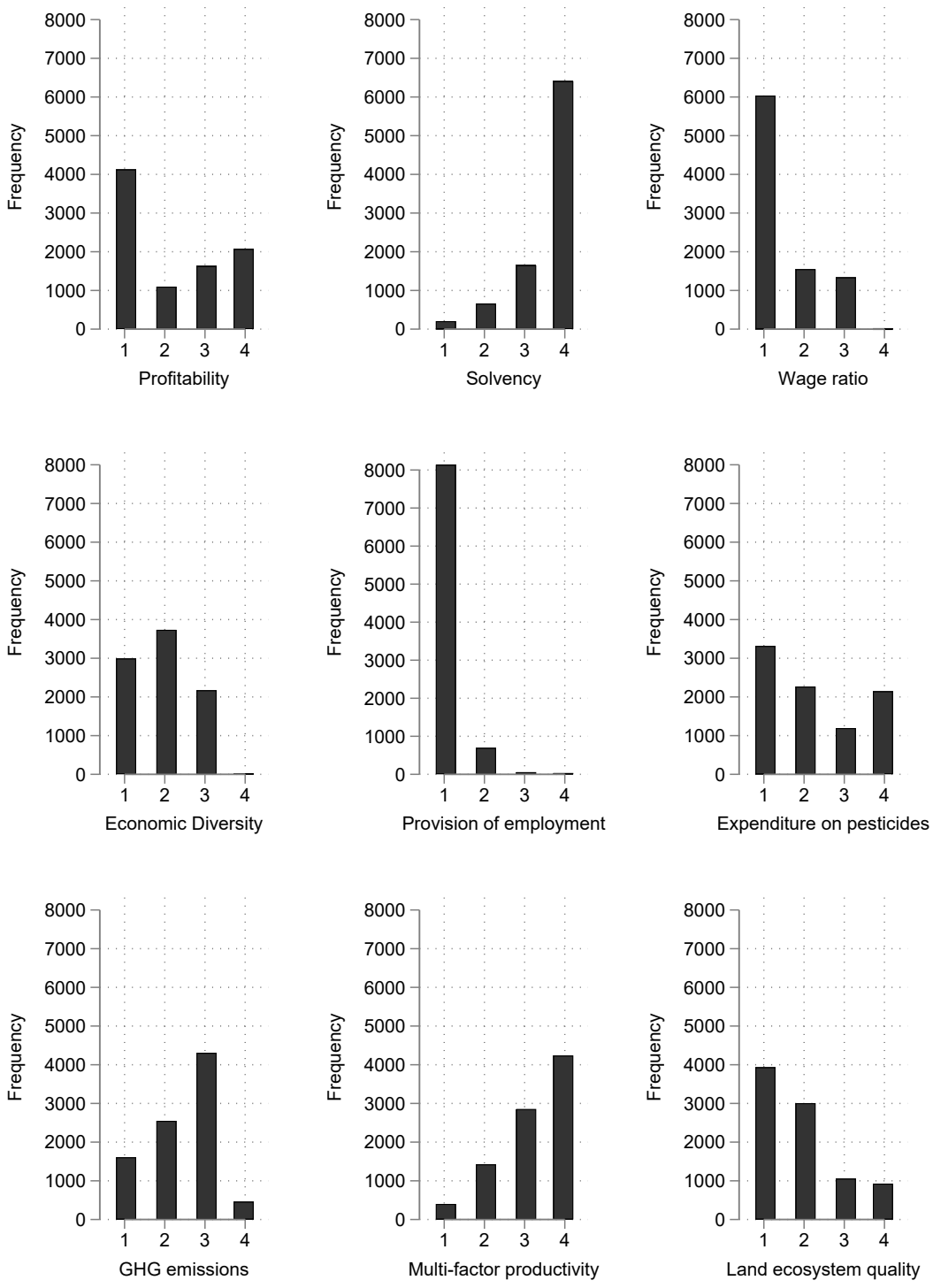

Figure A.1: Frequencies for each category of the nine sustainability items

A.3 Results

	ΔELPD	s.e.(ΔELPD)	ELPD	s.e.(ELPD)
Model with discrimination	0.0	0.0	-83991.7	192.8
Model without discrimination	-3148.7	77.1	-87140.4	188.8

Note: ELPD is the expected log pointwise predictive density. Δ ELPD is the difference in ELPDs. s.e. denotes the standard error.

Table A.4: Model comparison using LOO-CV

j	θ	$P(y_{2j} = \text{“}very\ sustainable''$)	$P(y_{5j} = \text{“}very\ sustainable''$)
1	1.573	0.7712	0.00013
2	1.263	0.7605	0.000105
3	-0.5826	0.6896	0.000032

Table A.5: Parameter estimates for three sample farms

Farm type	P(Y = very unsustainable)	P(Y = unsustainable)	P(Y = sustainable)	P(Y = very sustainable)
Mixed	0.2373	0.2725	0.3335	0.1568
	(0.0661)	(0.0545)	(0.0421)	(0.0430)
Granivores	0.3523	0.2937	0.2554	0.0986
	(0.0654)	(0.0550)	(0.0425)	(0.0444)
Other grazing livestock	0.3180	0.2922	0.2776	0.1122
	(0.0640)	(0.0551)	(0.0439)	(0.0476)
Milk	0.3396	0.2935	0.2635	0.1034
	(0.0649)	(0.0552)	(0.0429)	(0.0456)
Other permanent crops	0.2437	0.2748	0.3290	0.1525
	(0.0603)	(0.0507)	(0.0507)	(0.0555)
Wine	0.3682	0.2932	0.2455	0.0931
	(0.0661)	(0.0545)	(0.0421)	(0.0430)
Horticulture	0.3014	0.2900	0.2888	0.1198
	(0.0634)	(0.0548)	(0.0452)	(0.0493)
Fieldcrops	0.2231	0.2661	0.3434	0.1673
	(0.0579)	(0.0483)	(0.0530)	(0.0576)

Table A.6: Posterior means of the predicted probabilities for each sustainability category with standard errors in parentheses, by farm type.

Farm economic size class	P(Y = very unsustainable)	P(Y = unsustainable)	P(Y = sustainable)	P(Y = very sustainable)
25,000-50,000	0.2979	0.3003	0.2879	0.1140
	(0.0711)	(0.0536)	(0.0498)	(0.0510)
50,000-100,000	0.3073	0.3015	0.2815	0.1097
	(0.0718)	(0.0536)	(0.0495)	(0.0498)
100,000-250,000	0.2884	0.2987	0.2945	0.1183
	(0.0700)	(0.0534)	(0.0501)	(0.0520)
250,000-500,000	0.2732	0.2955	0.3053	0.1261
	(0.0685)	(0.0531)	(0.0509)	(0.0538)
500,000-750,000	0.2491	0.2885	0.3226	0.1399
	(0.0661)	(0.0523)	(0.0524)	(0.0570)
750,000-1,000,000	0.2105	0.2717	0.3504	0.1674
	(0.0616)	(0.0503)	(0.0555)	(0.0627)
1,000,000-1,500,000	0.1875	0.2581	0.3664	0.1880
	(0.0583)	(0.0487)	(0.0579)	(0.0661)
1,500,000-3,000,000	0.1434	0.2230	0.3922	0.2415
	(0.0510)	(0.0443)	(0.0628)	(0.0733)
>3,000,000	0.1272	0.2067	0.3981	0.2680
	(0.0477)	(0.0427)	(0.0644)	(0.0760)

Table A.7: Posterior means of the predicted probabilities for each sustainability category with standard errors in parentheses, by farm size.

Region	Est	Error	Q10	Q90
Stuttgart	0.1207	0.0508	0.0582	0.1879
Karlsruhe	0.1278	0.0521	0.0625	0.1976
Freiburg	0.1263	0.0523	0.0613	0.1973
Tübingen	0.1246	0.0515	0.0615	0.1934
Oberbayern	0.1224	0.0512	0.0589	0.1909
Niederbayern	0.1139	0.0492	0.0536	0.1798
Oberpfalz	0.1274	0.0521	0.0618	0.1967
Oberfranken	0.1167	0.0498	0.0555	0.1842
Mittelfranken	0.1268	0.052	0.0616	0.1958
Unterfranken	0.134	0.0535	0.0668	0.206
Schwaben	0.1085	0.0478	0.0496	0.1718
Brandenburg	0.2043	0.0642	0.1198	0.2874
Hamburg	0.0883	0.0431	0.0379	0.1465
Darmstadt	0.1273	0.0522	0.0626	0.1965
Gießen	0.1177	0.0502	0.0558	0.1852
Kassel	0.1095	0.0482	0.0507	0.1748
Mecklenburg-Vorpommern	0.1862	0.0621	0.1062	0.2662
Braunschweig	0.1521	0.0569	0.0799	0.2277
Hannover	0.1313	0.053	0.0653	0.2025
Lüneburg	0.1307	0.0528	0.0646	0.2013
Weser-Ems	0.1017	0.0463	0.0458	0.163
Düsseldorf	0.1183	0.0503	0.0563	0.1859
Köln	0.1223	0.0513	0.0596	0.1904
Münster	0.0955	0.0447	0.0421	0.1554
Detmold	0.1199	0.0506	0.0574	0.1878
Arnsberg	0.1163	0.0497	0.0543	0.1826
Koblenz	0.1158	0.0497	0.0538	0.1825
Trier	0.1004	0.046	0.0447	0.1624

Region	Est	Error	Q10	Q90
Rheinhessen-Pfalz	0.1151	0.0494	0.0548	0.1813
Saarland	0.118	0.0506	0.0564	0.1849
Dresden	0.1696	0.0597	0.0913	0.2474
Chemnitz	0.1639	0.0588	0.0882	0.2431
Leipzig	0.197	0.0638	0.1125	0.2804
Sachsen-Anhalt	0.1942	0.0631	0.1114	0.2757
Schleswig-Holstein	0.1236	0.0513	0.0599	0.1923
Thüringen	0.1984	0.0635	0.1155	0.2792

Table A.8: Regional averages of the predicted probability for a random farm achieving the "very sustainable" category, with standard errors and scores for top and bottom 10% of the sample.

	Percent of farms with missing		
Number of missing items	10	30	50
1	0.9918	0.9806	0.9691
2	0.9868	0.9646	0.9449
3	0.9823	0.9500	0.9178

Table A.9: Correlation coefficients for missing item tests

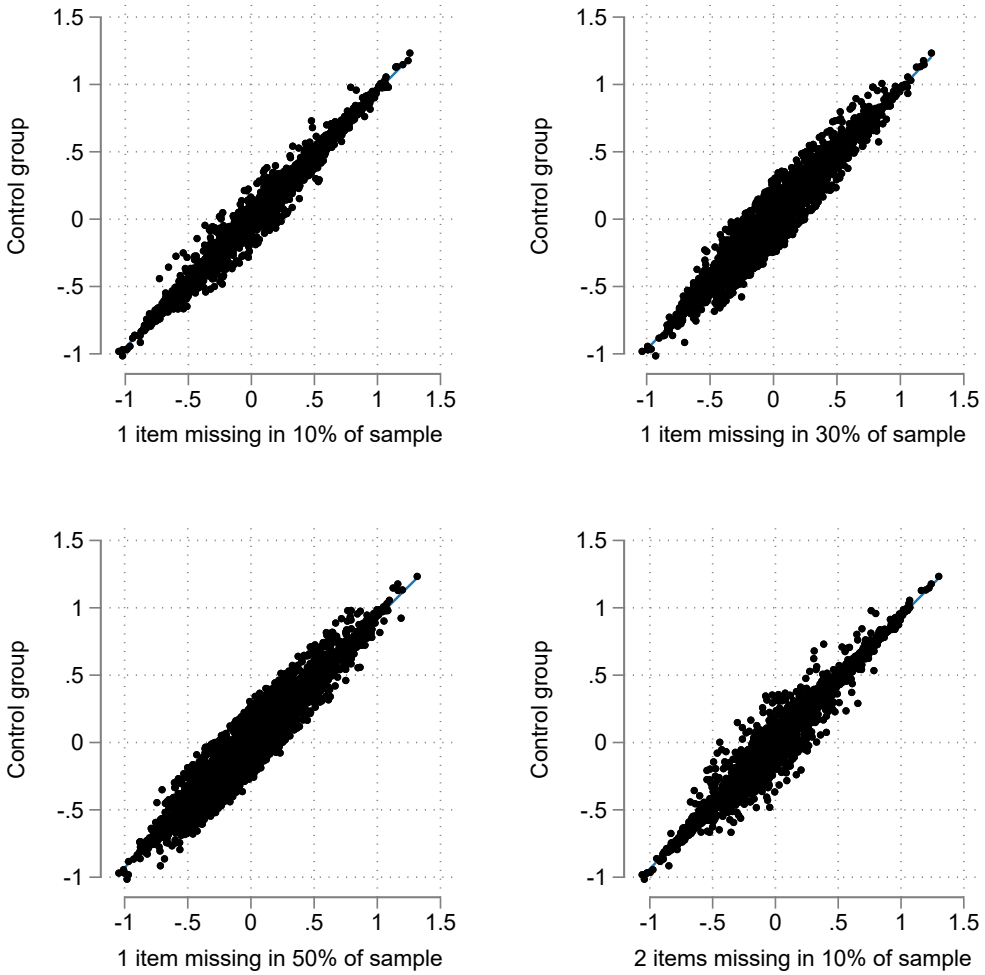

Figure A.2: Scatter plots for missing item tests (1/2)

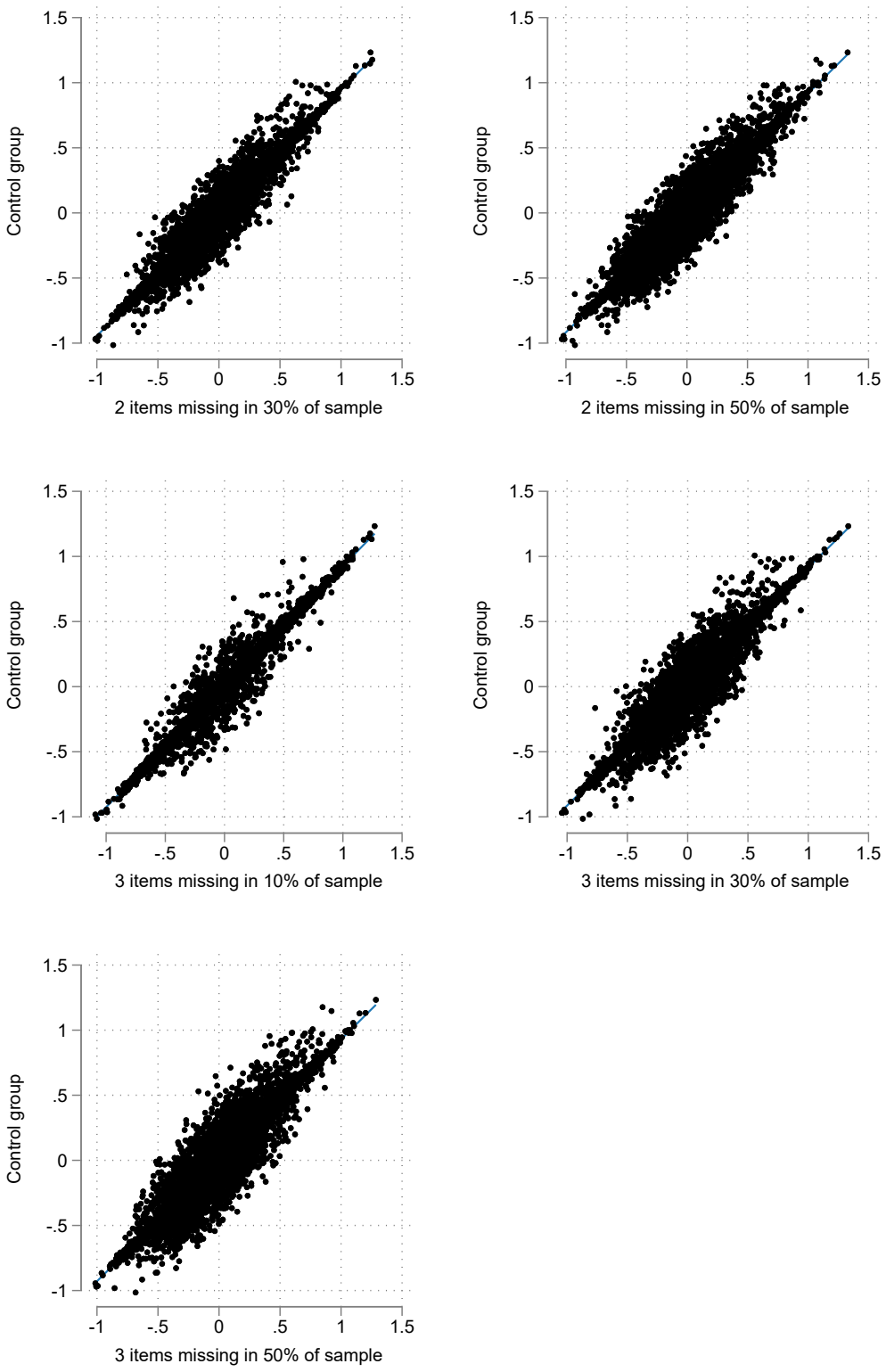

Figure A.3: Scatter plots for missing item tests (2/2)

Item missing	Item No.	1	2	3	4	5	6	7	8	9
Profitability	1	X								
Solvency	2	0.9742	X							
Wage ratio	3	0.9474	0.9538	X						
Economic diversity	4	0.8633	0.8683	0.8362	X					
Provision of employment	5	0.9691	0.9732	0.9440	0.9429	X				
Expenditure on pesticides	6	0.9784	0.9741	0.9629	0.9529	0.9625	X			
GHG emissions	7	0.8865	0.8864	0.8580	0.8572	0.8644	0.8855	X		
Multi-factor productivity	8	0.9524	0.9445	0.9300	0.9240	0.9257	0.9469	0.9354	X	
Land ecosystem quality	9	0.9593	0.9523	0.9477	0.9439	0.9430	0.9545	0.9528	0.9552	X

Table A.10: Correlation coefficients for concurrent scale linking simulations between East (columns) and West (rows) Germany. All data sets are correlated with AS scores estimated with the full data set.

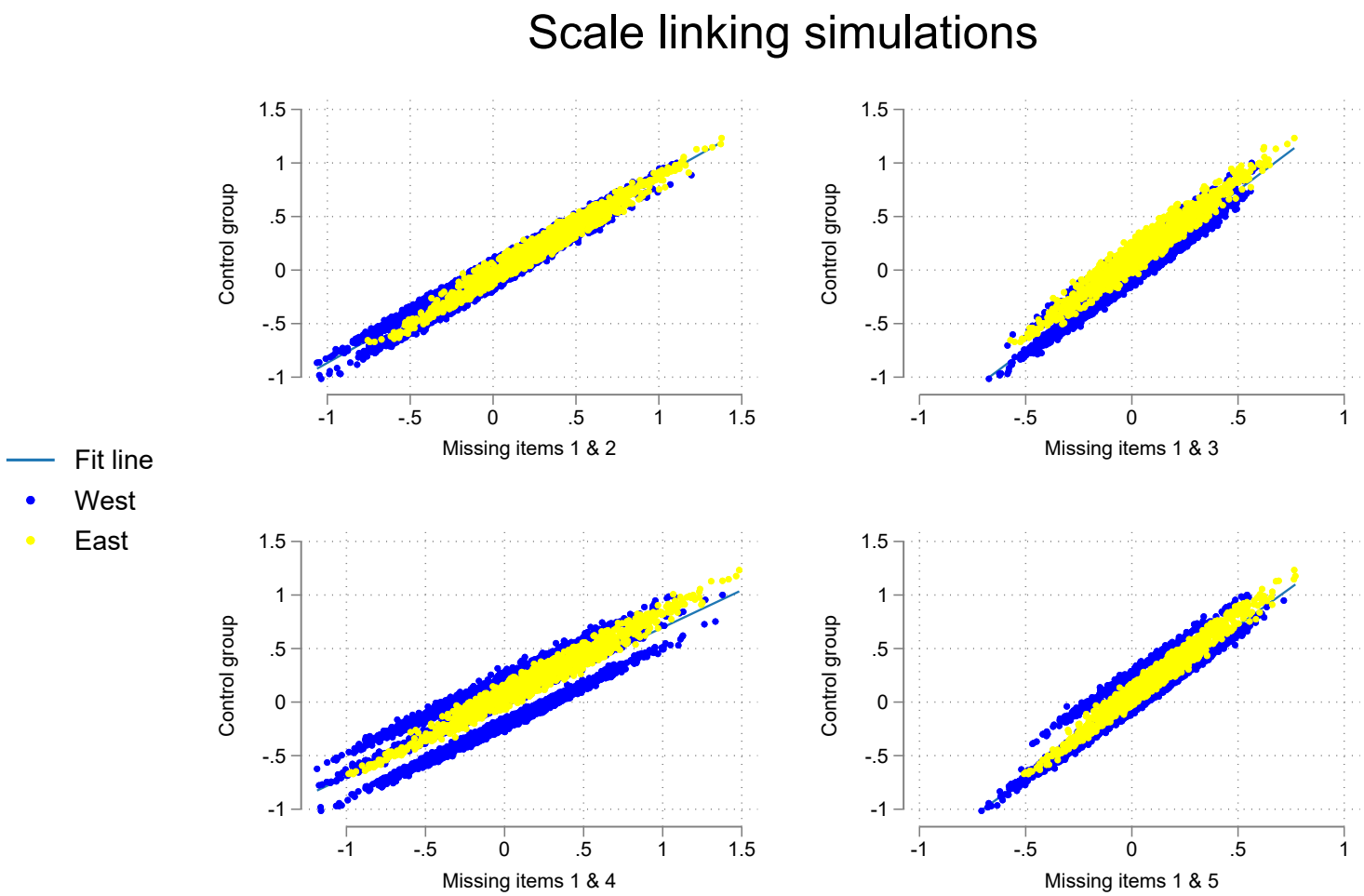

Figure A.4: Scatter plots for concurrent scale linking simulations between East and West Germany (1/9)

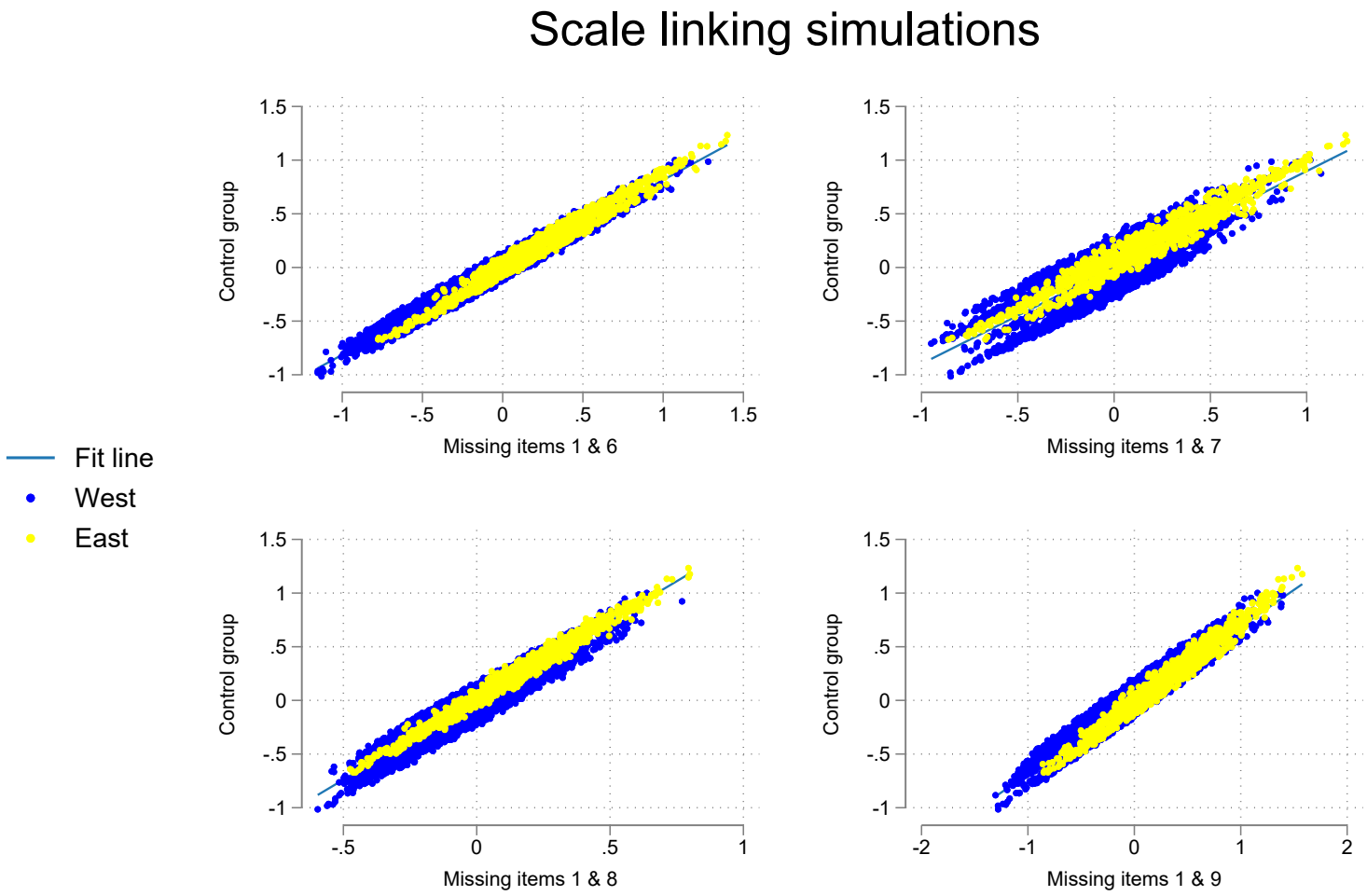

Figure A.5: Scatter plots for concurrent scale linking simulations between East and West Germany (2/9)

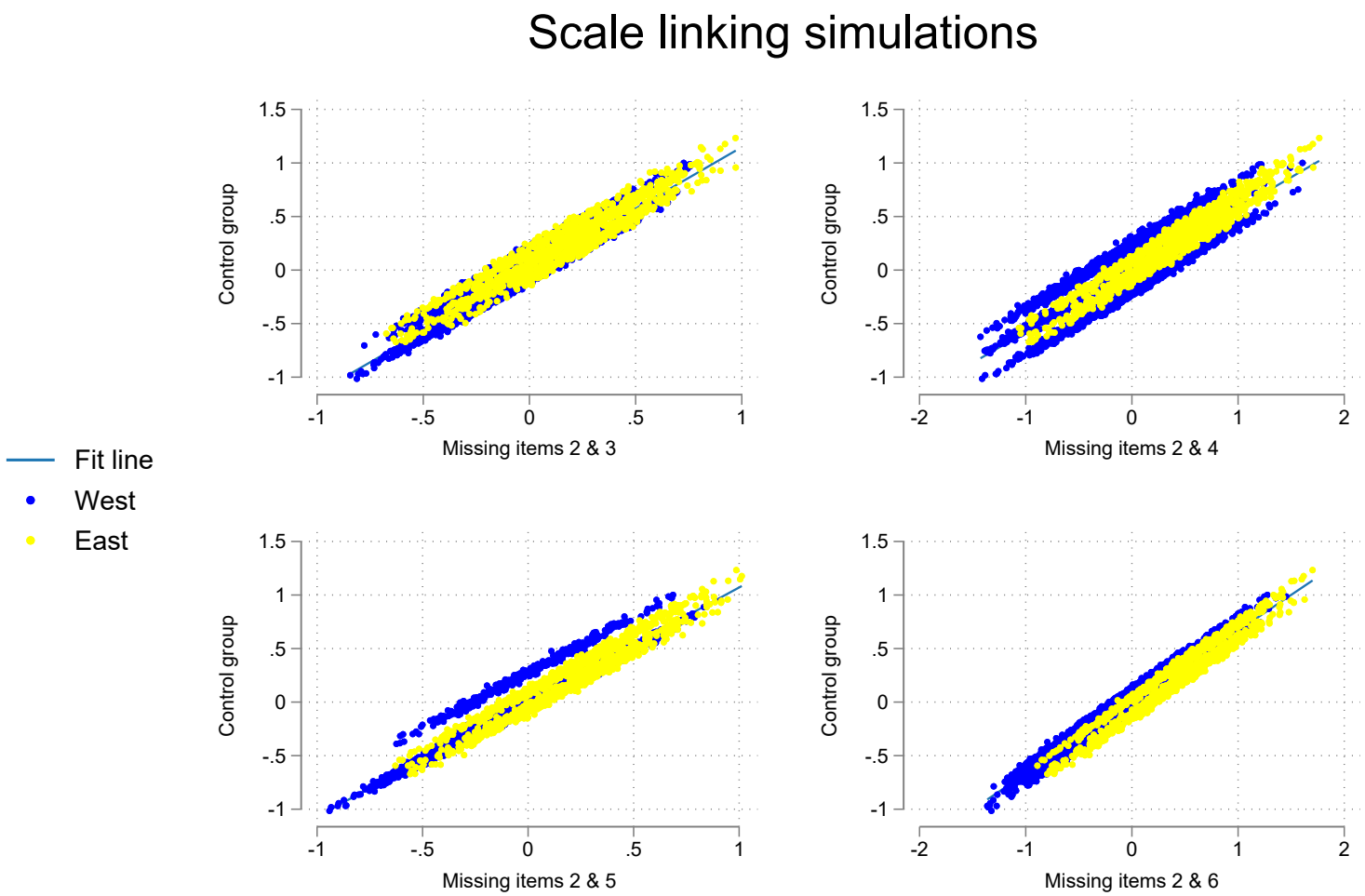

Figure A.6: Scatter plots for concurrent scale linking simulations between East and West Germany (3/9)

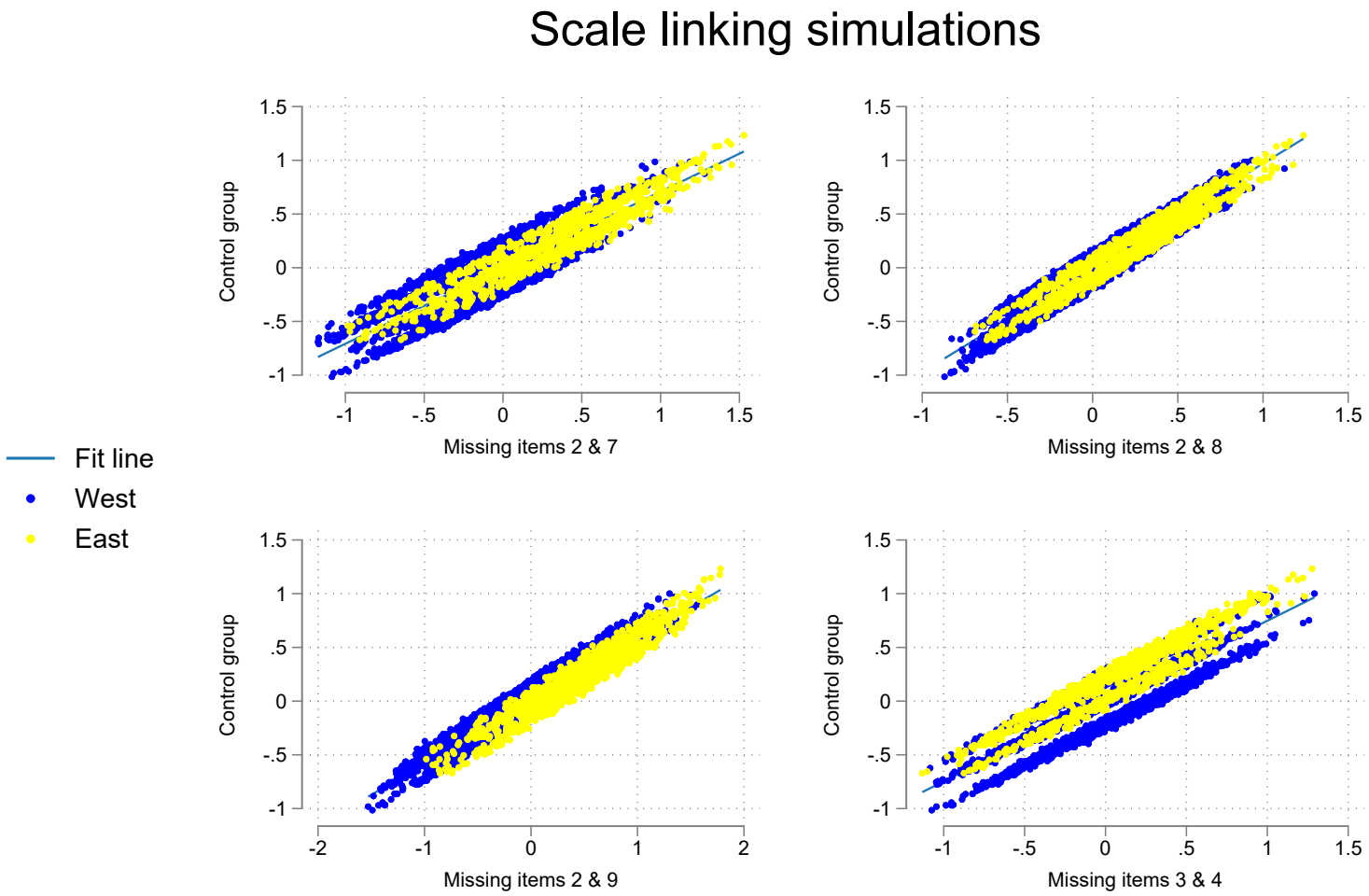

Figure A.7: Scatter plots for concurrent scale linking simulations between East and West Germany (4/9)

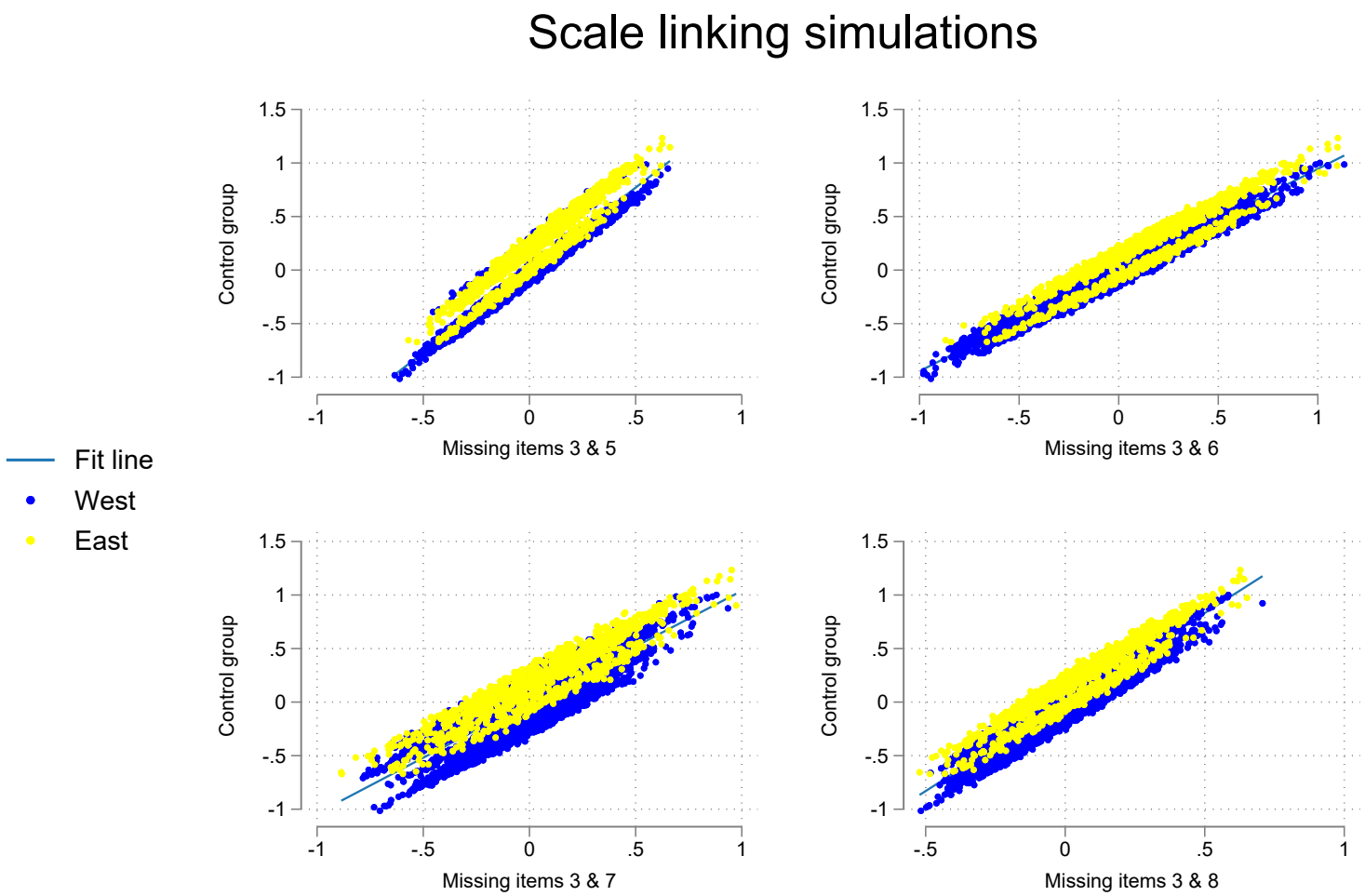

Figure A.8: Scatter plots for concurrent scale linking simulations between East and West Germany (5/9)

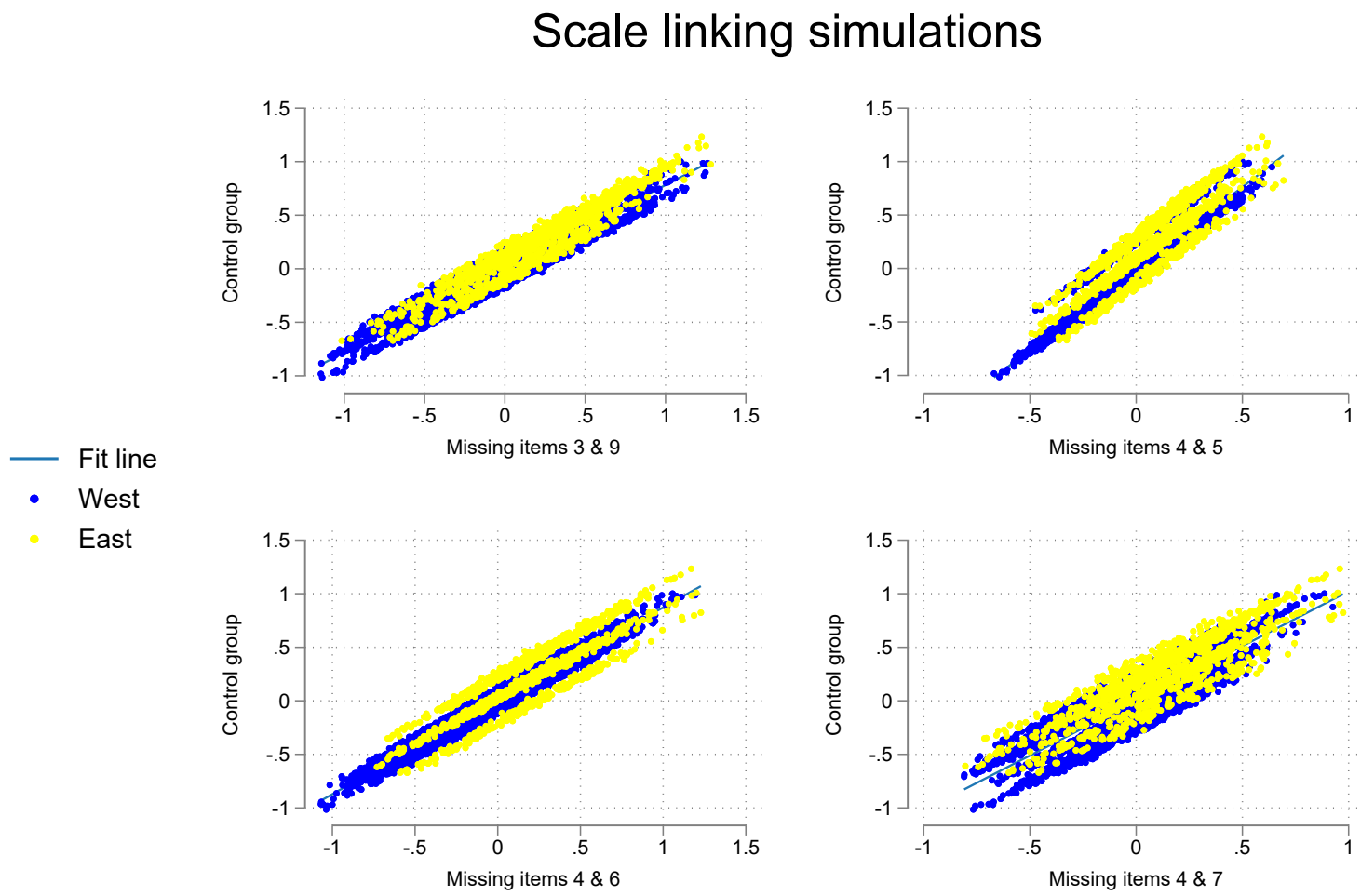

Figure A.9: Scatter plots for concurrent scale linking simulations between East and West Germany (6/9)

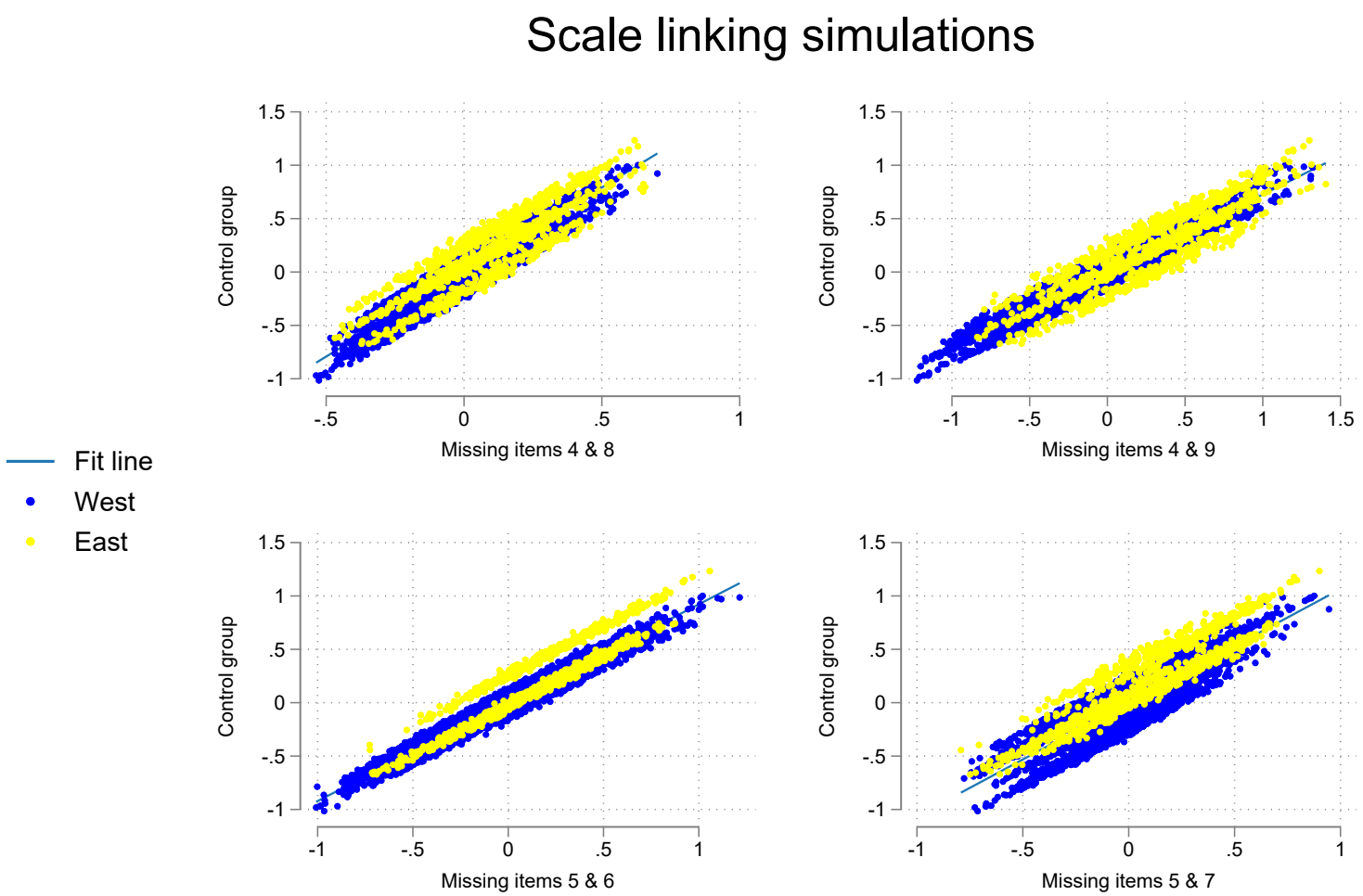

Figure A.10: Scatter plots for concurrent scale linking simulations between East and West Germany (7/9)

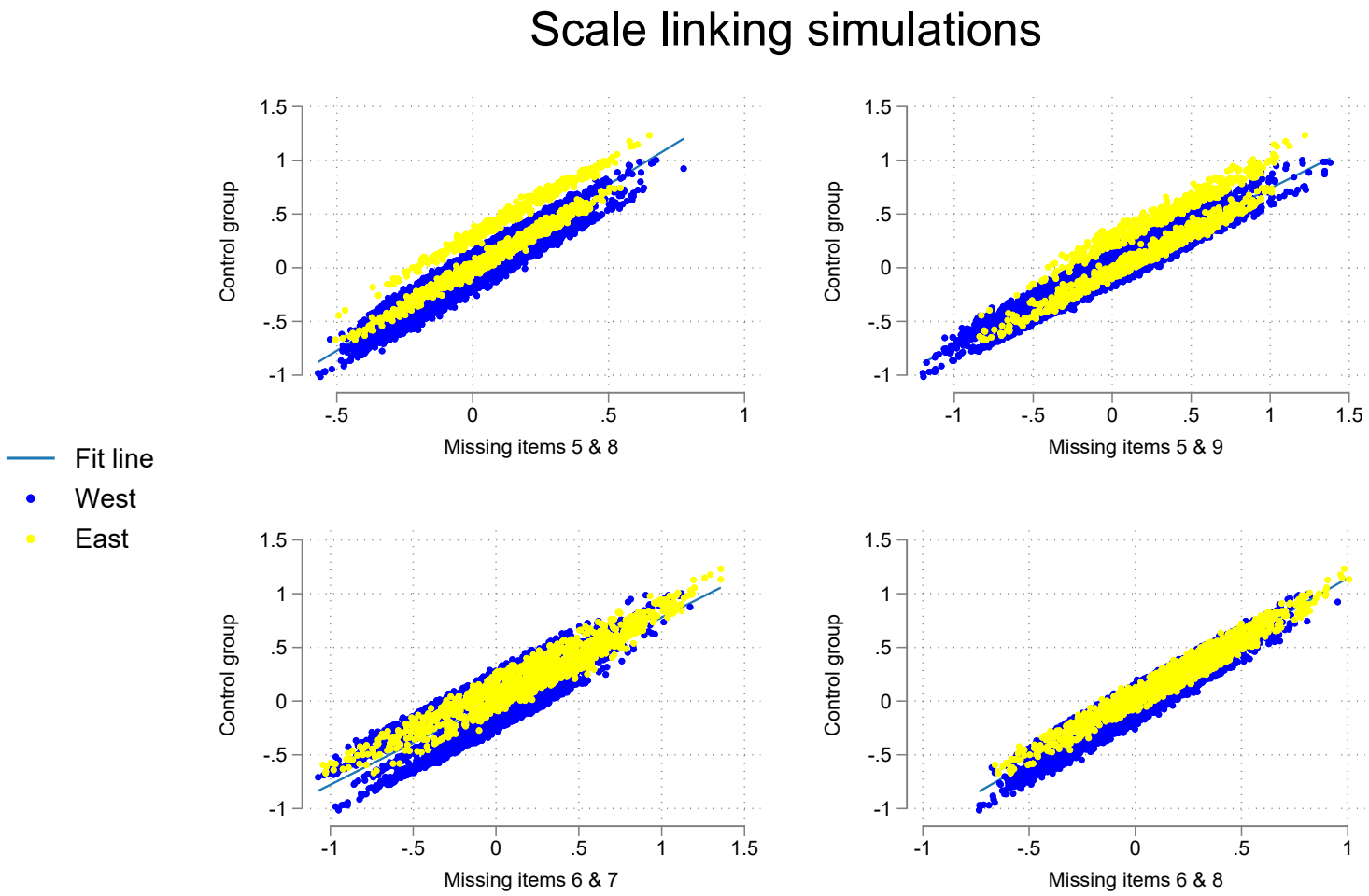

Figure A.11: Scatter plots for concurrent scale linking simulations between East and West Germany (8/9)

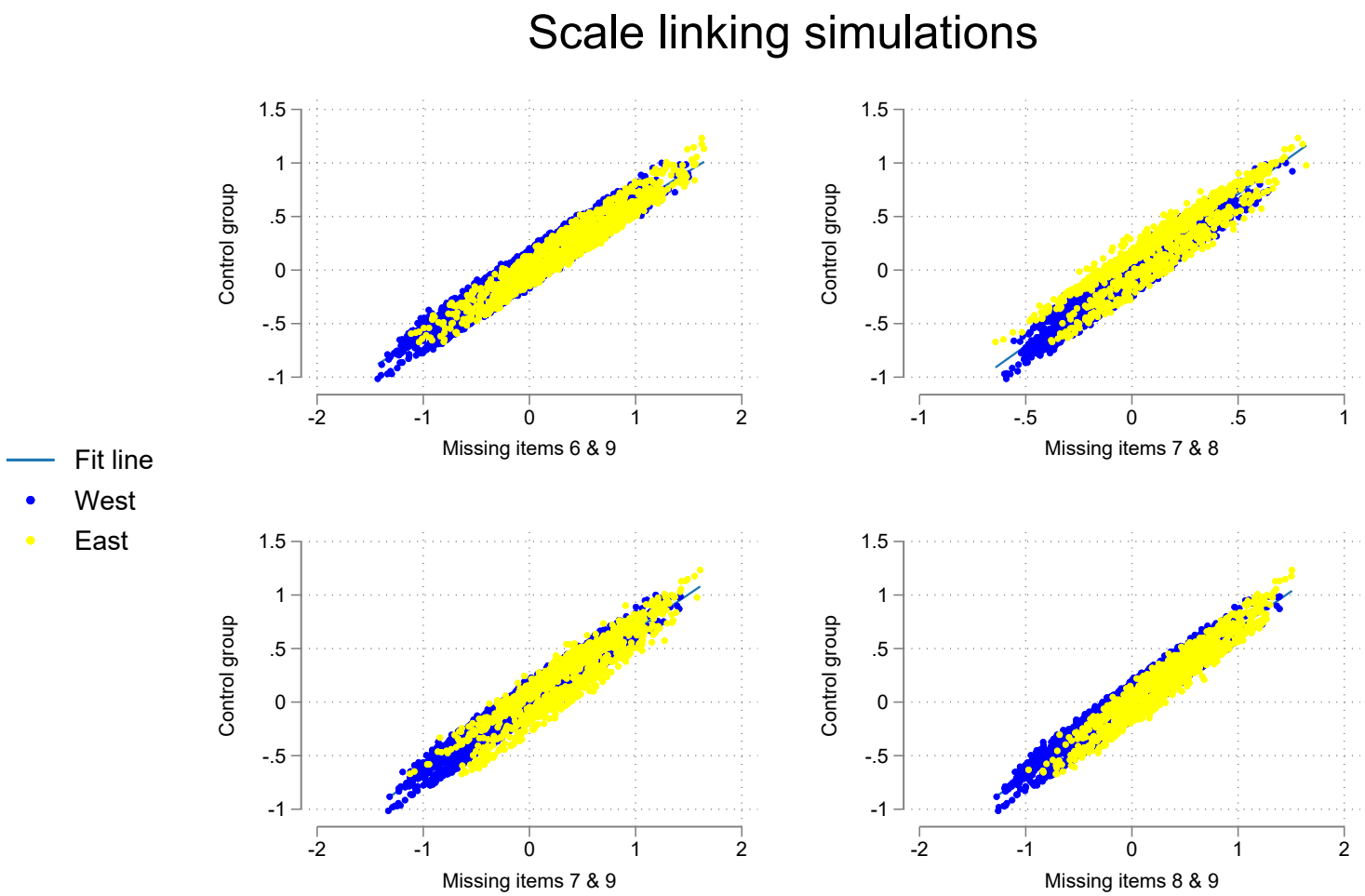

Figure A.12: Scatter plots for concurrent scale linking simulations between East and West Germany (9/9)

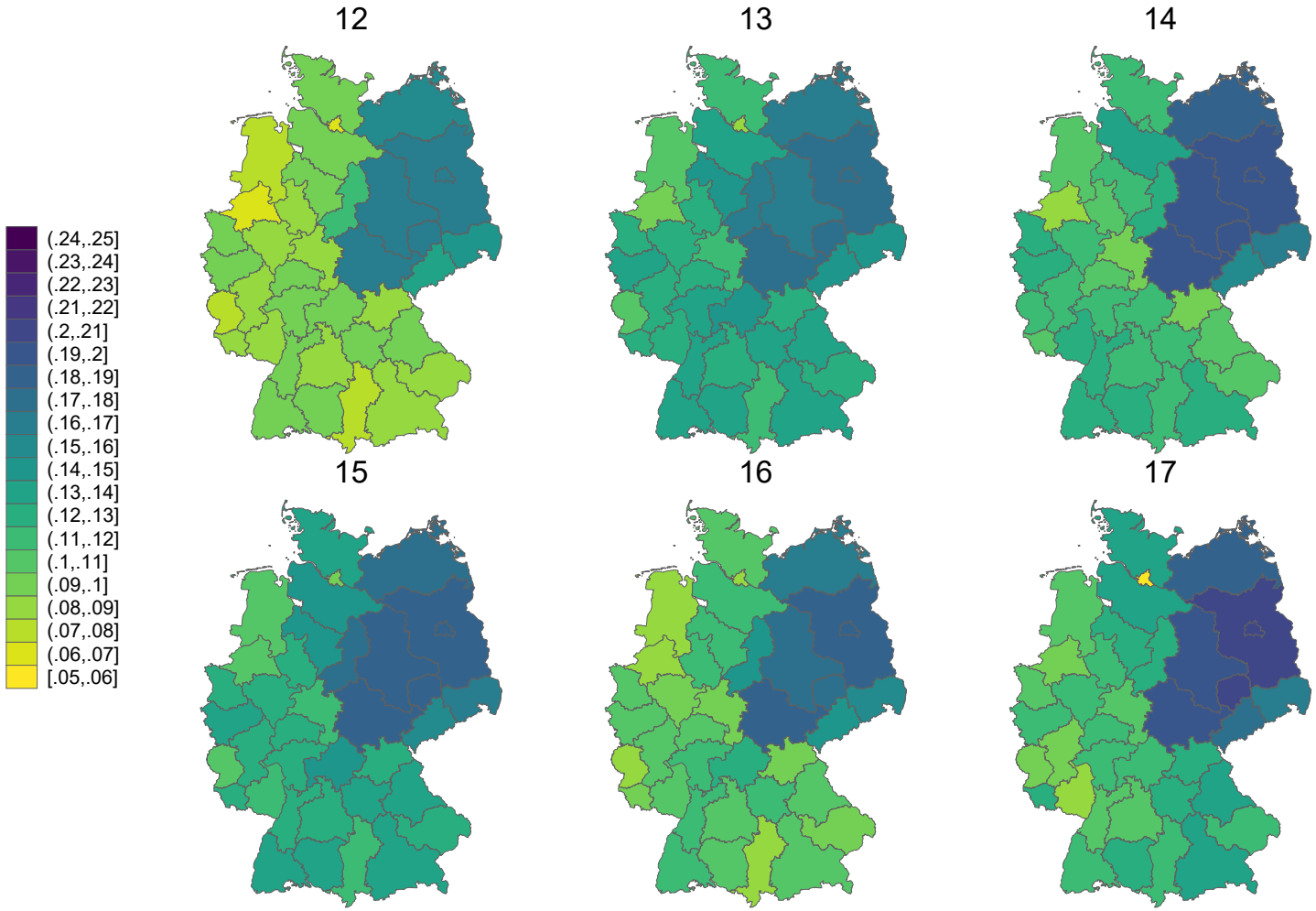

Figure A.13: Maps of scale linking tests by NUTS 2 region (1/6). The two digit number refers to the items missing from the samples. As an example, map number 12 means that item 1 (profitability) is missing in East Germany and item 2 (solvency) is missing in West Germany.

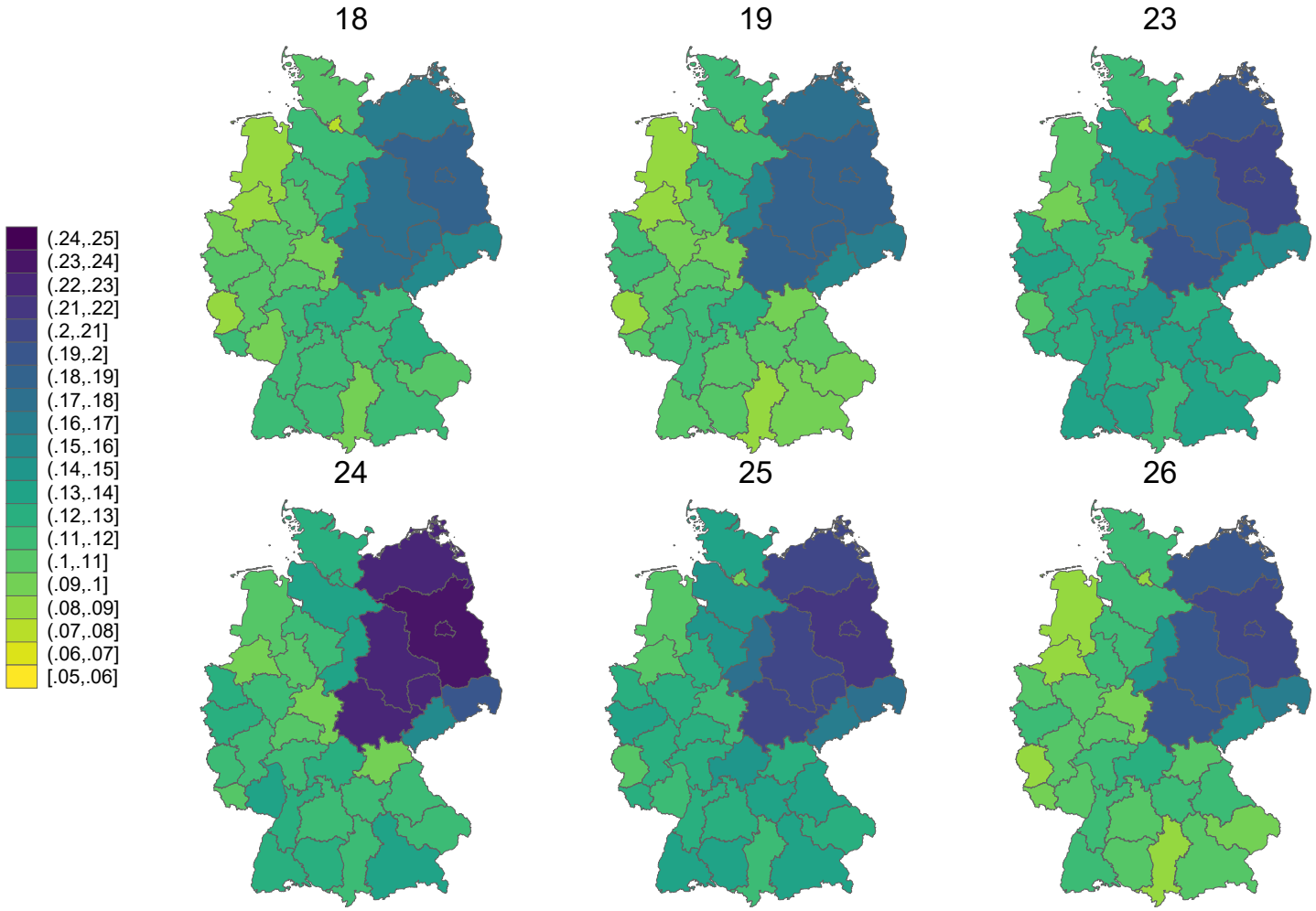

Figure A.14: Maps of scale linking tests by NUTS 2 region (2/6). The two digit number refers to the items missing from the samples. As an example, map number 18 means that item 1 (profitability) is missing in East Germany and item 8 (multi-factor productivity) is missing in the West Germany.

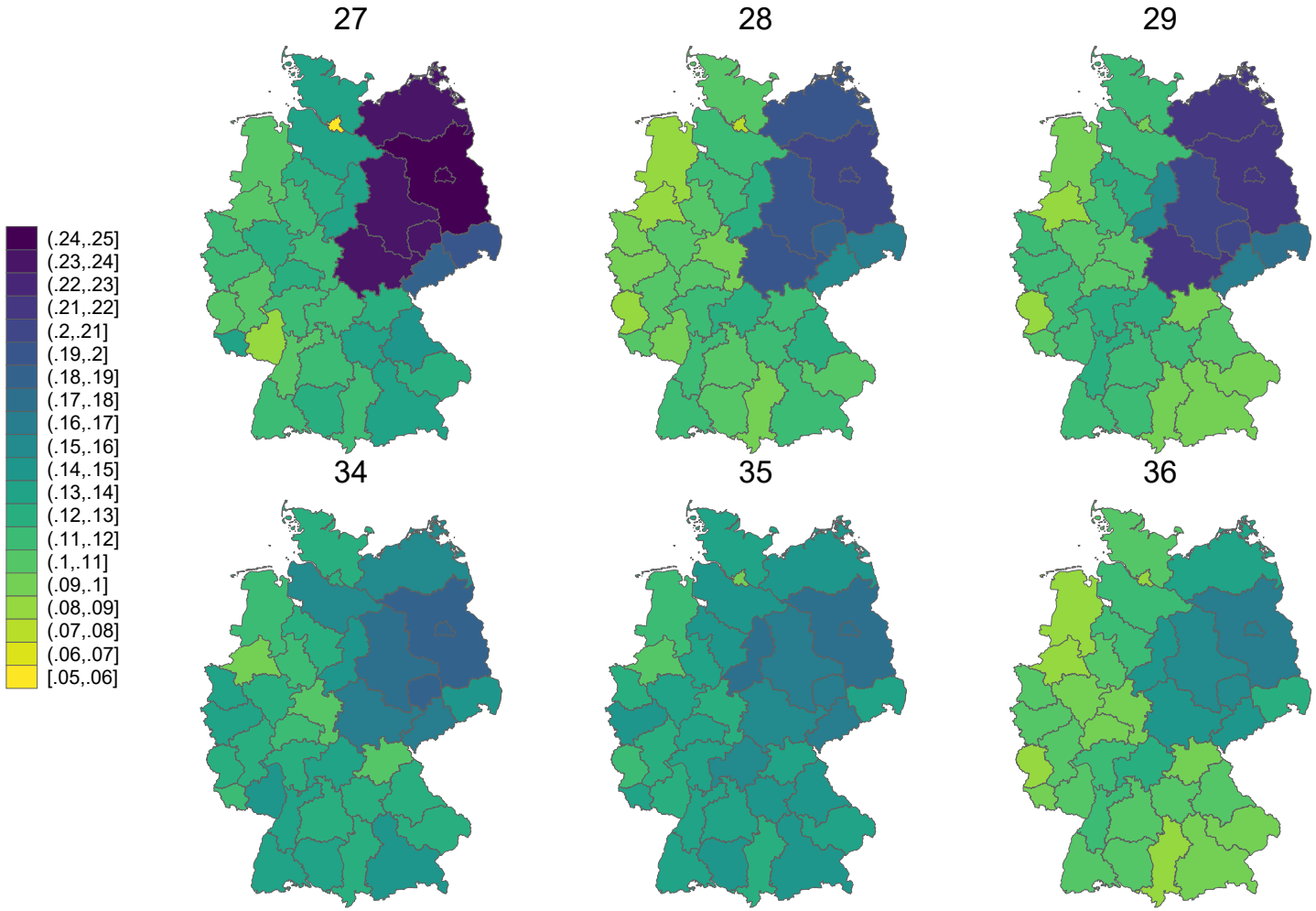

Figure A.15: Maps of scale linking tests by NUTS 2 region (3/6). The two digit number refers to the items missing from the samples. As an example, map number 27 means that item 2 (solvency) is missing in East Germany and item 7 (GHG emissions) is missing in West Germany.

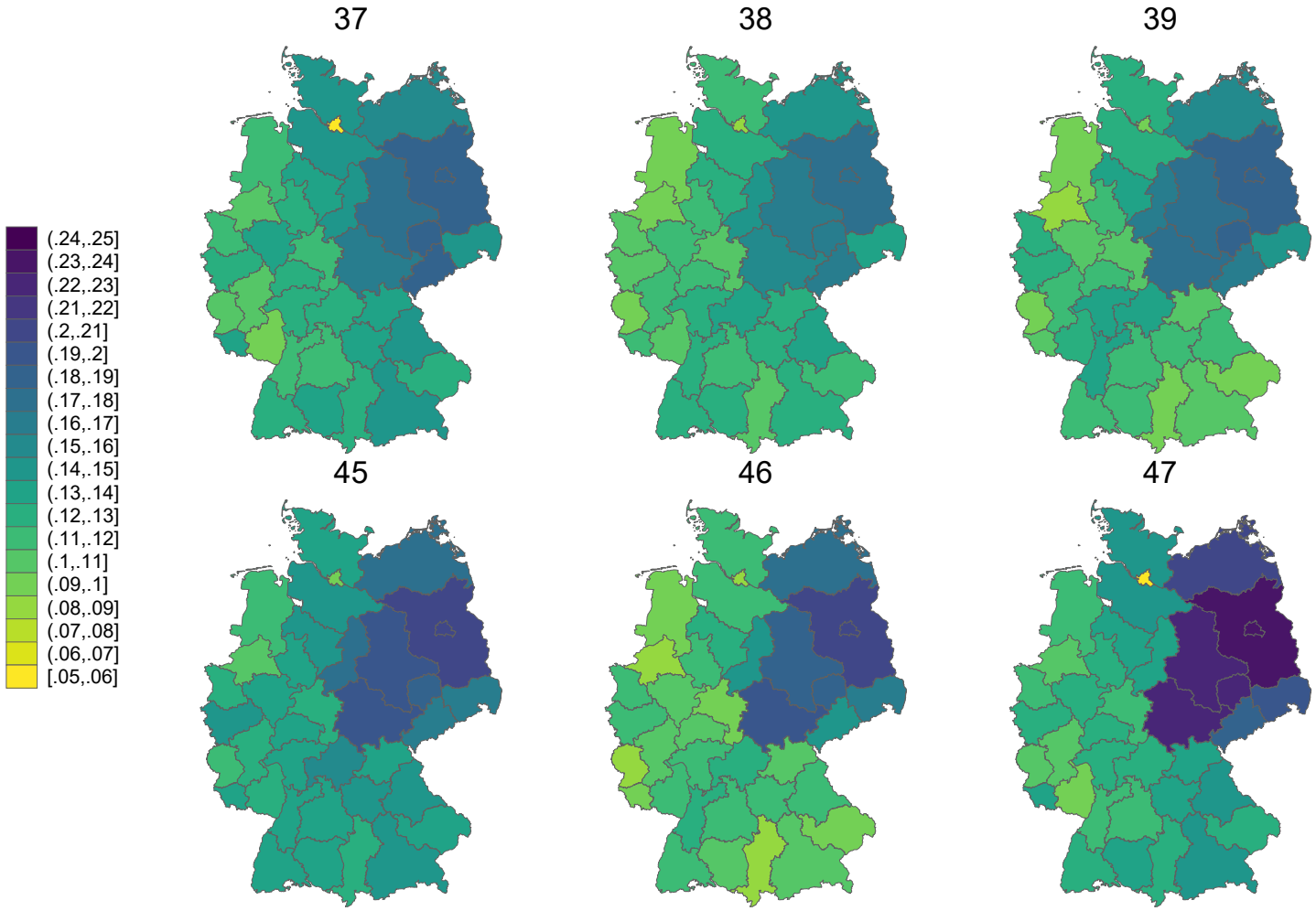

Figure A.16: Maps of scale linking tests by NUTS 2 region (4/6). The two digit number refers to the items missing from the samples. As an example, map number 37 means that item 3 (wage ratio) is missing in East Germany and item 7 (GHG emissions) is missing in West Germany.

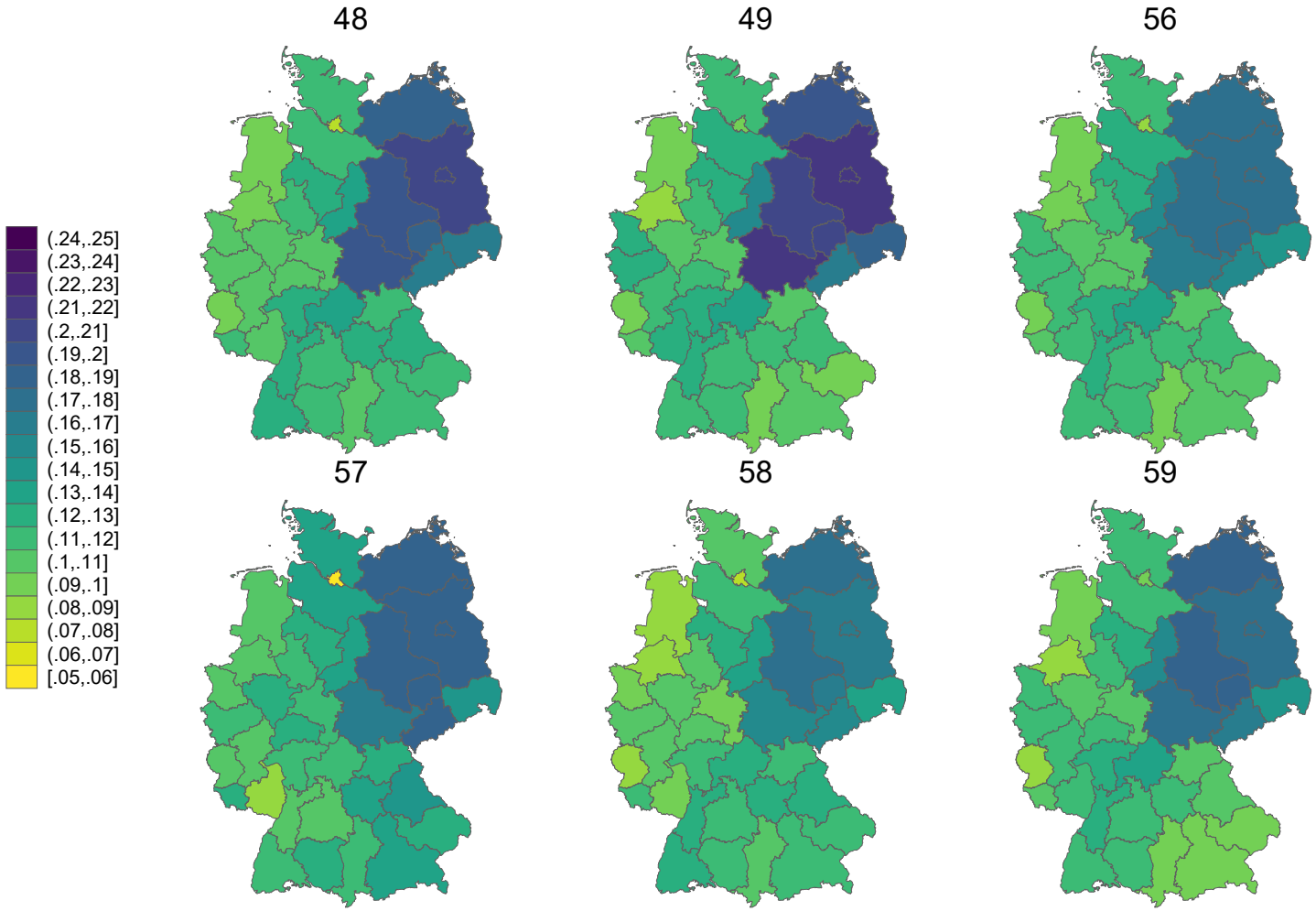

Figure A.17: Maps of scale linking tests by NUTS 2 region (5/6). The two digit number refers to the items missing from the samples. As an example, map number 48 means that item 4 (economic diversity) is missing in East Germany and item 8 (multi-factor productivity) is missing in West Germany.

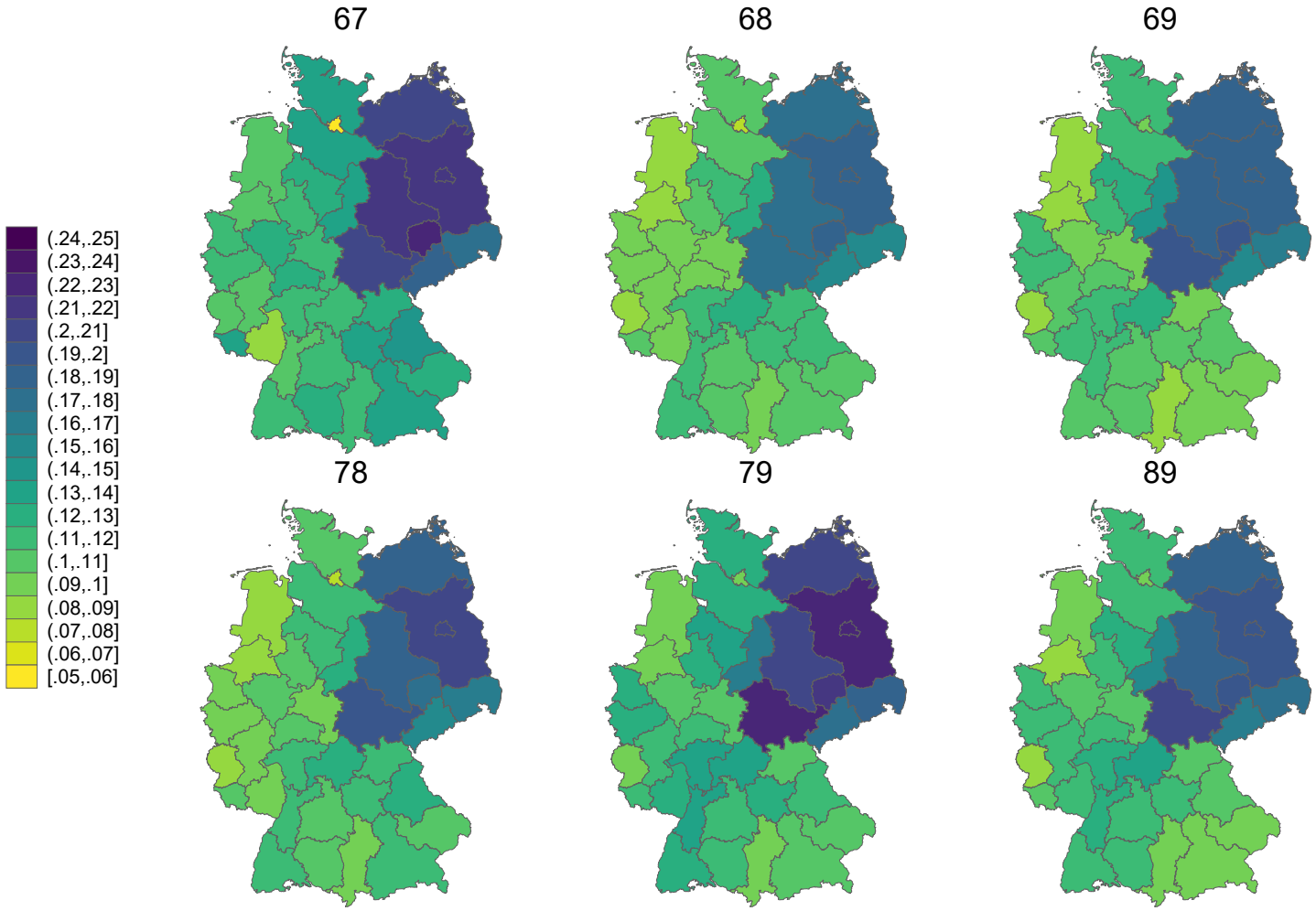

Figure A.18: Maps of scale linking tests by NUTS 2 region (6/6). The two digit number refers to the items missing from the samples. As an example, map number 67 means that item 6 (expenditure on pesticides) is missing in East Germany and item 7 (GHG emissions) is missing in West Germany.

Appendix B

B.1 Descriptive statistics

	Category	Mean	SD	Min	Max	N=
Not on marginal land ($ml = 0$)						
No NFC production	A	-	-	-	-	5010
Energy crops, % of total output	B	11.62	14.63	0.00	92.49	464
Industrial crops, % of total output	C	35.89	32.01	0.01	97.21	94
On marginal land ($ml = 1$)						
No NFC production	D	-	-	-	-	3130
Energy crops, % of total output	E	11.80	15.09	0.02	76.99	178
Industrial crop, % of total output	F	21.53	27.41	0.04	94.47	51

Table B.1: Descriptive statistics of NFC production variables by marginal land classification.

B.2 Results

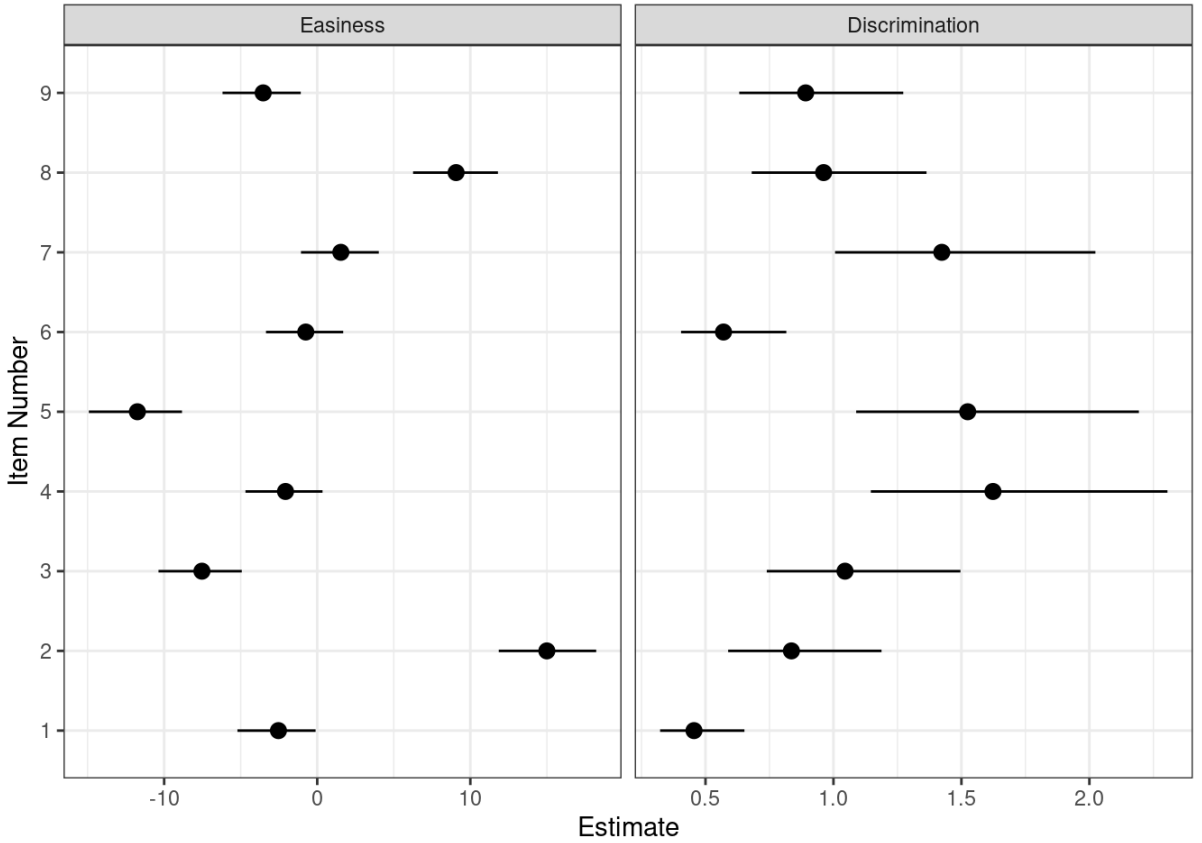

Figure B.1: Posterior means and 95% credible intervals for easiness and discrimination parameters.

	Category	Mean	SD	95% lower	95% upper
Not on marginal land ($ml = 0$)					
Energy crop output	B	1.89	0.19	1.53	2.27
Industrial crop output	C	1.93	0.31	1.32	2.54
On marginal land ($ml = 1$)					
No NFC production	D	0.12	0.07	-0.02	0.27
Energy crop output	E	1.32	0.24	0.87	1.80
Industrial crop output	F	1.68	0.41	0.89	2.47

Table B.2: Parametric regression results for the farm types

Appendix C

C.1 Methods

	Year Y						
Farm	1	2	3	4	5	Group	$t = 0$
1	C	P/T	P/T	P/T	O	Starter (S)	Year 2
2	C	C	P/T	O	P/T	Starter (S)	Year 3
3	C	C	C	C	C	Always conventional (AC)	-
4	C	C	P/T	C	P/T	Always conventional (AC)	-
5	O	O	O	O	O	Always organic (AO)	-
6	O	O	P/T	O	P/T	Always organic (AO)	-
7	O	O	O	C	C	Quitter (Q)	-
8	C	P/T	O	O	C	Quitter (Q)	-
9	C	C	C	O	O	Disqualified (DQ)	-

Table C.1: Examples of farm groupings based on different combinations of organic classifications over time.

		Group			
	Category	S	AC	AO	Total
Farm type (TF8)	Fieldcrops	5	2158	88	2251
	Horticulture	2	558	16	576
	Wine	7	438	8	453
	Other permanent crops	1	221	9	231
	Milk	21	2057	108	2186
	Other grazing livestock	13	852	69	934
	Granivores	2	1413	18	1433
	Mixed	8	2568	113	2689
Economic size class	25,000 - <50,000	8	935	70	1013
	50,000 - <100,000	20	2352	148	2520
	100,000 - <250,000	34	4351	167	4552
	250,000 - <500,000	19	2936	88	3043
	500,000 - <750,000	6	998	35	1039
	750,000 - <1,000,000	0	414	8	422
	1,000,000 - <1,500,000	0	372	11	383
	1,500,000 - <3,000,000	1	425	6	432
	>= 3,000,000	1	238	3	242
	Total	49	7798	323	**8170**

Note: Frequencies are reported for the data set with a minimum five obs. per farm.

Table C.2: Frequencies of organic classifications by farm type and economic size class

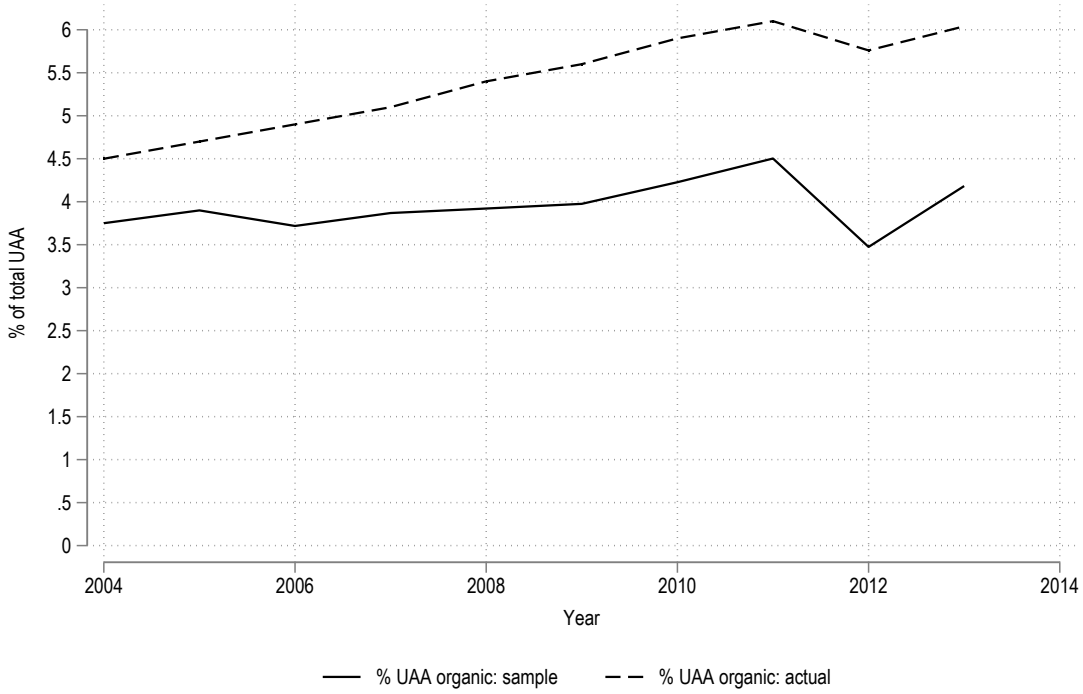

Figure C.1: UAA under organic production, percent of total. Data source: Eurostat (2022c)

No.	Item	Description	(+/-)	Unit	Mean	SD
1	Profitability	Farm net income less an allowance for unpaid labor based on the regional median agricultural wage	+	€	13123	104732
2	Solvency	Ratio of total debts to total assets	-	Ratio	0.2378	0.3453
3	Economic diversity	The maximum percentage of a single agricultural product to total output	-	Ratio	0.6454	0.1898
4	Expenditure on pesticides	Ratio of total pesticide expenditure to total utilized agricultural area (UAA)	-	€/ha	233.4	882.1
5	GHG emissions	CO_2 intensity on the farm: ratio of annual CO_2 equivalent gases emitted/absorbed to gross value added	-	Tonnes/€	0.4234	2.2437
6	Land ecosystem quality	Estimated quality of land as a percent relative to pristine (untouched) natural landscape	+	Percent	0.1680	0.1300
7	Wage ratio	Ratio of average wages on the farm to the median wage in the region	+	Ratio	0.3239	0.3794
8	Provision of employment	Ratio of total expenditure on wages and contract work to total output of the farm	+	Ratio	0.1052	0.5542
9	Multi-factor productivity	Ratio of total value added to factor inputs for land, labor, and capital	+	Ratio	0.8867	0.5799

Table C.3: Agricultural sustainability items and descriptions. The (+/-) refers to the relationship between the value of the variable and its effect on farm sustainability. Adapted from Table A.3

C.2 IRT model specification

In this section, we specify the model for the time-series AS used in Chapter 4. The model is an extension of the cross-sectional models used in Chapters 2 and 3. In the time-series model, the outcome of the ith sustainability item at time t for farm j, y_{itj}, is measured with C categories representing the ratings of sustainability on an ordered scale. We use the same ordinal items with the same $C = 4$ categories as in the previous chapters.

In the time-series model, the probability of a particular category is

$$P(y_{itj} = c|\boldsymbol{\tau}, \psi_{itj}) = F(\tau_c - \psi_{itj}) - F(\tau_{c-1} - \psi_{itj}), \tag{C.1}$$

where F denotes the CDF of the standard logistic distribution and $\boldsymbol{\tau}$ is a vector of $C - 1$ unknown thresholds.[1] The distributional parameter ψ_{itj} can be expressed as a function of farm parameters, θ_{tj}, and item parameters, ξ_i:

$$\psi_{itj} = \theta_{tj} + \xi_i. \tag{C.2}$$

The farm parameter θ_{tj} represents the latent construct of agricultural sustainability for farm j at time t (i.e. the AS index). The larger the value of the AS index is, the larger the probability that the farm is classified as "very sustainable" for each of the items.

The AS index is allowed to vary over time, as we observe our nine sustainability items repeatedly during the sampling period. We specify a flexible model for the

[1] The threshold parameters τ_1, τ_2, and τ_3 are freely estimated whereas τ_0 and τ_4 are set to $-\infty$ and $+\infty$, respectively.

possibly nonlinear trajectories of the AS index by including a set of time dummies:

$$\theta_{tj} = \gamma_{1j} + \gamma_{2j} d2_t + ... + \gamma_{Tj} dT_t, \tag{C.3}$$

where γ_{1j} is a farm-specific intercept, γ_{2j} through γ_{Tj} are farm-specific time effects, and $d2_t$ through dT_t represent a set of $T-1$ time dummies. The item parameter ξ_i in equation C.2 represents the easiness of item i. For items with a high value of ξ_i, each farm has a higher probability of being classified as "very sustainable".

The specification of the IRT model in equation C.2 relies on the unrealistic assumptions that the effect of farm-specific agricultural sustainability on each item probability is constant. To relax this assumption, we introduce an item-specific discrimination parameter, α_i, that reflects that some items can better differentiate among farms with different degrees of agricultural sustainability than others:

$$\psi_{itj} = \alpha_i(\theta_{tj} + \xi_i) = \alpha_i \theta_{tj} + \delta_i. \tag{C.4}$$

We fit the model in a Bayesian framework using the `brms` package (Bürkner 2017) in R (R Core Team 2021), which allows to interface with the probabilistic programming language Stan (Carpenter et al. 2017).[2] We use weakly informative prior distributions that help to improve convergence of the sampling algorithm while they do not have a strong influence on the posterior distribution because of the large amount of sample data available from the FADN. Following Bürkner (2019), we impose two constraints to ensure identification. First, we restrict the discrimination parameters α_i to be positive because a change of the sign of α_i can be offset by a change of the sign of $\theta_j + \xi_i$. This constraint is not overly restrictive because higher categories of the items represent always a higher degree of sustainability.[3] Second, we fix the

[2]The model was fit using R version 3.6.3, `brms` version 2.16.3, and Stan version 2.21.0.
[3]A negative sign of α_i would imply that a higher degree of AS is associated with a decrease in

standard deviations of the farm-specific parameters to 1, as the multiplicative relationship of α_i and θ_j does not allow to freely estimate the scale of the farm-specific parameters. That is, the scale of the farm-specific parameters is determined by the scale of the discrimination parameters.

Parameter	Prior distributions	Constraints
Thresholds: τ_1, τ_2	Student-t$(3, 0, 2.5)$	$\tau_1 < \tau_2$
Discrimination parameters: α_i	Normal$(\mu_\alpha, \sigma_\alpha)$	$\alpha_i > 0$
	$\mu_\alpha \sim$ Normal$(0, 1)$	
	$\sigma_\alpha \sim$	
	Half-Normal$(0, 1)$	
Item-specific effects: δ_i	Normal$(\mu_\delta, \sigma_\delta)$	
	$\mu_\delta \sim$ Normal$(0, 1)$	
	$\sigma_\delta \sim$	
	Half-Normal$(0, 1)$	
Farm-specific effects: $\gamma_{1j}, ..., \gamma_{Tj}$	Normal(μ_γ, Σ)	$\sigma_\gamma = 1$
	$\mu_\gamma \sim$ Normal$(0, 1)$	
	Σ is parameterized	
	using a Cholesky	
	factor with a LKJ(1)	
	prior.	

Table C.4: Prior distributions

the probability for the category "sustainable" and an increase in the probability for the category "unsustainable".

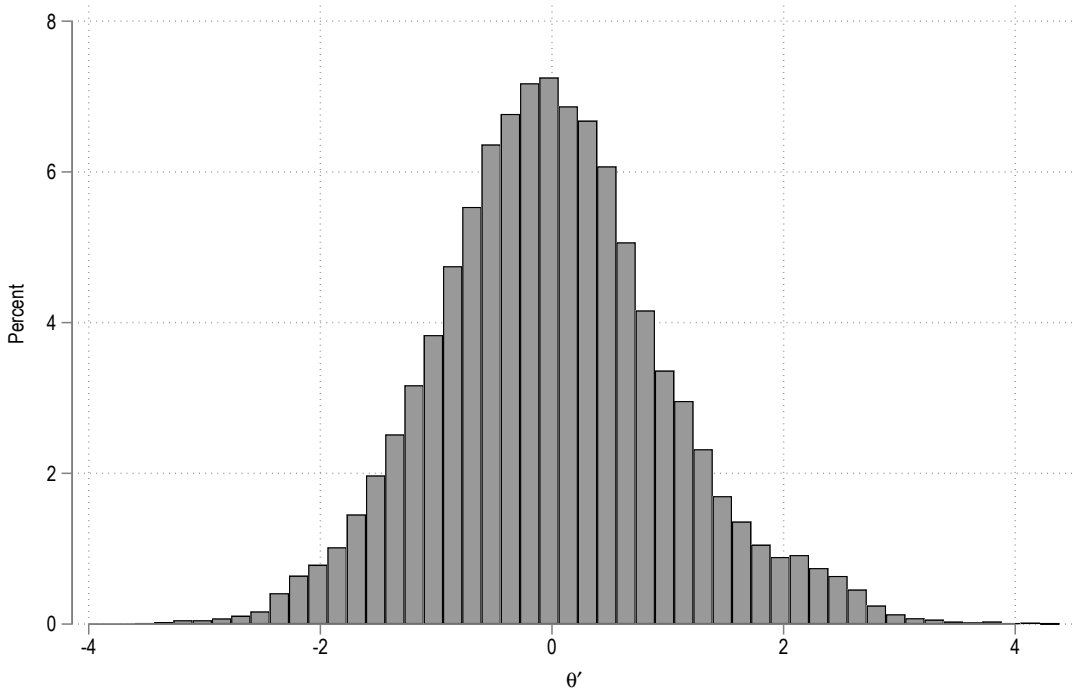

Figure C.2: Histogram of standardized AS index score

C.3 Descriptive analyses

Descriptive analysis for all observations, 2004

Group	Mean	s.e.	obs.
AC	0.0458	0.5180	5031
Pre-S	0.1248	0.5091	42
AO	0.5447	0.5619	189

Table C.5: Summary statistics for all observations, 2004

Source	SS	df	MS	F	Prob > f
Between groups	45.49	2	22.74	84.26	0.000
Within groups	1419.64	5259	0.2699		
Total	1465.13	5261	0.2785		

Table C.6: ANOVA results for all observations, 2004

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0786	0.0805	0.592	-0.1102	0.2673
AO vs AC	0.4989	0.0385	0.000	0.4086	0.5891
AO vs pre-S	0.4203	0.0886	0.000	0.2125	0.6281

Table C.7: Tukey pairwise comparisons of means with equal variances

Descriptive analysis for fieldcrops, 2004

Group	Mean	s.e.	obs.
AC	0.3836	0.3753	806
Pre-S	0.8066	0.2628	2
AO	0.8521	0.4413	40

Table C.8: Summary statistics of fieldcrop farms

Source	SS	df	MS	F	Prob > f
Between groups	8.68	2	4.3420	30.31	0.000
Within groups	121.05	845	0.1436		
Total	129.73	847	0.1532		

Table C.9: ANOVA results of fieldcrop farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.4230	0.2680	0.255	-0.2061	1.0521
AO vs AC	0.4685	0.0613	0.000	0.3245	0.6124
AO vs pre-S	0.0455	0.2742	0.985	-0.5984	0.6893

Table C.10: Tukey pairwise comparisons of means with equal variances of fieldcrop farms

Descriptive analysis for horticulture, 2004

Group	Mean	s.e.	obs.
AC	-0.1005	0.5679	442
Pre-S	-0.0642	0.0884	2
AO	0.4248	0.3565	4

Table C.11: Summary statistics of horticulture farms

Source	SS	df	MS	F	Prob > f
Between groups	1.10	2	0.5479	1.71	0.1821
Within groups	142.61	445	0.3205		
Total	143.70	447	0.3215		

Table C.12: ANOVA results of horticulture farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0363	0.4012	0.995	-0.9072	0.9797
AO vs AC	0.5253	0.2843	0.156	-0.1433	1.1939
AO vs pre-S	0.4890	0.4903	0.579	-0.6639	1.6419

Table C.13: Tukey pairwise comparisons of means with equal variances of horticulture farms

Descriptive analysis for wine, 2004

Group	Mean	s.e.	obs.
AC	-0.4075	0.3522	206
Pre-S	-0.3518	0.2067	3
AO	-0.3985	0.1275	4

Table C.14: Summary statistics of vineyards

Source	SS	df	MS	F	Prob > f
Between groups	0.01	2	0.0047	0.04	0.9620
Within groups	25.56	210	0.1217		
Total	25.57	210	0.1206		

Table C.15: ANOVA results of vineyards

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0557	0.2029	0.959	-0.4232	0.5346
AO vs AC	0.0090	0.1761	0.999	-0.4067	0.4247
AO vs pre-S	-0.0466	0.2665	0.983	-0.6756	0.5823

Table C.16: Tukey pairwise comparisons of means with equal variances of vineyards

Descriptive analysis for other permanent crops, 2004

Group	Mean	s.e.	obs.
AC	-0.0359	0.4169	115
Pre-S	-0.1951	-	1
AO	0.6295	0.2673	4

Table C.17: Summary statistics for other permanent crops

Source	SS	df	MS	F	Prob > f
Between groups	1.744	2	0.8721	5.09	0.0076
Within groups	20.03	117	0.1712		
Total	21.77	119	0.1830		

Table C.18: ANOVA results for other permanent crops

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	-0.1594	0.4155	0.922	-1.1457	0.8272
AO vs AC	0.6654	0.2104	0.006	0.1659	1.1650
AO vs pre-S	0.8247	0.4626	0.180	-0.2734	1.9228

Table C.19: Tukey pairwise comparisons of means with equal variances for other permanent crops

Descriptive analysis for milk, 2004

Group	Mean	s.e.	obs.
AC	-0.0762	0.4080	1338
Pre-S	0.0355	0.5179	20
AO	0.1837	0.4295	60

Table C.20: Summary statistics of milk farms

Source	SS	df	MS	F	Prob > f
Between groups	4.08	2	2.0399	12.10	0.0000
Within groups	238.50	1415	0.1685		
Total	242.58	1417	0.1712		

Table C.21: ANOVA results of milk farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.1117	0.0925	0.449	-0.1053	0.3287
AO vs AC	0.2599	0.0542	0.000	0.1328	0.3871
AO vs pre-S	0.1482	0.1060	0.342	-0.1005	0.3969

Table C.22: Tukey pairwise comparisons of means with equal variances of milk farms

Descriptive analysis for other grazing livestock, 2004

Group	Mean	s.e.	obs.
AC	0.0229	0.5545	314
Pre-S	0.0949	0.3886	8
AO	0.7231	0.6592	29

Table C.23: Summary statistics for other grazing livestock farms

Source	SS	df	MS	F	Prob > f
Between groups	13.02	2	6.5115	20.70	0.0000
Within groups	109.47	348	0.3146		
Total	122.50	350	0.3500		

Table C.24: ANOVA results for other grazing livestock farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0721	0.2008	0.931	-0.4006	0.5447
AO vs AC	0.7004	0.1089	0.000	0.4441	0.9566
AO vs pre-S	0.6283	0.2240	0.015	0.1011	1.1555

Table C.25: Tukey pairwise comparisons of means with equal variances for other grazing livestock farms

Descriptive analysis for granivores, 2004

Group	Mean	s.e.	obs.
AC	-0.2711	0.4207	666
Pre-S	0.2834	-	1
AO	0.3428	0.4967	7

Table C.26: Summary statistics of granivore farms

Source	SS	df	MS	F	Prob > f
Between groups	2.91	2	1.4554	8.19	0.0003
Within groups	119.19	671	0.1776		
Total	122.10	673	0.1814		

Table C.27: ANOVA results of granivore farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.5544	0.4218	0.388	-0.4364	1.5451
AO vs AC	0.6139	0.1601	0.000	0.2378	0.9901
AO vs pre-S	0.0596	0.4506	0.990	-0.9988	1.1179

Table C.28: Tukey pairwise comparisons of means with equal variances of granivore farms

Descriptive analysis for mixed farms, 2004

Group	Mean	s.e.	obs.
AC	0.2878	0.5240	1144
Pre-S	0.6473	0.4922	5
AO	0.7766	0.4451	41

Table C.29: Summary statistics of mixed farms

Source	SS	df	MS	F	Prob > f
Between groups	10.05	2	5.0241	18.48	0.000
Within groups	322.79	1187	0.2719		
Total	332.84	1189	0.2799		

Table C.30: ANOVA results of mixed farms

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.3597	0.2337	0.273	-0.1888	0.9081
AO vs AC	0.4890	0.0829	0.000	0.2945	0.6835
AO vs pre-S	0.1293	0.2470	0.860	-0.4504	0.7090

Table C.31: Tukey pairwise comparisons of means with equal variances of mixed farms

Descriptive analysis for size class 25,000 -<50,000, 2004

Group	Mean	s.e.	obs.
AC	-0.0466	0.4141	203
Pre-S	-0.2957	0.6846	2
AO	0.3242	0.4462	26

Table C.32: Summary statistics for size class 25,000 -<50,000

Source	SS	df	MS	F	Prob > f
Between groups	3.34	2	1.6681	9.49	0.0001
Within groups	40.09	228	0.1758		
Total	43.42	230	0.1888		

Table C.33: ANOVA results for size class 25,000 -<50,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	-0.2491	0.2980	0.681	-0.9521	0.4538
AO vs AC	0.3707	0.0873	0.000	0.1647	0.5768
AO vs pre-S	0.6199	0.3077	0.111	-0.1062	1.3458

Table C.34: Tukey pairwise comparisons of means with equal variances for size class 25,000 -<50,000

Descriptive analysis for size class 50,000 -<100,000, 2004

Group	Mean	s.e.	obs.
AC	-0.0709	0.4223	1006
Pre-S	0.0242	0.6737	13
AO	0.2484	0.4538	54

Table C.35: Summary statistics for size class 50,000 -<100,000

Source	SS	df	MS	F	Prob > f
Between groups	5.30	2	2.65	14.51	0.000
Within groups	195.62	1070	0.183		
Total	200.92	1070	0.1828		

Table C.36: ANOVA results for size class 50,000 -<100,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0951	0.1194	0.705	-0.1850	0.3753
AO vs AC	0.3194	0.0597	0.000	0.1792	0.4595
AO vs pre-S	0.2242	0.1321	0.207	-0.0859	0.5342

Table C.37: Tukey pairwise comparisons of means with equal variances for size class 50,000 -<100,000

Descriptive analysis for size class 100,000 -<250,000, 2004

Group	Mean	s.e.	obs.
AC	-0.0277	0.4581	2102
Pre-S	0.1701	0.3571	15
AO	0.5922	0.4347	61

Table C.38: Summary statistics for size class 100,000 -<250,000

Source	SS	df	MS	F	Prob > f
Between groups	23.27	2	11.6327	55.72	0.000
Within groups	454.10	2175	0.2088		
Total	477.35	2177	0.2193		

Table C.39: ANOVA results for size class 100,000 -<250,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.1978	0.1184	0.217	-0.0799	0.4755
AO vs AC	0.6199	0.0593	0.000	0.4807	0.7591
AO vs pre-S	0.4221	0.1317	0.004	0.1133	0.7309

Table C.40: Tukey pairwise comparisons of means with equal variances for size class 100,000 -<250,000

Descriptive analysis for size class 250,000 -<500,000, 2004

Group	Mean	s.e.	obs.
AC	-0.0138	0.4712	1034
Pre-S	0.2867	0.3899	8
AO	0.7055	0.5853	27

Table C.41: Summary statistics for size class 250,000 -<500,000

Source	SS	df	MS	F	Prob > f
Between groups	14.25	2	7.1226	31.73	0.000
Within groups	239.30	1066	0.2245		
Total	239.30	1066	0.2245		

Table C.42: ANOVA results for size class 250,000 -<500,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.3005	0.1682	0.174	-0.0942	0.6952
AO vs AC	0.7193	0.0924	0.000	0.5025	0.9360
AO vs pre-S	0.4188	0.1907	0.072	-0.0289	0.8664

Table C.43: Tukey pairwise comparisons of means with equal variances for size class 250,000 -<500,000

Descriptive analysis for size class 500,000 -<750,000, 2004

Group	Mean	s.e.	obs.
AC	0.1978	0.4802	241
Pre-S	-0.1120	0.3300	3
AO	1.1512	0.7259	12

Table C.44: Summary statistics for size class 500,000 -<750,000

Source	SS	df	MS	F	Prob > f
Between groups	10.76	2	5.3812	22.53	0.000
Within groups	60.42	253	0.2388		
Total	71.18	255	0.2792		

Table C.45: ANOVA results for size class 500,000 -<750,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	-0.3098	0.2840	0.520	-0.9792	0.3595
AO vs AC	0.9533	0.1445	0.000	0.6126	1.2941
AO vs pre-S	1.2632	0.3155	0.000	0.5194	2.0069

Table C.46: Tukey pairwise comparisons of means with equal variances for size class 500,000 -<750,000

Descriptive analysis for size class 750,000 -<1,000,000, 2004

Group	Mean	s.e.	obs.
AC	0.3271	0.5814	65
AO	1.4969	0.2248	2

Table C.47: Summary statistics for size class 750,000 -<1,000,000

Source	SS	df	MS	F	Prob > f
Between groups	2.65	1	2.6548	7.96	0.0063
Within groups	21.68	65	0.3336		
Total	24.34	66	0.3688		

Table C.48: ANOVA results for size class 750,000 -<1,000,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
AO vs AC	1.1697	0.4146	0.006	0.3416	1.9978

Table C.49: Tukey pairwise comparisons of means with equal variances for size class 750,000 -<1,000,000

Descriptive analysis for size class 1,000,000 -<1,500,000, 2004

Group	Mean	s.e.	obs.
AC	0.5689	0.6306	103
AO	1.3739	0.2703	4

Table C.50: Summary statistics for size class 1,000,000 -<1,500,000

Source	SS	df	MS	F	Prob > f
Between groups	2.50	1	2.4952	6.42	0.0127
Within groups	40.78	105	0.3884		
Total	43.27	106	0.4082		

Table C.51: ANOVA results for size class 1,000,000 -<1,500,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
AO vs AC	0.8050	0.3176	0.013	0.1753	1.4347

Table C.52: Tukey pairwise comparisons of means with equal variances for size class 1,000,000 -<1,500,000

Descriptive analysis for size class 1,500,000 -<3,000,000, 2004

Group	Mean	s.e.	obs.
AC	0.8925	0.5227	178
AO	1.1660	-	1

Table C.53: Summary statistics for size class 1,500,000 -<3,000,000

Source	SS	df	MS	F	Prob > f
Between groups	0.07	1	0.0744	0.27	0.6025
Within groups	48.36	177	0.2732		
Total	48.43	178	0.2721		

Table C.54: ANOVA results for size class 1,500,000 -<3,000,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
AO vs AC	0.2735	0.5242	0.602	-0.7609	1.3079

Table C.55: Tukey pairwise comparisons of means with equal variances for size class 1,500,000 -<3,000,000

Descriptive analysis for size class >3,000,000, 2004

Group	Mean	s.e.	obs.
AC	0.9830	0.4113	99
Pre-S	0.9910	-	1
AO	1.2302	0.0940	2

Table C.56: Summary statistics for size class >3,000,000

Source	SS	df	MS	F	Prob > f
Between groups	0.1198	2	0.0599	0.36	0.7004
Within groups	16.59	99	0.1676		
Total	16.71	101	0.1654		

Table C.57: ANOVA results for size class >3,000,000

Group	Contrast	s.e.	p-value	95% lower	95% upper
pre-S vs AC	0.0080	0.4114	1.000	-0.9710	0.9869
AO vs AC	0.2472	0.2924	0.676	-0.4485	0.9426
AO vs pre-S	0.2392	0.5014	0.882	-0.9538	1.4321

Table C.58: Tukey pairwise comparisons of means with equal variances for size class >3,000,000

C.4 DiD model results

Time period	ATT	95% lower	95% upper
Pre-treatment	0.018	-0.122	0.157
Post-treatment	0.125	-0.125	0.375
-6	0.197	-0.748	1.141
-5	-0.190	-0.732	0.352
-4	0.006	-0.385	0.398
-3	0.073	-0.286	0.431
-2	-0.116	-0.447	0.215
-1	0.137	-0.177	0.451
0	0.040	-0.278	0.358
1	0.096	-0.249	0.441
2	0.102	-0.242	0.446
3	-0.024	-0.355	0.307
4	0.056	-0.351	0.463
5	0.172	-0.373	0.717
6	0.214	-0.330	0.758
7	0.179	-0.578	0.936
8	0.288*	-0.035	0.610

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.59: CSDiD model results with the AC control group and a minimum of five observations per farm

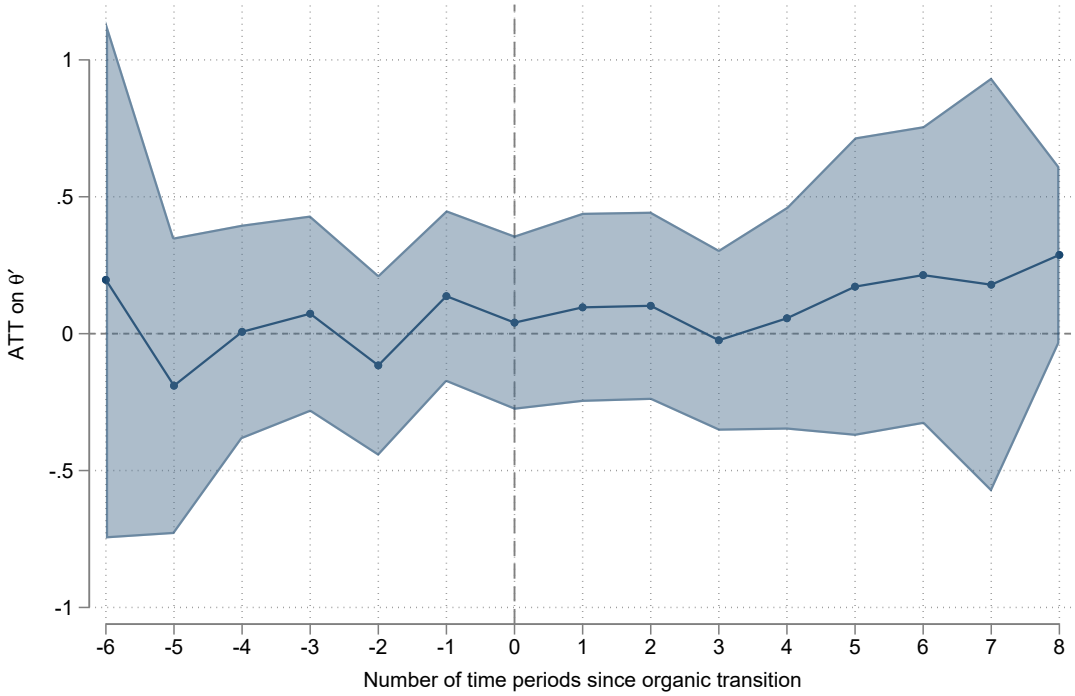

Figure C.3: DiD results with "not yet treated" (pre-S) control group

Time period	ATT	95% lower	95% upper
Pre-treatment	0.018	-0.122	0.157
Post-treatment	0.125	-0.125	0.375
-6	0.197	-0.748	1.141
-5	-0.190	-0.732	0.352
-4	0.007	-0.385	0.398
-3	0.073	-0.286	0.432
-2	-0.116	-0.447	0.215
-1	0.137	-0.177	0.451
0	0.040	-0.278	0.359
1	0.096	-0.249	0.441
2	0.102	-0.242	0.446
3	-0.024	-0.355	0.307
4	0.056	-0.351	0.463
5	0.172	-0.374	0.717
6	0.214	-0.330	0.758
7	0.179	-0.578	0.936
8	0.288^{*}	-0.035	0.610

* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

Table C.60: CSDiD model results with the "not yet treated" control group and a minimum of five observations per farm

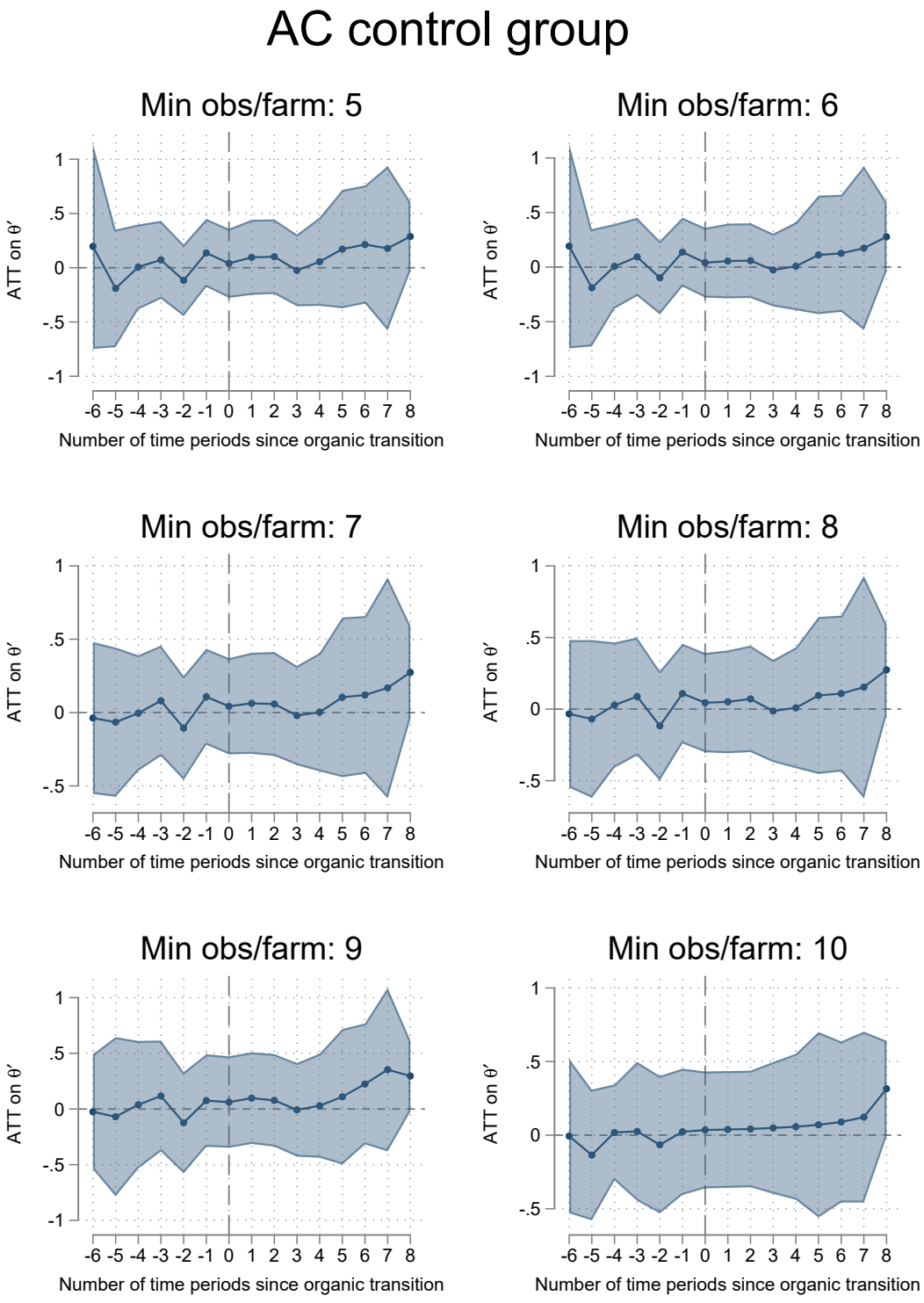

Figure C.4: DiD results with varying minimum observations per farm and always conventional control group

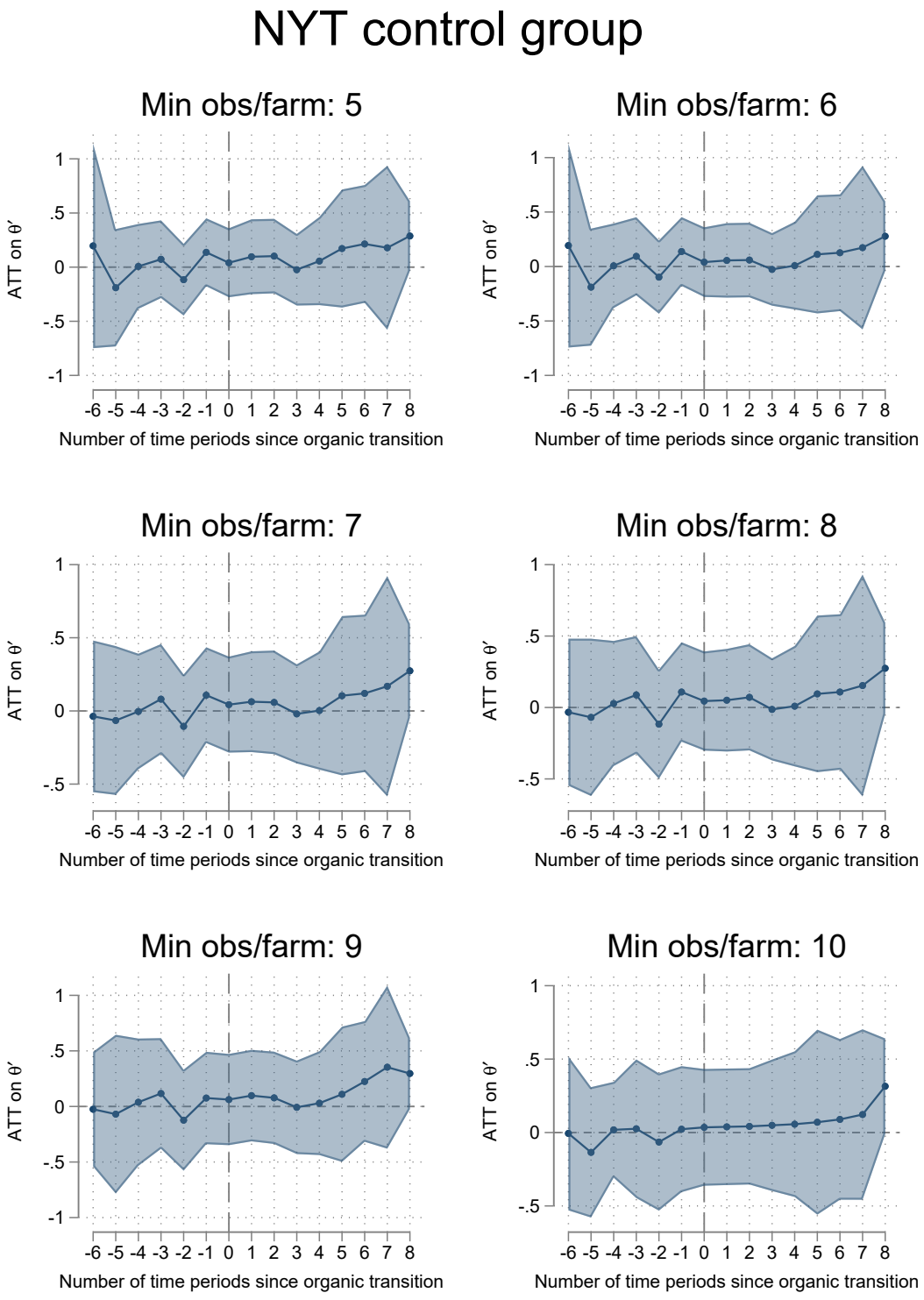

Figure C.5: DiD results with varying minimum observations per farm and "not yet treated" control group