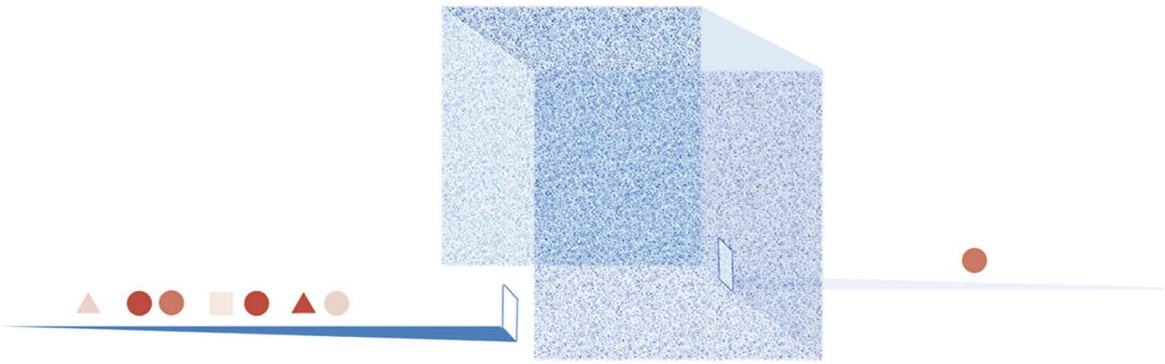


From raw numbers to robust evidence

Finding fact, avoiding fiction

Sam Desiere

UNIVERSITEIT GENT

FACULTEIT BIO-INGENIEURSWETENSCHAPPEN

Promoter:	Prof. dr. ir. Marijke D'Haese
Dean:	Prof. dr. ir. Marc Van Meirvenne
Rector:	Prof. dr. Anne De Paepe

Sam Desiere

From raw numbers to robust evidence

Finding fact, avoiding fiction

Thesis submitted in the fulfilment of the requirements
for the degree of Doctor (PhD) in Applied Biological Sciences

This research was funded by a PhD grant of the
Bijzonder Onderzoeksfonds (BOF).

Dutch translation of the title:
Van ruwe data tot betrouwbare feiten.

Suggested way of citation:
Desiere, S. 2015. From raw numbers to robust evidence. Finding fact, avoiding fiction. Doctoral dissertation, Ghent University.

ISBN 978-90-5989-848-6

Cover by Louize Perdieus.

Members of the examination board

Prof. dr. ir. Patrick Van Damme (Chairman)
Department of Plant Production
Ghent University, Belgium

Prof. dr. Carl Lachat (Secretary)
Department of Food Safety and Food Quality
Ghent University, Belgium

Prof. dr. ir. Jeroen Buysse
Department of Agricultural Economics
Ghent University, Belgium

Prof. dr. ir. Miet Maertens
Division of Bioeconomics
KU Leuven, Belgium

Prof. dr. Koen Schoors
Department of Economics
Ghent University, Belgium

Dr. Alberto Zezza
Development Research Group
World Bank, USA

Prof. dr. ir. Marijke D'Haese (Promoter)
Department of Agricultural Economics
Ghent University, Belgium

Acknowledgments

Nu dit doctoraat voor mij ligt, kijk ik er met een tevreden gevoel naar terug. Hoewel verschillende hoofdstukken erg kritisch staan tegenover statistieken, blijf ik geloven dat kwantitatief onderzoek interessant en leerrijk kan zijn. Bovendien is nadenken over de analyse en interpretatie van data ook gewoon leuk en geweldig uitdagend. Bijna vier lang heb ik dit ongestoord kunnen doen, en daar wil ik verschillende mensen voor bedanken.

In de eerste plaats, Marijke, promotor van dit werk, voor het aanbieden van verschillende kansen de afgelopen jaren. Een nieuw onderwerp of methodologie ontdekken en uitdiepen, samenwerken met andere universiteiten, rustig een (te) moeilijk artikel lezen, het delen van contacten, een congres meepikken of op het laatste moment een paar dagen vakantie nemen, het kon allemaal. Deze flexibiliteit zorgde voor een unieke werkomgeving, waarin er ruimte was om bij te leren en te experimenteren. De leden van de doctoraatsjury bedank ik graag voor hun waardevolle commentaren, het delen van interessante literatuur en de suggesties voor verder onderzoek. Tenslotte wil ik ook Annick en Sibylle bedanken voor hun hulp de afgelopen weken en maanden met alle administratieve en praktische zaken die bij het afronden van een doctoraat komen kijken.

Zonder de collega's binnen de vakgroep zou het dagelijkse doctoraatsleven maar een saaie bedoeling zijn. De kleine rituelen – lunch om 11u45, koffie met koekjes rond 15u30, badminton van tijd tot tijd, taart bij verjaardagen, en vaak bier op vrijdag - zorgden voor een aangename werksfeer en, indien nodig, voor een uitlaatklep wanneer het werk niet even snel vorderde als gehoopt. Jullie zijn met teveel om hier iedereen te vermelden, maar de grappige anekdotes tijdens de

lunch over huizen kopen, politiek, taal, moeilijke samenwerkingen en conflicten, kinderen, fietsen en eco-cheques zal ik niet snel vergeten. Een speciaal woord van dank aan Wytse en Lotte. Het was fijn om met jullie een kantoor te delen, te keuvelen, praktische problemen aan te pakken, te vloeken bij een 'rejected' paper en samen papers te schrijven. Ik heb veel aan jullie gehad. Sanctus wil ik graag bedanken voor de samenwerking rond verschillende projecten in Burundi. Zonder Sanctus zou de data wellicht nog steeds niet in Gent zijn aangekomen. Ik hoop dat de situatie in Bujumbura snel stabiliseert en kijk er naar uit de laatste papers over Burundi af te ronden.

Werken is belangrijk, ontspannen is dat evenzeer. Anton, Bart, Griet, Maarten, Maurits, Pepijn, Pieter en Simon, dank jullie wel voor de onvergetelijke reizen, de roadtrip naar Roemenië, de weekends in de Ardennen en de kaartavonden. Bedankt ook voor de filosofiche discussies over de uitdagingen van doctoreren en de meer praktische hulp van Bart, die zonder aarzelen zijn Latex-files voor de layout van dit doctoraat met mij deelde. Aan alle JNM'ers, te talrijk om hier allemaal op te noemen zonder iemand te vergeten, dank jullie wel voor de tomeloze energie en engagementen voor doldwaze acties en kleine en grote projecten. Ik denk niet dat ik ergens meer heb geleerd dan van jullie. Zonder mijn huisgenootjes van de voorbije jaren – Céline, Louize, Marie, Matti, Ole, en Sofie - was dit doctoraat er wellicht nooit geweest. Elk van jullie stond mij met raad en daad bij en zorgde voor gezellige avonden, heerlijke maaltijden en verrijkende gesprekken. Dankzij Céline hoorde ik dat de vakgroep landbouweconomie bestond, Louize ontwierp de voorkaft, en dankzij de aanmoedigingen van Ole publiceerde ik een kort artikel over carbon emissions. Eva, tenslotte, dank je wel voor de fijne tijd. Ik kijk uit naar wat de toekomst zal brengen.

Veel van de papers in dit doctoraat staan erg kritisch ten opzichte van data en statistiek. Stilletjes vraag ik mij af of die (te) kritische houding niet met de paplepel is ingegeven. Onbewust hebben mijn ouders en mijn broer een grote invloed op dit doctoraat gehad. Dank voor alle kansen van de voorbije jaren.

Ik hoop, beste lezer, dat je (een deel van) dit doctoraat leest of doorbladert en het boeiend vindt, dat het leidt tot nieuw onderzoek en dat sommige aspecten beleidsmakers bereiken. Alle feedback, nieuwe ideeën of bedenkingen zijn meer dan welkom.

Sam Desiere
Gent, december 2015

Contents

CHAPTER 1

Introduction

1.1 Introduction

Numbers are at the heart of many social science PhDs. Researchers collect and analyze quantitative data to test hypotheses and to refine theories. In contrast to experimental research, socio-economic research is bound to work with non-experimental data. This means that work in the social sciences has specific characteristics mostly derived from the uncertainties that can arise from working with non-experimental data. One classic example is the difficulty in establishing whether there are causal relations, rather than simple correlations, between variables. This is a phenomenon that economists refer to as the *endogeneity problem*. The issue of endogeneity is, in some respects, the bread and butter of modern economics and has led to the proliferation of sophisticated econometric models. However economists have paid far less attention to a second important aspect of non-experimental data, that is, its quality. Since non-experimental data needs to be collected in the 'real' world, the way data is collected and processed matters and will influence the outcomes of any subsequent analysis. In short, numbers are socially constructed and can only be interpreted within the context in which they are generated or gathered.

Applied economic research aims to establish causal links between concepts and there is an implicit assumption that concepts and data on which they are based have been correctly measured (Boumans, 2005a; Morgan, 1991). For example, when economists empirically study the impact of economic growth on poverty, it is under the assumption that these two concepts have been reliably measured. This is in sharp contrast to other disciplines within social research, such as applied educational and psychological research, which are primarily concerned with accurately measuring concepts. Economists tend to leave the arduous task of measurement to national and international institutions and have little interest in the measurement procedures employed or their accuracy (Reiss, 2013)[1]. This PhD challenges the assumption that economic concepts are adequately measured; it examines the process of transforming raw data into measured concepts and explores how measurement error can bias the claimed relationships between concepts. More specifically, it addresses three inter-related research questions: (i) how is data collected? (ii) how can we quantify concepts? and (iii) given that the concepts are always imperfectly measured, how do these imperfections affect the claimed causal links between the two concepts? Since the pitfalls in the measurement process oc-

[1] This is a strong generalization. Some academic economists do devote time and energy to gathering reliable data. For instance, the most important academic contribution of Piketty's bestseller Capitalism in the twenty-first century is the presentation of new, accurate time series on inequality (Piketty, 2014). Many other economists have pointed out the risks related to using secondary datasets in academic research. Prominent examples are Atkinson (2001); Jerven (2013a); Devarajan (2013). Moreover, in recent years there has been more academic interest in improving data quality. See, for instance, the Living Standard Measurement Surveys (LSMS) of the World Bank and increased attention for compulsory publication of the data along a journal article (Hanson et al., 2011).

cur due to many different reasons, I do not limit this thesis to a specific topic, research field or methodology. Rather, this study draws from different fields ranging from (agricultural) economics, to statistics, educational sciences and political economy.

Many researchers are concerned about the quality of their raw data or devote themselves to developing new measurement instruments to quantify concepts. Much effort goes into constructing reliable datasets and developing measurement instruments that can quantify concepts. What is novel about this PhD is that it studies the whole process of data transformation - from the initial data collection, through the transformation of data into measurements, to using the outcomes of measurement to establish relations between concepts - as a topic in its own right, rather than as a necessary step in every research process.

The type of data I work with in the different case studies presented in this PhD are very similar. All data were collected at household level in developing countries (mainly in Burundi and Rwanda, but also in Colombia and Ecuador) through door-to-door interviews by trained enumerators, commissioned by Statistical Offices, NGOs or universities. The surveys dealt with poverty, food security and/or agriculture. The main aim of the surveys was to provide the necessary background to evaluate or design rural policies. In this perspective, the surveys served the broader purpose of enhancing the 'evidence base' in order to develop evidence-based policies. I work with data that are (publicly) available and I had no say in the design of the studies. This limited the research questions that could be asked. It does, however guarantee that the 'errors' in the data are likely to occur in other settings as well. The data I used are classic data gathered using conventional methods. They have nothing to do with the new trend of 'big data', such as satellite data or data from mobile phones, which are currently receiving much attention in the academic world and the media (Mayer-Schönberger and Cukier, 2013; Varian, 2014). While 'big data' may be the future, traditional sources of data (i.e. censuses and surveys) still remain the most important source of information in the developing world today (Carletto et al., 2015b). As such studying the quality of household survey data remains an essential task.

A second common feature to most of the case studies in this PhD is that they take the definition of concepts as given. In other words, I do not question the relevance of a concept nor do I discuss its limitations. I recognize Koopman's famous dictum that measurement without theory is not feasible or, at least, not efficient (Deaton, 2010; Koopmans, 1947). Yet, I do not discuss theories, but focus on how concepts are quantified. I do accept that the concepts discussed in this PhD can be quantified, which is also an assumption that is certainly debatable. Using the well-known concept of GDP as an example, I illustrate the difference between criticizing a concept because it is deemed not fit for the purpose at hand (something I avoid in this thesis), and criticizing a concept because it is not being measured accurately and precisely, which is what I do for several concepts in

this thesis. GDP has widely been criticized because it does not take inequality into account and is a not a good proxy for well-being (Stiglitz, 2009). This may or may not be the true, but this thesis would not pursue this line of enquiry. Instead, it would criticize the concept of GDP on the following grounds (i) the data required to measure it are unavailable or unreliable (Jerven, 2013a); (ii) GDP is inaccurately measured as the black market is systematically excluded, and its size varies between countries (Enste and Schneider, 2000) and; (iii) poorly measured GDP may create spurious correlations between, for instance, economic growth and volatility of growth rates (Dawson et al., 2001; Woods, 2014). As far as possible, I avoid the debate about the 'correct' definition of concepts such as poverty and food security, which are two important concepts in this PhD.

In the remainder of the introduction, I briefly explain why numbers are so widely used in the social sciences and, even more so, in policy circles. The extent of their use and the reliance that is placed upon them is the main motivation for undertaking this research. Because numbers are so ubiquitous, understanding the process behind their construction is essential. I emphasize that numbers are not objective facts, but are man-made and, as such, are subjective. After this discussion I discuss how raw data are transformed, through measurement, into quantitative concepts and highlight several pitfalls in this process. In the final section of the introduction, I present the structure of the thesis.

1.2 Why are numbers so popular?

The popularity of numbers in the social sciences, and particularly in economics, almost goes without saying. Kelvin remarked in the 19th century, that *"when you cannot measure, your knowledge is meager and unsatisfactory"*. While he made this statement in relation to physics it has also become a standard dictum in the social sciences. This has not always been the case. In the natural sciences measurement devices were developed as early as the 17th and 18th centuries. Yet economists only started to develop their own measurement devices at the end of the 19th century (Morgan, 2001). It is worth recalling that Adam Smith's magnum opus 'The Wealth of Nations', published in 1776, is not based upon empirical models or measurement (Blaug, 2002). It is arguable that measurement became popular in the social sciences because researchers longed to lay claim to the same standards as those of, say, physics and the accompanying prestige (Kuhn, 1961; McCloskey, 1983). Measurement was seen as a first step towards a mathematical representation of social realities (Porter, 2001).

Measurement and formalization do indeed help in making elusive concepts, such as poverty or food security, more tangible and clear-cut. They reduce much of the clutter associated with qualitative arguments and make implicit assumptions more explicit. Even today, showing that there exists a statistically significant associa-

tion between two variables is often considered irrefutable evidence of a particular theory, while substantiating a theory with qualitative arguments is considered less convincing (Head, 2008; Porter, 1996; Ziliak and McCloskey, 2008). Numbers are considered to be more objective, value-free, neutral and harder to manipulate than qualitative arguments[2].

In his seminal work 'Trust in numbers' Porter argues that the perception of numbers as being objective and value free explains their proliferation (Porter, 1996). In his view, it is not scientific rigour that leads to quantification, but pressure from the outside to generate 'objective' knowledge. He argues that the strict rules and procedures associated with measurement and statistics lend numbers an air of objectivity. Standardization of measurements limits the scope for personal judgment and, as such, serves as a check on subjectivity. The strategy of impersonality, as Porter calls it, helps to build trust in the people and institutions who produce the knowledge, who may not otherwise be trusted by the outside world. The quantification of the value of ecosystems, a currently popular research topic, can be considered such a strategy. While it is extremely difficult to assign a monetary value to ecosystem services, doing so seems to be viewed as a more value-free argument than moral arguments about the importance of nature conservation (Diamond and Hausman, 1994). Ecologists and environmental economists, two groups whom the general public may not fully trust, attempt to enhance their status and the 'truth' of their knowledge by using quantitative arguments.

Numbers are thus a communication strategy. They travel well within the public domain and are often instantly quoted by researchers and the media (Howlett and Morgan, 2010). This is well illustrated by a report, published in 1996 by the FAO, which estimated the number of undernourished people to be 841 million (FAO, 1996; Smith, 1998). A simple search on Google reveals that this figure had already been cited over 33 000 times by January 2015[3]. One can think of other 'famous numbers' from different disciplines that have a life of their own. Numbers, particularly indicators, are sometimes even created, not because they contain valuable information, but more to promote a cause and get media attention (Bateman, 2001; Kelley and Simmons, 2015). The famous quote of Bill Gates *"if you can't measure, measure anyhow"* summarizes this strategy. It has led to an explosion of indicators and rankings the added value of which are questionable (The Economist, 2014). The power of socio-economic indicators to shape policies, criticize governments or advocate for a particular cause cannot be underestimated (Kelley and Simmons, 2015).

[2] An excellent example is the slogan of the Belgian employers' organizations Agoria protesting against the strikes by trade unions in Belgium in November/December 2014: "Geen slogans, wel cijfers (numbers instead of slogans). This nicely illustrates the point that numbers are believed to be more 'true' and more difficult to manipulate than words.

[3] Search term "800 million people are undernourished": 33 100 hits; "800 million undernourished" (without quotations) 87 600 hits; "800 million undernourished" (without quotations in Google Scholar) 17 400 hits (searches: 20 January 2015)

The increasing call for accountability, transparency and evidence-based policies has further contributed to an explosion of numbers in the public sphere (Pawson et al., 2011; Reiss, 2013; Young et al., 2002). This evolution is often associated with an increase in demand for quantitative data and careful impact evaluations of development programmes (Ravallion, 2014). Within the development sector, the Paris Declaration on Aid Effectiveness (2005) and the Accra Agenda for Action (2008) both emphasized the need to measure the impact of development aid (OECD, 2005/2008). The recent interest of philanthropists in development aid has further strengthened the idea that 'returns on aid' have to be measured (The Economist, 2015b). For instance, more and more donors in the developing world want to measure the impact that their funding has on poverty alleviation (Pérouse de Montclos, 2012). Impact evaluations are highly data-intensive and, at the least, require the collection of good quality data before and after the implementation of a project. Yet, the cost of a thorough impact evaluation is often prohibitive, leading to less expensive, yet also less informative, data collection efforts to monitor development programmes.

In academic socio-economic research, there has also been a shift away from theory building and towards more empirical research, especially in development economics (Angrist and Pischke, 2010; Bromley, 2008). Research now focuses on what works instead of why it works. This is partly the result of the increasing availability of data at low cost, in combination with the decreasing cost of computer power over recent decades, but is also due to a genuine paradigm shift towards more empirical work (Rodrik, 2008). Randomized controlled trials (RCTs), which are currently an extremely popular way to measure the impact of interventions by government or NGOs, are a classic example of this trend (Banerjee and Duflo, 2011). The growing trend of systematic reviews, based on meta-analyses of all the available quantitative studies within a given research topic, also illustrates the growing interest in systematic and standardized empirical research (International Initiative for Impact Evaluation (3ie), 2015).

In sum, there is a growing trend towards quantification in the social sciences. Notwithstanding the many advantages of numbers in studying and describing facts, socio-economic research data should be treated with many caveats. In contrast to the widely-held view, measurements and quantitative knowledge are not value-free, objective facts (Schedler, 2012). This is because data not only needs to be correctly recorded, but also has to be transformed into measurements. In other words, while raw data may be objective, measurements require assumptions and as such, quantitative knowledge can never be value free (Reiss, 2014). To go back even further, the choice of which data are collected, and which are not is already a subjective decision in itself. In the following section, I discuss how measurement tools in the social sciences are developed.

1.3 Measurement in the social sciences

Theory posits relations between concepts. Before one can empirically test if these relations hold, the concepts need to be measured. Yet, by definition, concepts are unobservable. This is true in both natural sciences and social sciences. Yet, some concepts are more easily definable, and thus observable, than others. In the natural sciences, for instance, it is easier to measure the concept of 'land area' than that of 'temperature'. The easier it is to observe or define a concept, the less difficulty there is in developing a measurement instrument that quantifies the concept. In this sense, it is no surprise that the development of the thermometer to measure the unobservable concept of temperature was more arduous than the development of simple tools to measure 'land' (Chang, 2004). Similarly, in the social sciences, it is easier to measure the simple concept of 'family size' than the intangible concept of 'poverty'.

Once a measurement instrument has been developed and is considered to be reliable by the research community, the status of a concept changes from 'unobservable' to 'observable'. Temperature is by its nature an unobservable concept, but became observable when a reliable measurement instrument, the thermometer, was developed to quantify it. In a way, temperature does not define the thermometer, but the thermometer defines temperature. In this sense, the concept and the measurement instrument are intrinsically linked and cannot be considered separately.

In the natural sciences, the measurement procedures for many important concepts are long established and no longer subject to debate. In the social sciences, by contrast, there is less agreement about measurement procedures. There are still disagreements over the best approach for quantifying key concepts, such as poverty, in socio-economic research. As a result, numerous measurement instruments for quantifying poverty have been elaborated and the literature about poverty indicators is still very active (Alkire and Santos, 2010). The debate about the 'best' measurement instrument can often be traced back to different views of the correct definition of the concept. In other words, the research community is still questioning whether the different measurement instruments really capture the concept they are intended to measure. This is called the 'validity' of the measurement instrument. A second, less contentious, question concerns the accuracy and precision of the measurement instruments. This does not question the validity of a measurement instrument, but questions whether or not a specific instrument systematically mismeasures the concept under particular circumstances (inaccurate measurement) or produces a random error that is so large that the instrument cannot be used for most practical purposes (imprecise measurement).

Section 1.3.1 defines and formalizes the measurement process and goes on to discuss valid, accurate and precise measurement. This section presents a simple

model that links the true, unobservable value of the concept with the observable outcome of a measurement. The process is paved with obstacles, widespread in the social sciences, that can lead to invalid, inaccurate and imprecise measurement. In the next section (1.3.2), I define the validity, accuracy and precision of measurement instruments and highlight the central role of 'gold standards' in assessing their quality. Even if a measurement instrument is in principle valid, accurate and precise, the quality of the raw data may still be an important influence, a point which is discussed in section 1.3.3. Following on from this, in section 1.3.4, I discuss how these factors influence researchers' endeavours to establish relations between concepts and to develop 'evidence-based policies' and whether these relations can be confidently established if the concepts have been imperfectly measured.

1.3.1 The measurement process

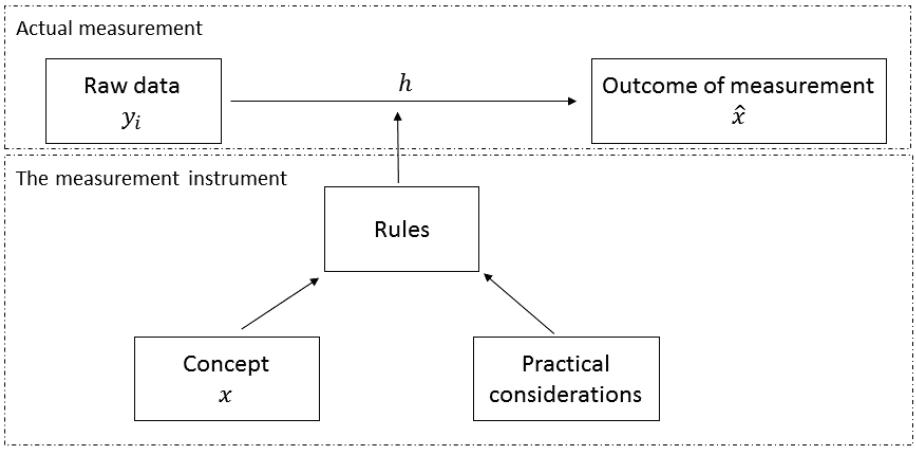
Measurement theory defines measurement as a *"process of assigning numbers to attributes of objects and events of the real world according to rules, in such a way to represent them, or to describe them"* (Finkelstein, 2005; Schedler, 2012). In other words, measurement translates concepts into numbers that can be interpreted and processed for further use.

Figure 1.1 shows the measurement process in the social sciences. To transform data into measurements, one needs instruments for measuring. While measurement instruments in the natural sciences such as the thermometer need to be 'build', 'rules' are at the heart of the measurement process in the social sciences. These rules specify how raw data are transformed into the outcome of the measurement. They have been, and are, developed by social scientists who aim to capture the concept in a way that is compatible with practical considerations, such as the cost of data collection and the ease of data processing.

The measurement process, as illustrated in figure 1.1, can also be examined more formally. A formal treatment helps to identify the key conditions for precise and accurate measurement[4] Equation 1.1 links the true value of the concept, x, to the outcome of the measurement process, $\hat{x}$. This outcome variable is obtained through a transformation of a vector of observable raw data, y_i, according to pre-defined rules, represented by the function h. The key assumption of the measurement process is that the observable data, y_i, are correlated with the true, unobservable value of the concept, x. This correlation is represented by the function g. Observational errors enter the measurement process through the context, also referred to as 'other circumstances (OC)' in the literature, in which the measurement is operationalized. As a consequence, there is a direct relation between the true value of the concept, x, and the measurement result, $\hat{x}$. However this relation

[4]This section draws heavily from the framework developed by Boumans (2005b, 2009, 2013, 2015); Finkelstein (2005) and from the excellent book *Science outside the laboratory: measurement in field science and economics* (Boumans, 2015).

Figure 1.1: Measurement in the social sciences



is imperfect because of disturbances from the 'noisy environment' in which the measurement takes place.

$$\hat{x} = h(y_i) = h(g(x)) = f(x, OC) \tag{1.1}$$

Presenting this equation as a differential equation provides additional insights about the measurement process and helps to show the conditions that needs to be fulfilled for 'perfect measurement'.

$$\Delta\hat{x} = \frac{\partial f}{\partial x}\Delta x + \frac{\partial f}{\partial OC}\Delta OC \tag{1.2}$$

This shows that a change in the outcome of a measurement, $\Delta\hat{x}$, is determined by two factors: either the true value of the concept, Δx, or the context, ΔOC, may have changed.

An ideal measurement system requires that the second term in the equation, $\frac{\partial f}{\partial OC}\Delta OC$, equals zero. When this condition is met the outcome of the measurement is solely a function of the true value of the concept and is not affected by any noise in the data. The measurement instrument is then working independently from the context in which it is used. This goal can be achieved in a laboratory setting where measurement occurs in a controlled environment. In such a setting the noise due to changing circumstances can be controlled and limited. Consequently, the famous *ceteribus paribus* condition holds: $\Delta OC = 0$ or even $OC = 0$. Yet, even in a controlled environment, it is often not possible to eliminate all 'noise'.

A second strategy to obtain reliable measurements is to design measurement instruments in such a way that background noise has only a limited effect on the

measurement. This requires that the *ceteribus neglectis* condition, $\frac{\partial f}{\partial OC} \approx 0$, is met. If this condition holds, the noise in the environment does not affect the measurement. Social scientist aim to design measurement instruments that satisfy this condition.

However, when designing such instruments, the so-called *problem of the passive observer* haunts the social scientist (Haavelmo, 1944). Once again, this problem occurs because social researchers observe phenomena in the real world and cannot control the circumstances. If a stable relation between the true value of the concept, x, and the measurement result, $\hat{x}$, is observed, it can be concluded that $\frac{\partial f}{\partial OC}\Delta OC = 0$. This implies that $\frac{\partial f}{\partial OC} = 0$ or that $\Delta OC = 0$. If $\frac{\partial f}{\partial OC} = 0$, then the instrument is a valid tool under changing circumstances. However, given that one can only observe if $\frac{\partial f}{\partial OC}\Delta OC = 0$, one can never be certain that $\frac{\partial f}{\partial OC} = 0$ since it may be the case that the environment did not change (i.e. that $\Delta OC = 0$). In practice, it is often difficult to determine if there was sufficient variation in the environment, that is, if ΔOC changed substantially, to conclude that the measurement instrument is robust enough to adapt to changing circumstances. One way to deal with this issue is to test a measurement instrument under many different circumstances. A second approach, easier in practice, is to identify in which particular contexts a particular measurement instrument works (Friedman, 1953), although this approach does not guarantee that the measurement instrument will also work under new circumstances.

1.3.2 Valid, accurate and precise measurement

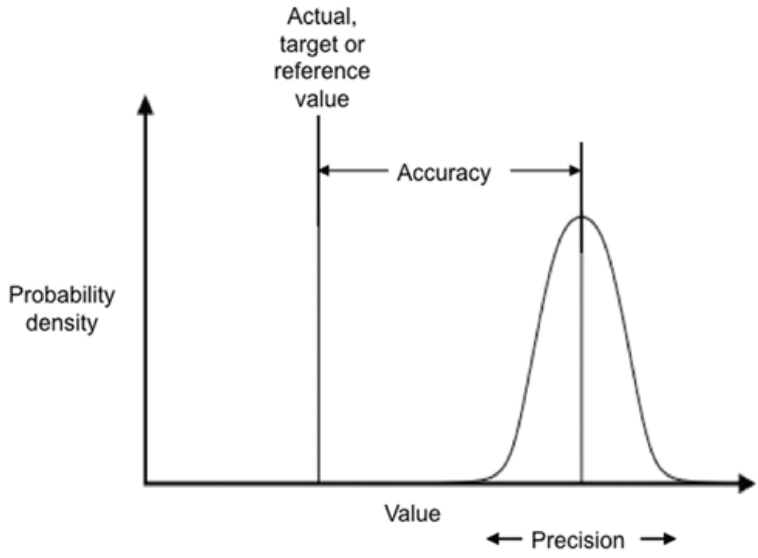
As discussed above, measurement in the social sciences can be challenging because of a combination of the difficulties of controlling errors and the problem of the passive observer. As a result, the true value of the concept, x, will differ from the outcome of the measurement, $\hat{x}$. These differences cause invalid, inaccurate and imprecise measurements.

While the term 'precision' is well-defined, there is much more confusion about the definitions of 'validity' and 'accuracy' (Boumans, 2009; Cafiero et al., 2014). In this thesis, validity relates to the degree to which a measurement instrument captures an unobservable concept. In other words, a measurement instrument is valid if it measures the concept that it is intended to measure. For instance, it has been argued that defining a poor household as 'a household with an income below a certain threshold' is not a valid approach to measuring poverty since it reduces the multidimensional concept of poverty to income poverty. In sum, validity is a qualitative concept which relates to the correct definition of the concept that is being measured. Assessing the validity of a measurement instrument requires careful reflection on the concept because the measurement instrument cannot formally be tested in isolation.

Accuracy and precision have close links with two statistical terms, systematic error

and random error. A measurement instrument that is both accurate and precise is called a reliable measurement instrument. Accuracy requires that the measurement instrument is not systematically biased under certain conditions or for some subgroups of the population. For instance, a poverty instrument that systematically underestimates poverty in rural areas and overestimates it in urban areas is systematically biased and, hence, inaccurate. Precision refers to the spread or variation of the estimates around the central value. Poverty is precisely estimated if repeating the measurement under the same conditions with the same instrument gives a similar outcome. The difference between precision and accuracy is illustrated in figure 1.2.

Figure 1.2: Accuracy versus precision



Source: Wikipedia https://commons.wikimedia.org/wiki/File:Accuracy_and_precision.svg

While validity and accuracy are clearly defined in principle some confusion nonetheless occurs because the term 'validity' is only meaningful if the concept is unobservable, i.e. not clearly defined. The term 'accuracy', on the other hand, can be applied to both observable and unobservable concepts. The key difference between an observable and unobservable concept is that observable concepts can be perfectly characterized by a gold standard. A gold standard is defined as some knowledge of the true value, x, of the unobservable concept, at least under certain conditions. Temperature, for instance, has a gold standard, while poverty does not. A gold standard allows researchers to calibrate new measurement instruments or to test their validity (Bland and Altman, 1986).

Gold standards are rare in the natural sciences and even more so in social sciences. If there is a gold standard, however, the term 'validity' is not meaningful. To illustrate this point, consider the measurement of temperature, which has a gold standard, and poverty, which does not. If one develops a new thermometer, one can calibrate it against other thermometers that are known to perfectly measure temperature under certain conditions. Because of the existence of a gold standard, there is no need to define validity since the concept of temperature is exactly determined by the gold standard. A thermometer can, however, be inaccurate in certain circumstances. The validity of a measurement instrument can thus only be discussed if there is uncertainty about the definition of the concept, that is, if there is no existing gold standard.

Poverty is an example of a concept for which there is no gold standard. What is typically done to validate poverty instruments is to measure the same concept using different measurement instruments and to compare the findings. This is also called concurrent validity (Cafiero et al., 2014). For instance, the poverty status of a household can be determined using indicators based on income or multidimensional indicators. If, for most households and in most situations, both indicators classify households in the same category, both indicators measure the same concept equally well. However, we do not know if the concept that is being measured is a valid proxy for poverty. We only know that both indicators measure the same latent variable. If the different measurement instruments for poverty result in a different classification of households, either one (or both) of the instruments is invalid or both are valid, but they are measuring a different aspect of poverty. Hence, whatever the correlation between the two different measurement instruments for poverty, their validity cannot be proven as long as there is no gold standard. Only qualitative arguments, based upon a conceptualization of poverty, can determine which instrument is more valid.

The term accuracy is defined for concepts with and without a gold standard. Accuracy relates to 'the closeness between the results of the measurement,$\hat{x}$, and its underlying true value, x' (JCGM, 2008). The problem is that the accuracy of a measurement instrument can only be examined if one first defines a gold standard against which any new measurement instrument can be benchmarked. The gold standard of poverty is often defined as a household consuming less than a predefined quantity of different goods, which is calculated by applying the 'cost of basic needs approach', using consumption expenditure data from a representative household survey (Deaton, 1997). Once the gold standard is selected, one can test whether, for example, recall by the household head or personal diaries more accurately estimate consumption expenditure (Beegle et al., 2012) or whether larger households systematically underreport food consumption (Gibson and Kim, 2007)

As the examples above illustrate, there are no mechanical procedures for assessing the validity of a measurement instrument. Assessing the accuracy of a measure-

ment instrument is also challenging since it requires a (partly arbitrary) choice of a gold standard. Hence, assessing the validity and, to some extent, the accuracy of measurement instruments is based on the personal judgement of experts, as it requires critical thinking and an excellent, qualitative understanding of the concept to be measured (Boumans, 2015; Schedler, 2012). It is as this point that a certain degree of subjectivity can sneak in. Subjectivity does not imply every expert developing their own measurement instrument. Rather, it implies that debate (eventually leading to consensus) among experts in the field is required to develop an accurate measurement instrument. Even if no consensus is reached, the debate can contribute to refining existing theories about a concept or the design of better measurement instruments (Jick, 1979).

While the precision of measurement is often discussed, the accuracy of the measurement is sometimes neglected, especially in economics. It may then happen that the rules of the measurement instrument replace the concepts. Many researchers and policymakers concerned with poverty in developing countries, for instance, link poverty to those people that live on less than one-dollar-a-day, which is one of the most widely used definitions of poverty[5]. In so doing there is a risk that they forget that poverty is a much broader concept. The French statistician and philosopher, Alain Desrosières, refers to this phenomenon as 'the paradox of statistics simultaneously being the referent and reality' (Desrosières, 2002/1993)[6]. As such, statistics can acquire a life of their own.

Imprecise or inaccurate measurement can limit the usefulness of indicators. Poverty and food security indicators, for instance, are only suitable to monitor the impact of development programs if they satisfy certain validity criteria. Within the framework of this thesis, three different validity criteria of food insecurity and poverty indicators are examined: cross-sectional validity, inter-temporal validity and internal validity. Cross-sectional validity means that a good indicator distinguishes poor (food insecure) from non-poor (food secure) households in a cross-sectional setting. In statistical terms, cross-sectional validity corresponds to indicators with high 'sensitivity' (true positive rate) and 'specificity' (true negative rate). This is an important condition for development programmes that aim to reach out to poor households and exclude the non-poor households. Inter-temporal validity implies sensitivity of the indicator to changes in poverty (food security) status over time. This is required to monitor the impact of a development programme over time. Finally, most socio-economic indicators consists of several questions

[5] Although not completely arbitrary, the choice of one-dollar-a day as the threshold for poverty is by itself subject to debate. Since 2005, the World Bank has set the threshold at $1.25 a day (Ravallion et al., 2009) This shows, once again, why defining rules to measure a concept requires subjective assumptions.

[6] In his book, 'The politics of large numbers: a history of statistical reasoning', he gives the example of 'les cadres', a term introduced to classify managers of a firm (and to distinguish them from the ordinary 'employées') in statistical reports in France. Over time, some individuals in society started to identify themselves as 'les cadres' as a distinct, 'higher' social class in society.

that are related to the concepts that one is seeking to measure. Internal validity requires that these different questions measure the same underlying construct. This is often evaluated by Cronbach's alpha. Imprecise or inaccurate measurement or a combination of both can limit the cross-sectional, inter-temporal or internal validity of an indicator.

In contrast to multidimensional concepts such as poverty, some concepts can more easily be translated into rules and, thus, measured. Even in those cases, careful data collection is still essential to obtain precise results. Noisy data will result in imprecise estimates. Moreover, data collection may introduce systematic measurement errors which reduce the accuracy of the results. If households, for instance, systematically underreport their wages because they do not like to share sensitive information, poverty estimates will be overestimated and thus inaccurate. Thus, the difficulties involved in data collection are an important aspect in this thesis. This is because data collection in itself requires many implicit and explicit assumptions which play an important role when data is finally converted into measurement. Data collection is therefore discussed at length in the next section.

1.3.3 Data collection

Although often neglected, it is a simple but crucial fact that one needs reliable data to get reliable results (Maier and Imazeki, 2012; Woods, 2014). As shown in figure 1.1, the measurement process always starts with the input of raw data, y_i. Data collection is, however, never a trivial – and is often an expensive – task. Funding of data collection, overcoming logistical and technical challenges in gathering the data, avoiding measurement error in key variables and ensuring that the purpose of the data collection does not create incentives to manipulate the numbers are all integral parts of the process behind data generation[7].

Data collection in developing countries is expensive (Jerven, 2013b). Data collection is often (partially) funded by donors and has often been project-based. As such there is often a lack of the long-term commitment needed to assemble good quality data spanning several years or decades (Upton et al., 2015). At the same time, core funding for national statistical offices has declined in the last decades. As a result, there is often a lack of reliable data on key indicators such as population growth, GDP and agricultural production in developing countries, particularly in Africa (Carletto et al., 2013a; World Bank and United Nations and Food and Agricultural Organization, 2010). Most figures, published annually by international institutions, are predictions or estimates that are only irregularly updated when a new survey becomes available. Devarajan (2013), currently chief economists at the World Bank's Middle East and Northern African region, refers to 'Africa's Statistical Tragedy'. The Millennium Development Goals (MDGs),

[7] Schoors' (2000) account of his quest to build a dataset of Russian banks in 1995 is fun reading. It demonstrates how difficult and time-consuming it can be to collect data and to check their reliability.

for instance, set clear and measurable targets for progress in human development, but the progress made towards their realization is unclear as the data needed to evaluate this is absent in many regions (Attaran, 2005). Scholars have also pointed out that the Sustainable Development Goals (SDGs), which are due to replace the MDGs at the end of 2015, will substantially increase the demand for data to monitor them (Sachs, 2012). Once again, the poorest regions will probably lack the necessary capacity to collect and process the data or the cost of data collection may exceed the benefits (Jerven, 2014a).

Surveys and censuses are important sources of information in developing countries (Carletto et al., 2015b). However, the logistic and technical challenges of setting up these surveys are, enormous. The conceptually simple prerequisites for collecting good quality data turn out to be difficult to apply in practice. One well-known example is the need to have a 'representative sample' to enable extrapolation of research findings to an entire population. This raises important practical questions about how to randomly draw a sample of respondents from a region when there is no complete list of all the inhabitants, or how to reach people in remote areas. Moreover, it is also well known that the definition of a household as an economic unit is problematic in some contexts (Randall and Coast, 2014).

Obtaining accurate measurements of key variables from surveys can be hard. Measuring land area, for instance, is an essential part of agricultural surveys. But there is an ongoing debate about whether one should use GPS, tape and compass methods or farmer' self-reported estimates of land area by farmers (Carletto et al., 2014). Similarly, household food consumption and food expenditure data are considered crucial measurements of poverty and food security, but measuring them accurately is notoriously difficult. Measurements based on recall or on personal diaries are known to differ substantially across settings (Beegle et al., 2012). In addition the answers may be affected by psychological or contextual factors (Groves and Couper, 2012). Respondents may not be able to recall how much food they consumed in the last two weeks or may be unwilling to provide the enumerator with sensitive information about their wages or ethnicity.

Data are collected for many purposes and some of these may create incentives to manipulate the numbers (Jerven, 2014b). Historically, the registration of economic transactions was closely tied to administrative procedures and conducted by bureaucracies (Porter, 2001). Research or policy analysis were not the prime reasons for data collection. Data collection served political purposes such as tax collection or economic reforms. This is often still the case today. As rational actors in the data collection process realize the purpose of the survey, they may have an incentive to manipulate the numbers. This aspect of data quality is known as the political economy of data (Sandefur and Glassman, 2015) and can have serious consequences for data quality and availability (The Economist, 2015a). For instance, when population censuses serve as the basis for sharing power between regions, as was the case in Nigeria (Jerven, 2013b), many key players may over-

estimate the actual number of inhabitants. Similarly, when local officials have to reach targets set by central government, they may overestimate the positive effect of certain policies (Sandefur and Glassman, 2015). These data may then enter the national statistical system and be used for new research, with the researcher unaware of the purpose for which the data were originally collected. Obviously, their research findings will not reflect reality.

1.3.4 From measurement to (causal) relations

Measurement is only rarely an objective by itself. Broadly speaking, researchers are interested in measuring concepts because they want to explore relations between two or more concepts. For instance, they may be interested in investigating the link between economic growth and poverty alleviation. Therefore, they measure both concepts and use the outcome of the measurements to establish causal links between those concepts. However, most economists take measurements as given and do not investigate or discuss whether, and to what extent, inaccurate or imprecise measurement might affect their findings (Jerven, 2013b; Reiss, 2013; Woods, 2014). As we have seen in the previous sections, there are many reasons – such as data collection and data processing – why the outcome of the measurement $\hat{x}$, is not perfectly correlated with the unobservable, true value of the concept x. This section discusses how imperfect measurement can influence the quest for deriving causal relations.

Implicitly (and on rare occasions explicitly), economists assume that the measurement is accurate and that measurement errors occur randomly. Random error reduces the precision of the descriptive statistics and makes it harder to detect patterns in the data (Carroll et al., 2012). However, it does not affect the accuracy of the results. If the dataset only contains random error, and researchers still detect statistical significant relations using standard econometric approaches, it can be shown that this relation is 'real' and not the result of measurement error (Carroll et al., 2012). For instance, random measurement error in the dependent variables causes attenuation bias in linear regressions, that is, the coefficient of the noisily measured variable is biased towards zero. As such, the strength of the effect may be underestimated, but is never overestimated (Hyslop and Imbens, 2001).

However, measurement error can also be systematic. This occurs when the error is correlated with a characteristic of the measured concept or with the 'true value' of the measured concept. In contrast to random measurement error, systematic errors yield inaccurate measurements. Moreover, systematic errors can generate spurious correlations in the data (Imai and Yamamoto, 2010). For instance, larger households may systematically underreport food consumption, leading researchers to assume a spurious correlation between household size and food consumption (Gibson and Kim, 2007).

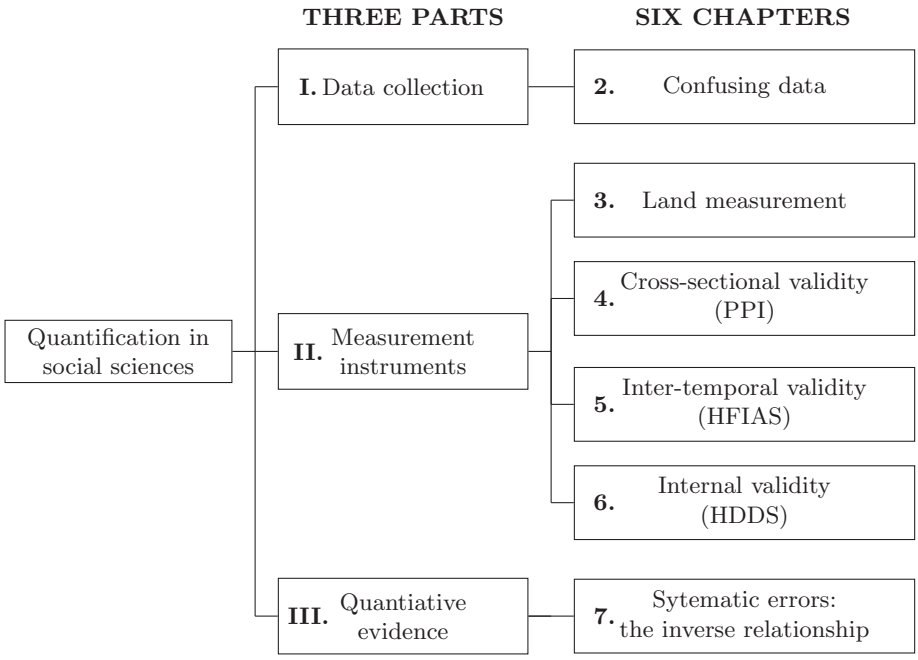
It is therefore important to investigate whether the way the concept is measured matters when establishing relations between concepts. The simplest approach is to examine if the correlations still hold if the concepts are measured with a second measurement instrument. This requires that every imperfectly measured concept is also measured with a second measurement instrument. The critical assumption here is that the measurement errors of both measurement instruments are orthogonal. In other words, it is assumed that the measurement instruments do not measure the concept imprecisely for the same reasons. In econometric terms, this corresponds to an instrumental variable approach (Carroll et al., 2012). In practice, however, a concept is often only measured once. This makes it difficult to assess the extent to which measurement error affects the results as the 'true' value of the mismeasured value is rarely known. Perhaps, this difficulty is one reason why systematic measurement error is only rarely discussed in academic papers. When discussing income and expenditure data, Chesher and Schluter (2002) have noted that: *"measurement error is an ever-present, generally significant, but usually neglected, feature of survey based income and expenditure data"*.

Given that concepts are rarely measured twice, other, second-best, approaches are needed to test the sensitivity of results to random and systematic measurement error. Second best approaches demand creativity and a thorough understanding of the dataset. The simplest, and most widespread approach, is to use the variables in the dataset that are the least likely to be incorrectly or noisily measured. If one believes that those variables correctly represent reality, the final information is also likely to be accurate. Another approach uses complex econometric models to handle measurement error. However, these models often require context-specific assumptions about the distribution of the measurement error, that are hard to satisfy (Fuller, 2009; Stefanski, 2000). This is, perhaps, another reason why systematic measurement error is often ignored in practical applications (Blackwell et al., In press).

1.4 Outline of the thesis

This PhD examines how raw data are transformed into measurements and measurements into quantitative knowledge. More specifically, the research addresses three related questions: (i) how is data collected? (ii) how can we quantify concepts? and (iii) given that concepts are always imperfectly measured, how do these imperfections affect causal links between concepts? This thesis is paper-based, but the different papers are related and structured into three parts that correspond to the three research questions. Each part consists of different case studies, structured as chapters (see figure 1.3 for an overview). These case studies stand alone and the case-specific findings have relevance that goes beyond the overarching theme of this thesis. In addition, the three parts of the thesis contribute to three different strands of academic literature.

Figure 1.3: Outline of the thesis



The first part, data collection, focuses on the challenging process of data collection. In the first case study agricultural yields in Rwanda are estimated from different data sources, leading to very different results **(chapter 2)**. I attempted to reconcile these different estimates and argue that difficulties of data collection combined with the desire to show a positive impact of large scale agricultural reforms on yields may explain the discrepancies. This illustrates why careful data collection is an essential first step in obtaining reliable measurements. Moreover, in the case of Rwanda, there is a risk that these 'wrong' numbers will become accepted facts as they become embedded within the national and international system of data management. This case study contributes to the small, but growing literature about data quality in Sub-Saharan Africa. Most of the literature in this field currently focuses on GDP, while this case study focuses on the agricultural sector.

The second part, measurement instruments, reflects upon the hazardous task of developing precise and accurate measurement instruments. In this part, I assess the validity, accuracy and precision of several existing measurement instruments. I used two different strategies to evaluate measurement instruments. First, to test the validity and accuracy of an instrument, I compared the result of the measurement with its gold standard or, at least, with another proxy of the concept.

Second, to test the robustness of an instrument to errors caused by changing circumstances, I tested the same instruments under different circumstances. In other words, referring back to equation 1.2 in section 1.3.1, I varied ΔOC as much as possible.

Chapter 3 is a classic example of the first strategy. It compares crop area measurement done with GPS against the gold standard, the compass and rope method. It concludes that GPS accurately measures land area, but the precision increases with plot size. This seemingly simple question has haunted statistical offices in many developing countries.

The other three case studies in this part assess indicators of poverty and food security, which are multidimensional concepts, which are not directly observable. These studies contribute to the literature on poverty and food security indicators. The different case studies illustrate the cross-sectional, inter-temporal and internal validity of indicators. **Chapter 4** examines the robustness of the Progress Out of Poverty Indicator (PPI) under changing circumstances. It turns out that this poverty indicator is cross-sectionally valid and remains a useful tool even when used in circumstances that differ substantially from the context in which it was initially developed. **Chapter 5** assesses the Household Food Insecurity Access Scale (HFIAS). Using food production as a gold standard, it shows that this food security indicator can be used to assess food insecurity over the same time period, but cannot be used to track food security over time. This questions the inter-temporal validity of the indicator. **Chapter 6** deals with yet another indicator of food security, the Household Dietary Diversity Score (HDDS). Borrowing a methodology from psychometrics to study internal validity, Rasch analysis is applied to demonstrate that the HDDS does not accurately measure the concept of food security.

In the third part, quantitative evidence, it is accepted that concepts are always imperfectly measured. This part consists of one case study (**chapter 7**) which examines the stylized fact of the inverse productivity-size relationship in Burundi. It shows how rounding errors, i.e. the tendency to report production numbers as a round numbers, strengthen the inverse productivity-size relationship. This illustrates that systematic errors can bias statistical analyses. This study not only offers a new (partial) explanation for the inverse productivity-size relationship, but also contributes to the small literature that studies systematic measurement error.

1.5 Overview of the different datasets

Table 1.1 provides an overview of the five different household survey datasets used in this thesis and outlines the main similarities and differences between them.

Further information is provided in the relevant chapters. All the surveys were administered recently and focused on poverty, food security and the agricultural production of rural households in developing countries. Data were collected by door-to-door interviews by trained enumerators. Two datasets are from Burundi, two from Rwanda and one from a development programme in Colombia and Ecuador.

All the surveys adopted a random (stratified) sampling design, but the specific sampling procedures depended on the reasons for collecting the data. Three surveys (two in Rwanda and one in Burundi) are representative of the (rural) population and followed a two-stage stratified cluster design. These surveys were set up by National Statistical Offices with (technical) support from international donors. Because these datasets are meant to be nationally representative, their sample size is relatively large. For example, the household survey conducted in Rwanda in 2010/11 included more than 14 000 households and was representative at the district level.

The household surveys in Rwanda, known by their French acronym EICV, were developed to monitor living conditions and poverty and are conducted every five years. The flagship reports of the Government of Rwanda about poverty reduction are based on these datasets (GoR, 2012b, 2015c). In this thesis, data from the second and third round (2005/06 and 2010/11, respectively) of the survey are used. Households were visited several times and data on socio-economic household characteristics, household assets, living conditions, agricultural production and employment were collected. Of all the datasets used in this thesis, the household surveys in Rwanda are the only ones that are publicly available. They can be downloaded from the website of the National Institute of Statistics of Rwanda (NISR). The data from the household surveys in Rwanda are analyzed in chapters 2 and 4.

The survey from Burundi, known as the ENAB survey, is an agricultural survey conducted in 2010/11 which focused on accurately measuring crop area and food production. It was the first nationally representative survey in Burundi since the 1970s and was carried out with the aim of updating agricultural statistics and national accounts. Households were visited at least once during each of the three agricultural seasons. Detailed production data was collected, by crop, at plot level, and the size of more than 50 000 plots was measured. In addition to gathering detailed production data, the survey also collected some basic information on socio-economic household characteristics and living conditions. This survey was a pilot project and one of its objectives was to build statistical capacity at the National Institute of Statistics in Burundi (Isteebu). Training enumerators, developing questionnaires and an appropriate sampling design were integral parts of the project. To capitalize on the experience acquired during this pilot project, the Ministry of Agriculture of Burundi aims to conduct an annual agricultural survey to monitor food production. Similar agricultural surveys, with a revised design, were conducted in 2012/13 and 2013/14. BTC, the Belgian Technical Cooper-

ation, partly funded the data collection in 2011/12 and allowed a joint research project, between Ghent University, the University of Burundi and Isabu (Institut des Sciences Agronomiques du Burundi), access to this dataset. The data from the agricultural survey in Burundi is used in chapters 3 and 7. Research, based on this dataset, is still ongoing.

The dataset collected in Ngozi, a province in the north of Burundi, is the only panel dataset used in this thesis. The sample size is relatively small ($n = 340$). The data was collected in the framework of a VLIR (Flemish Inter-university Council) Own Initiative project on food security dynamics in densely-populated regions in the north of the country. It is also the only dataset that was explicitly set up for research purposes. Data collection was primarily coordinated by Sanctus Niragira and, to a lesser extent, by Professor Marijke D'Haese and Sam Desiere. This dataset is used to test the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale in chapter 5.

The dataset from Colombia and Ecuador was collected as baseline data for the evaluation of a development project targeting small-scale coffee producers. Data was collected by CIAT, the International Center of Tropical Agriculture (based in Cali, Colombia), in collaboration with the Catholic Relief Service (CRS). Wytse Vellema, a researcher at the Department of Agricultural Economics at Ghent University, was involved in the design of the questionnaire and the field work. Power calculations ensured a sufficient number of observations in two groups (beneficiaries of the project and a control group) to detect a relatively small impact of the development program on household income (for details see Vellema et al. (2015)). The development programme aimed to improve food security and household income and, as such, data on income, household characteristics and main sources of income, and food security was collected. This data is used to verify the validity of the Household Dietary Diversity Score in chapter 6.

For the sake of conciseness, the questionnaires used in the different surveys are not included in the appendix, but are available upon request.

The second chapter of this thesis also uses macro-level data from the FAO and from an agricultural survey conducted in Rwanda in 2012/13. These datasets are not included in the overview table because I did not have access to the micro-level household survey data. These datasets are discussed in more detail in chapter 2.

Table 1.1: Overview of the different datasets used in this thesis by chapter

Dataset	Chapter	Country	Year	Purpose of data collection	N	Sampling design	Collected by
EICV 2	2,4	Rwanda	2005/06	Monitoring poverty and living conditions	6900	Representative of Rwanda's population (stratification at district level, clustering at village level)	NISR
EICV 3	2,4	Rwanda	2010/11	Monitoring poverty and living conditions	14 308	Representative of Rwanda's population (stratification at district level, clustering at village level)	NISR
ENAB	3, 7	Burundi	2011/12	Updating agricultural statistics (to compile national accounts)	2560	Representative of Burundi's population (statification at provincial level, clustering at district level)	Statistical Office of Burundi, Ministry of Agriculture
Survey Ngozi (Burundi)	5	Burundi	2007 and 2012 (panel data)	Food security dynamics in densely populated provinces	340	4 households randomly selected from 10 randomly selected villages in each of the 9 communes in the province of Ngozi	Ghent University, University of Burundi
Survey CIAT	6	Colombia/ Ecuador	2012	Baseline data for a development project targeting small-scale coffee producers	1015	Random selection from a list of all coffee producers in the region (stratified at municipal level)	CIAT and CRS

EICV: Enqûete Intégrale sur les Conditions de Vie
ENAB: Enqûete Nationale Agricole du Burundi
CIAT: International Center for Tropical Agriculture (based in Cali, Colombia)
CRS: Catholic Relief Service
NISR: National Institute of Statistics in Rwanda

CHAPTER 2

Agricultural reforms and yield growth in Rwanda: different data, different answers

Abstract: Statistics describe realities, but they also shape them, since they are used to design or support policies. As such accurate statistics are important. Using the agricultural sector in Rwanda as a case study, we demonstrate that dubious statistics can spread quickly. According to data from FAO, yields have increased by 60% since the implementation of large scale agricultural reforms, while other datasets point towards more modest gains. Yet, estimates in line with those of the FAO dominate the official discourse. We suggest that the discrepancies between datasets may be explained by the difficulties of collecting accurate agricultural statistics combined with an incentive to overestimate yields to show that the reforms have worked.

This chapter will be published as:
Desiere, S., Staelens, L., D'Haese, M., 2016. When the data sources writes the conclusion: evaluating agricultural policies. Journal of Development Studies.

2.1 Introduction

Do statistics describe realities or do they create them? These contrasting views are a recurrent theme in social sciences and public debates. On the one hand, data and statistics are considered to be objective observation of facts (Kuhn, 1961; Reiss, 2013). On the other hand, it is recognised that data and statistics are 'man-made' and as such, can be based on questionable assumptions, are shaped by the context in which they were generated and are prone to manipulation. This latter idea was aptly summarised by the Scottish poet Andrew Lang (1844-1912): *"some people use statistics like a drunk uses lamp-posts, more for support than illumination"*, a quote that was recently repeated by Romano Prodi, a former president of the European Commission (Carletto et al., 2013a). In his main work, French historian of statistics and sociologist, Desrosières (2002/1993), elaborates on these contrasting views. He refers to the double role of statistics as being both a social fact and referring to social facts. He argues that statistics and the context which shapes them are intimately linked. This perspective is shared by contemporary researchers in philosophy of science and economic history (Jerven, 2014a; Jerven and Johnston, 2015; Mensink, 2012; Morgan, 2001). In sum, statistics, independent of their evidence base, can become a reality in themselves.

Because statistics describe realities and, at the same time, shape realities there have been confusions and debates about 'truth' in many settings. Regions with limited capacity to assemble good quality data are arguably more vulnerable to the dissemination of biased statistics. Similarly, if collecting accurate data is challenging for technical reasons, the statistics that enter the public arena are more likely to be misleading and possibly biased. Agricultural statistics in Sub-Saharan Africa are a case in point. Although widely recognised to be of poor quality, they continue to shape policy debates and rural policies (Jerven, 2013b; Whitfield, 2012).

This paper shows that the lack of reliable agricultural data contributes to the risk of dubious statistics becoming part of reality. We illustrate this point by using the reporting on agricultural reforms in Rwanda as a case study. We used several datasets to compare agricultural yields in Rwanda before and after the implementation of the Crop Intensification Program (CIP) in 2007-2008. This programme is part of a wider set of policies implemented by the government of Rwanda (GoR) which aims to launch a Green Revolution. The main objective of the programme was, and continues to be, an increase in yields and food production.

The reforms were considered a great success by the government. Official documents and newspaper articles reported substantial improvements in yields of staple crops (Altazin, 2014; Kalibata and Roy, 2015). Moreover, a regional economic outlook produced by the IMF states *"as a result [of CIP], yields have increased significantly, from being among the lowest to among the highest in Sub-Saharan*

Africa". This statement includes a figure that shows an increase in cereal yields from slightly below 1000 kg/ha in 2007 to 2000 kg/ha in 2011 (IMF, 2013b, p.50). A World Bank report on Rwanda is equally confident about robust growth in its agricultural sector. It asserts that *"... between 2006 and 2011, the food outturn increased by 9.8 percent [per annum], almost double of the 5.4 percent between 2001 and 2006* (World Bank, 2013, p. 61) and attributes the acceleration in growth rates to the CIP. Yet, both reports fail to discuss the data and methodology behind the numbers. We attempted to replicate their findings.

As we will show in this paper, the increase in yields since the implementation of the agricultural reforms depends on the dataset used to evaluate it: it ranges from an impressive 60% to a modest 10% increase. We argue that it is not possible to make strong statements about the success or failure of the reforms in increasing yields. The problem is not a lack of data availability - the GoR undertook significant and laudable efforts to make their datasets publicly available - but rather that different data sources contradict each other and there is no way of telling which dataset is more reliable. Yet, it is only the figures that show the largest increase of yields that have been taken up in official discourses as illustrated above. Statistics may thus partially have created their own 'reality'.

It is important to note from the outset that this paper does not aim to evaluate the agricultural reforms in Rwanda. Such an evaluation requires a more comprehensive approach – in which an increase in yields and food production is only one aspect. Furthermore, this study does not have a counterfactual design. In other words, we do not know how yields would have evolved without the Crop Intensification Program. We do, however, occasionally refer to the 'impact of the agricultural reforms' when we simply compare yields before and after the implementation of the reforms since this terminology is also used in the official discourse. In no way do we claim to observe the causal impact of the reforms on yields. Rather than evaluating the reforms in Rwanda, this study focuses on the (lack of) quality of agricultural statistics and the risk of using them to support controversial policies. The case of Rwanda is used to demonstrate that this is a real threat.

This paper contributes to the small, but growing literature about data quality in Sub-Saharan Africa (Beegle et al., 2012; Jerven, 2014b; Jerven and Johnston, 2015). As elsewhere in the literature, we demonstrate that, besides data availability, data quality is a serious concern. Most of the literature has focused on the unreliable measurement of GDP (Jerven, 2013a, 2014b). We will focus on the agricultural sector, one of the key sectors in developing counties, where data limitations are likely to be even more severe than for other sectors (Carletto et al., 2015b). Bookkeeping in the agricultural sector is uncommon because of the subsistence nature of production and the high prevalence of illiteracy among farmers, while the unique mixed cropping systems pose a challenge to accurately measuring production. We will argue that the difficulties in data collection combined with political incentives to over-estimate production figures may explain the discrepancies

in yields between different datasets.

The paper is structured as follows. In the next section, we briefly describe recent agricultural reforms in Rwanda. We then present in detail the different datasets we draw upon in this study. Next, we outline our methodology and define the notion of overall yields, our preferred indicator of successful agrarian transformation. In the results section, we estimate overall yields from every dataset, followed by comparing the levels and trends of estimated yields from different datasets. In the discussion, we explore two potential explanations for the discrepancies between datasets: the challenges related to collecting agricultural statistics and the political economy of statistics. We conclude by formulating policy implications for Rwanda as well as for the broader community involved in collecting, processing and analysing agricultural data.

2.2 Agricultural policy in Rwanda

After the 1994 genocide in Rwanda, a technocratic government took power, which quickly restored relative stability and achieved rapid institutional reconstruction (Reyntjens, 2004). Moreover, it easily managed to attract development aid and Rwanda became one of the donor darlings in the region (Marysse et al., 2007). This effort led to a rapid recovery and economic growth averaging 8% in the last decade (Ansoms and Rostagno, 2012). Today, GDP per capita at PPP equals $1486 (IMF, 2013a).

Rwanda is an agriculture-based economy, and the agricultural sector employs more than 80% of the population, accounts for 39% of GDP and is the main earner of foreign exchange. Coffee and, to a lesser extent, tea and sugar cane are the main export products, with coffee accounting for 50% of foreign earnings (GoR, 2009, 2012a). However, the high population density, more than 300 inhabitants/km^2, has posed significant challenges to the agricultural sector for many decades (André and Platteau, 1998; Clay et al., 1995; Cochet, 2004; Verwimp, 2013). Competition for land is fierce and the average landholding per household is 0.76 ha, which is often dispersed with most households cultivating approximately four different plots. A quarter of households own less than 0.20 ha of land (GoR, 2010). Soil erosion poses additional and significant threats to soil fertility and undermines the already low levels of agricultural productivity. At the same time, few households have access to fertilisers or improved seeds.

Faced with these challenges, the government of Rwanda (GoR) set out the main priorities for the country's economic development in its ambitious Vision 2020 document. This aims at transforming Rwanda into a middle income country and shifting away from an agrarian to a knowledge based society by 2020. The development of a market-oriented agricultural sector was one of the main pillars

of the Vision 2020 document which stated that annual growth rates of 4.5 to 5% in the agricultural sector were essential to overcome poverty (GoR, 2000).

Subsequent official government reports further elaborated on the new agricultural proposals, culminating in the 'organic law determining the use and management of land in Rwanda' signed in 2005 (GoR, 2005). This law encouraged land consolidation with the aim of exploiting increasing returns to scale. For instance, it stipulated that a plot smaller than one hectare cannot be subdivided (GoR, 2005; Pottier, 2006). Moreover, it resumed a process of 'villagisation'. Initially, this policy forced households to abandon and demolish their houses if those houses were situated in areas devoted to agriculture and to rebuild them in the village (Pritchard, 2013). After internal and external protest, the policy was relaxed and nowadays only new houses need to be built within designated areas (Ansoms and Hilhorst, 2014). In addition, the law launched the land tenure regularisation programme that aimed to formally register the land of smallholder farmers to reinforce tenure security (Ali et al., 2014).

The Crop Intensification Program (CIP) was one of the flagship initiatives (GoR, 2015a). It aimed to increase productivity by increasing access to improved seeds and fertilisers to smallholder farmers. Additionally, the GoR selected priority crops and designated areas where those crops should be planted based on the agro-ecological conditions of the area. The underlying rationale is that specialisation, instead of mixed cropping systems, will increase yields, boost exports and facilitate mechanisation in the long term. Hence, farmers were encouraged to plant the same crops as their neighbours within a given area (GoR, 2012).

2.3 Data and methods

2.3.1 Data

The analyses are based on four datasets. All the datasets contain information about agricultural production, cropped area and yields during different periods from 2005 to 2013 in Rwanda, but differ considerably with regards to the purpose as well as the method of data collection. It is important to note that we have data from before and after the implementation of the agricultural reforms in 2007-2008 in Rwanda.

We grouped the datasets in three categories according to the methodology used: yearly estimates disseminated by the FAO, household surveys and agricultural surveys (table 2.1). The first group of datasets consists of yearly estimates of yields and cropped area of all major crops provided by FAOSTAT. These statistics are collected by FAO from national statistical offices and ministries of agriculture and disseminated through FAO's website. Data is available for most developing

countries, including Rwanda, since 1961. The FAO has no mandate to check the reliability of the figures, but simply disseminates the official national statistics (FAO, 2012a, 2014). The FAO confirmed that this is also the case for Rwanda and the FAO statistics simply reflect the official statistics from the Ministry of Agriculture of Rwanda. Hence, it would be more correct to refer to this data as 'statistics from the Ministry of Agriculture of Rwanda'. However as it is common practice to refer to them as 'FAO statistics' we will also do so in the remainder of this paper. It should nevertheless be kept in mind that if FAO statistics are unreliable it is because the national ministries reported wrong numbers.

The second group of datasets are household surveys. Household surveys follow a pre-defined sample design and collect information through door-to-door interviews. They form the backbone of statistical information in developing countries. The household surveys used in this study included a section on agricultural production and land. Both these key variables are based on recall by the household head. From this type of survey, we used two representative household surveys from Rwanda (EICV 2 and EICV 3). The household surveys in Rwanda, known by their French acronym EICV[1], are conducted every five years by the National Institute of Statistics of Rwanda to monitor poverty and living conditions (GoR, 2012c). The micro data are freely available from its website. EICV 2 commenced in October 2005 and continued till October 2006. The survey included 6900 households and followed a stratified cluster design (GoR, 2006). After removing households living in urban areas or with incomplete information on land or agricultural production, only 2225 observations remained. EICV 3 (October 2010 to October 2011) used a similar methodology and questionnaire as EICV 2, but the sample was much larger and representative at the district level. The survey contains 14 308 households of which we kept 8878 observations for further analysis (GoR, 2012a,b,c). In appendix 2.A, we discuss in detail the criteria used to discard observations in the datasets. Moreover, we provide evidence that the household characteristics of discarded households do not differ substantially from the included households, although they do differ with regards to their farm size. We argue that missing information on food production occurred randomly and that there is no reason to assume that we discarded or included households with the lowest or highest yields.

[1]EICV: Enquête Intégrale sur les Conditions de Vie

Table 2.1: Overview of the different datasets

Three methodologies	Four datasets	Period	Collected by	Purpose	Sample size
Yearly estimates	FAOSTAT/ Ministry of Agriculture	Yearly since 1961	FAO in collaboration with national statistical offices and ministries of agriculture	Global statistics to monitor worldwide trends	Time series of all major crops
Household surveys	EICV 2	Oct. 2005 - oct. 2006	National Institute of Statistics in Rwanda (NISR)	Monitoring poverty and living conditions	6900 (2225 valid observations)
	EICV 3	Oct. 2010 - Oct. 2011	National Institute of Statistics in Rwanda (NISR)	Monitoring poverty and living conditions	14 308 (8878 valid observations)
Agricultural survey	Agricultural survey Rwanda	Nov. 2012 Sept. 2013	National Institute of Statistics in Rwanda (NISR)	Comprehensive agricultural statistics for planning; compilation of national accounts	>15 000 (no microdata available)

The third group of datasets are the agricultural surveys. Agricultural surveys are set up to gather detailed data about agricultural production, land use and yields. In contrast to household surveys, they are more concerned with estimating total production than with household characteristics. This translates into a different sampling design which randomly sampled fields instead of farmers. Consequently, an agricultural survey is representative for the agricultural sector, while the household surveys are representative for the population. For instance, households with more land are more likely to be included in the agricultural survey, while the probability of being included in the household survey is independent of the size of the land cultivated by the household. As agricultural surveys focus on agriculture, much attention is paid to carefully measuring production and land. An agricultural survey was conducted by the National Institute of Statistics of Rwanda from November 2012 till November 2013. More than 15 000 farmers were interviewed during the three agricultural seasons. As the micro data from this survey are not publicly available, we relied upon the numbers reported in the main report of the survey (GoR, 2013).

2.3.2 Methods

As mentioned earlier, datasets were classified in three categories according to the methodology used (FAOSTAT statistics, household surveys and agricultural surveys). Each category relied upon different approaches to estimate agricultural production, cropped area and yields.

If we want to assess the increase in yields since the implementation of the agricultural reforms of 2007-2008, it is sufficient to study the trends in yields over time using a similar category of data. This is possible with data from FAOSTAT and the household surveys, for which we have data before and after the implementation of the reforms. It is not possible to examine trends in yields from the agricultural survey as we only have information for one point in time. However, we also want to compare levels of yields across datasets. Comparing levels is more troublesome than trends because it requires a certain degree of equivalence between the datasets. In other words, we assume that, notwithstanding the vastly different methodologies, the datasets measure the same underlying concept related to food production and yields. Only if this assumption holds can the levels of estimated yields be compared between different categories of datasets (Przeworski and Teune, 1966).

To compare levels and trends of yields, we need an indicator that summarises this information from the raw data. Our preferred indicator is 'overall yields', defined as total food production converted into its energy content per hectare. To get a familiar expression of yields, that is in kg/ha, we divided by the calorific content of beans, one of the main staple crops in Rwanda[2]. There are three equivalent ap-

[2]It is common practice to aggregate total food production by adding up the calorific values of all food crops. For instance, the well-known 'Daily per-capita energy supply' indicator of

proaches to define overall yields that correspond to the three categories of datasets defined earlier. They are formally presented by the following equations:

$$Overall\ yields = \sum_{i=1}^{14} \frac{cal_i}{cal_{beans}} * \frac{A_i}{A_T} * yield_i \quad (2.1)$$

$$= Median_j \left[\sum_{i=1}^{14} \frac{cal_i}{cal_{beans}} * \frac{production_{ij}}{A_{Tj}} \right] \quad (2.2)$$

$$= \sum_{i=1}^{14} \frac{cal_i}{cal_{beans}} * share_i * yield_{i,season\ B} \quad (2.3)$$

First, overall yields can be defined as a weighted sum of crop-specific yields ($yield_i$), weighted according to the energy content of crop i (cal_i) relative to beans (cal_{beans}) and the share of the crop in the total cropped area $\frac{A_i}{A_T}$. This is the definition used to analyse the data from FAOSTAT which reports on crop-specific yields and cropped area.

Household surveys estimate total harvest by crop as well as the total landholdings of every household. They do not estimate the share of land devoted to each crop because mixed cropping systems make this very cumbersome. The overall yield of a household, j, is then defined as total aggregate production ($production_{ij}$) expressed in its energy content and divided by landholdings of the household (A_{Tj}). As every farmer is unique, this calculation gives us a distribution of overall yields. This distribution is interesting by itself and will be discussed in depth. Overall yields at the national level are then defined as the median value of the distribution of yields. We opted for the median instead of the mean value of this distribution as a proxy of overall yields because the median is less susceptible to outliers than the mean.

The third equation is used to calculate yields based on data from the agricultural survey. This survey estimates crop-specific yields in every season and the share of land devoted to every crop ($share_i$). We defined overall yields as a weighted average of crop-specific yields in season B (March to July), because this season contributes the most to total, annual, food production.

FAO uses this approach (Smith, 1998). We then divided total aggregated production in its energy content by the energy content of beans, one of the main staple crops in Rwanda, to get a familiar expression of yields, that is, expressed in kg/ha. This normalisation facilitates interpretation of the results, but does not influence the findings. This approach is similar to the well-known conversion to cereal equivalents (Rask and Rask, 2014). A second, common approach to aggregate food production is to convert total production into monetary value. We did not opt for this approach because we did not have good price data. Moreover, this approach requires tricky assumptions about inflation and regional differences in price levels.

To ensure comparability of overall yields between datasets, we selected 14 food crops that were included in all the datasets. This selection included cereals (maize, millet, sorghum, rice, taro and wheat), roots and tubers (cassava, potatoes and sweet potatoes), pulses (beans, peanuts, peas and soybeans) and banana/plantains. According to FAOSTAT, these crops accounted for more than 80% of the total cropped area in Rwanda in 2013. Cash crops, e.g. coffee and tea, were not included in the analysis because we focused on food crops and because these were not included in the agricultural survey.

It is important to point out a subtle, yet important, difference in the interpretation of overall yields depending on the definition of 'land' area. We can use two definitions of land: *arable land* or *harvested land*. Arable land is defined as total land cultivated during an agricultural year. Total harvested land measures the total land area that has been harvested during a year. Hence, land that is harvested twice a year will be counted twice according to the definition of total harvested area, but will only be counted once according to the definition of arable land. The distinction between total arable land and total harvested area is important, because there are three cropping seasons during the agricultural year in Rwanda. The same plot of land is frequently harvested more than once a year. Hence, the total harvested area is greater than the total arable land. FAOSTAT and the agricultural survey reports the total harvested area of every crop. The household surveys, on the other hand, report total landholdings and, hence, total arable land. As a result, the household surveys may, if anything, overestimate overall yields. This subtle difference in the definition of land is one example why caution is warranted when comparing estimates of overall yields between different categories of datasets.

Overall yields are a fairly good instrument to capture the evolution of 'average' yields since the implementation of the reforms for three reasons. First, all stakeholders agree that increasing yields, rather than expanding arable land, is the only way to increase food production since most arable land is already under cultivation in this densely populated country. Second, our indicator of overall yields takes into account all major food crops, but gives more weight to crops that account for a larger share of total cropped area. An increase of yields of frequently cultivated crops, such as beans, therefore has a larger positive impact on overall yields than increasing yields of niche crops such as soybeans. For this reason, focusing on overall yields rather than on crop-specific yields gives a more accurate assessment of yield growth in the agricultural sector. Third, overall yields are easier to estimate than total food production. Estimating total food production with household surveys requires aggregating the data at national level. Such an aggregation requires accurate sampling weights. Furthermore, it requires that the survey is representative for the agricultural sector, which is not necessarily true since the survey is representative for the population. In addition, data aggregation is more sensitive to outliers in production numbers at household level (for

instance, due to data entry errors) than our definition of overall yields.

2.4 Results

2.4.1 Yearly estimates from FAOSTAT

Figure 2.1 compares the evolution of overall yields in Rwanda since 1990 based on data from FAOSTAT. It is striking that yields reported by FAO have increased tremendously in Rwanda since 2007-2008, which is generally considered as the start of the implementation of the Crop Intensification Program (GoR, 2012). According to the FAO overall yields in Rwanda have increased from 1253 kg/ha in 2007 to 2077 kg/ha in 2013. In other words, yields have increased by 66% in six years. This corresponds to an annual growth rate of 8.8%. This is a high growth rate, but only slightly higher than the average growth rates observed during the green revolution in Asia (Foster and Rosenzweig, 1996; Evenson and Gollin, 2003). Most of this increase occurred, however, from 2007 to 2011, and yields have only increased marginally since then. If these statistics are reliable, the claims that the agricultural reforms in Rwanda are extremely successful are justified.

Figure 2.1: Overall yields in Rwanda since 1990

Source: FAOSTAT and own calculations.

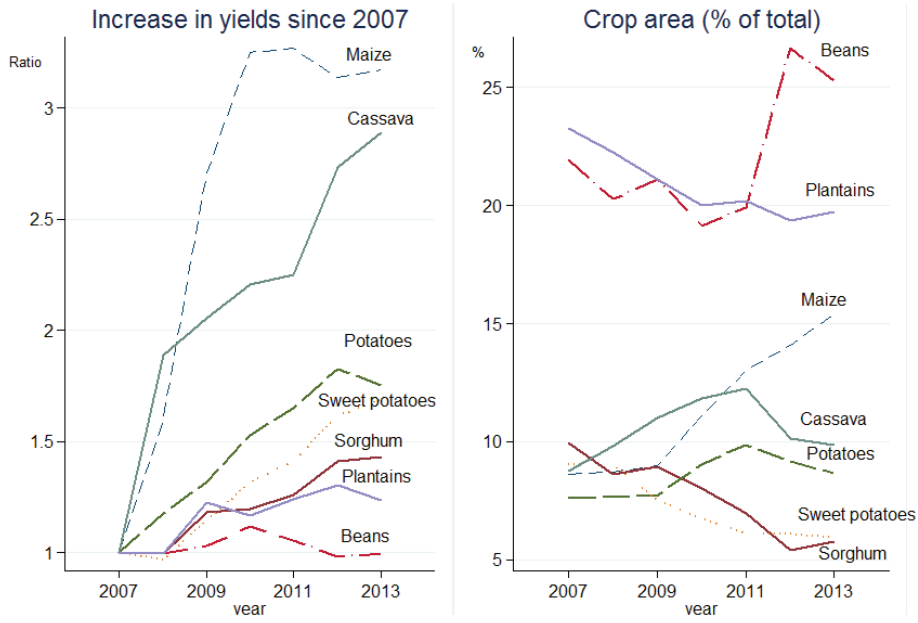
Examining the FAOSTAT data in more detail reveals that the increase in overall yields is driven by two complementary reasons. The most important one is the increase in yields of all crops since 2007 (figure 2.2, left panel). The increase in yields is most pronounced for maize, for which yields have more than tripled since 2005, and for cassava, for which yields have nearly tripled. This corresponds to an annual growth rate of 20%, which seems high. The strong yield growth of cassava is all the more surprising since the cassava mosaic disease occurs frequently in Rwanda and can severely reduce cassava harvests (Night et al., 2011). Cassava and maize are designated as 'priority crops' by the government of Rwanda (GoR, 2012c). Excluding cassava when estimating trends in overall yields, reduces the increase in overall yields from 66% to 42%. Excluding both cassava and maize from the estimation, reduces this increase to only 30%. Consequently, the sharp increase in overall yields since 2007 is mainly driven by strong yield growth of cassava and maize. In contrast, yields of beans, the most important staple crop in Rwanda and also one of the priority crops, remained constant. The second, and less important, explanation behind the increase in overall yields, is a shift over time in cropped area towards production of those crops with the greatest increase in yields (figure 2.2, right panel)[3]. For instance, land cultivated with maize accounted for 8% of the total cropped area in 2007 and this almost doubled to 15% in 2013. The share of land devoted to cassava increased slightly from 2007 to 2013. At the same time, land cultivated with sorghum, sweet potatoes and plantains has decreased over time. Only the evolution of land allocated to beans does not follow this trend since yields of beans remained constant, while land cropped with beans increased from 22% in 2007 to 25% in 2013.

It is interesting to compare the evolution of yields and cropped area between 'priority' and 'non-priority' crops. As part of its agricultural reforms, the government selected six priority crops, namely beans, cassava, potatoes, maize, wheat and rice (GoR, 2012c, 2015a)[4]. On average the yields of priority crops have more than doubled since 2007, while yields of non-priority crops have increased by 40%. Similarly, FAOSTAT statistics show a shift in cropped area from non-priority crops towards priority crops. Priority crops accounted for 50% of total cropped area in 2007 and 62% of total cropped area in 2013.

[3]To evaluate which of these two factors (i.e. an increase in crop-specific yields or a shift towards crops with the largest increase in yields), contributed most to the total increase in yields, we calculated overall yields in 2013 keeping the share of land devoted to each crop constant at 2007 levels. This calculation revealed that yields still increased by 60% from 1253 kg/ha to 2022 kg/ha. Hence, the increase in crop-specific yields is by far the most important factor explaining the increase of overall yields.

[4]Some sources also consider soybeans as a priority crop (GoR, 2012c). Including soybeans as a priority crop does not change the results because soybeans only accounted for 2% of total cropped area in 2013. Cash crops, such as coffee, are also priority crops.

Figure 2.2: Increase in yields since 2007 (left panel) and share of land devoted to each crop (right panel)



Only crops cultivated on more than 5% of total cropped area in 2013 are included in the figures. Left panels shows the ratio of yields per year to yields in 2007 by crop
Source: FAOSTAT and own calculations.

2.4.2 Household surveys

Table 2.2 shows median overall yields in Rwanda in 2006 and in 2011 estimated with household survey data. Overall yields in Rwanda in 2011, three to four years after the implementation of the reforms, were only 20% greater than yields in Rwanda in 2006, just before the implementation of the reforms. Household surveys thus point to a more modest increase in yields since the agricultural reforms than FAOSTAT-estimates. In sum, the success of the agricultural reforms depends on the data used to evaluate it.

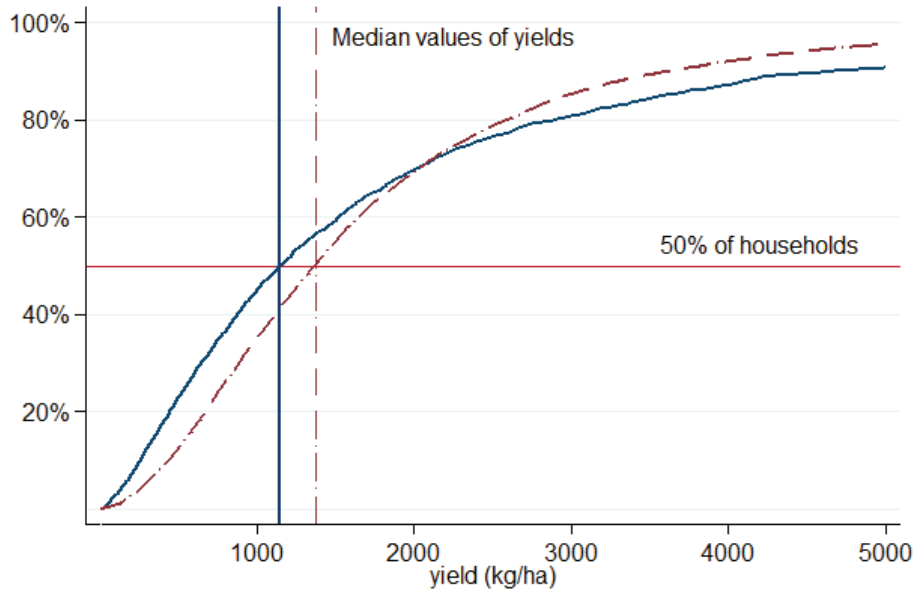
Table 2.2: Estimates of overall yields based on household surveys

	Rwanda (2006)	Rwanda (2011)
Median yields (kg/ha)	1140	1370
Number of observations	2225	8878

Source: Based on own calculations from EICV 2 and EICV 3.

Of the three categories of datasets, only for the household surveys did we have access to the underlying micro data. This allowed us to examine the distribution of yields between households. Figure 2.3 shows the cumulative distribution of yields of Rwanda in 2006 and 2011 and indicates the median values of the two distributions. The cumulative distributions are remarkable similar. An important feature of the distribution of yields is the enormous variation between farmers. For instance, our results show that 10% of the households in Rwanda in 2011 reported yields of less than 450 kg/ha, while another 10% of the households reported yields greater than 3600 kg/ha. This makes it extremely difficult to determine the 'representative' yield in Rwanda and explains why we preferred median rather than mean yields as the best proxy of overall yields in the country.

Figure 2.3: Cumulative distribution of yields in Rwanda in 2006 (blue, solid curve) and in 2011 (red, dashed curve) based on household survey data



Source: EICV 2 and EICV 3 and own calculations.
Vertical lines indicate median values of yields, horizontal line corresponds with 50% of the households.

Part of this huge variation is undoubtedly due to measurement errors in both production numbers and cropped area, which were based on farmers' estimates. Nevertheless, it probably also represents part of an agricultural reality. Yields are known to fluctuate significantly because of weather conditions, regional differences in soil quality and differences in inputs of fertiliser and labour. The two other sources of data, that is FAOSTAT-statistics and the agricultural survey, also undoubtedly required making assumptions on the distribution of yields to determine

'average' yields. However, neither FAOSTAT nor the report of the agricultural survey documented the variability of yields or the assumptions made to deal with this variability. This is unfortunate because we need this information to estimate the accuracy (95% confidence intervals) of 'average' yields.

2.4.3 Agricultural survey

Using the estimates of crop-specific yields and cropped area in season B of 2013 reported in the final report of agricultural survey in Rwanda, overall yields were 1478 kg/ha (GoR, 2013, table 72, p. 64 & table 73, p.64). Overall yields in season A and C were 1270 kg/ha and 1344 kg/ha, respectively. Consequently, our estimate of overall yields of 1478 kg/ha is an upper bound of average 'annual' overall yields. This estimate is slightly higher than overall yields estimated with household survey data, but still well below the FAO-estimate of 2077 kg/ha.

Table 2.3: Differences in yields and cropped area between FAOSTAT and the agricultural survey

	Yields (kg/ha)			Share of land (%)		
	Agricultural survey	FAO	Ratio	Agricultural survey	FAO	Ratio
Bananas[1]	8465	9223	1.09	17.8	23.11	1.30
Beans	853	913	1.07	17.4	25.30	1.45
Cassava	3176	15 766	4.96	15.9	9.86	0.62
Maize	1712	2285	1.33	5.5	15.41	2.80
Potatoes	9709	13 606	1.40	3.7	8.68	2.35
Sorghum	1355	1443	1.07	14.6	5.75	0.39
Sweet potatoes	8147	9616	1.18	9.7	5.93	0.61

Source: Report agricultural survey 2013, season B (GoR, 2013, table 72, p. 64 & table 73, p.64) and FAOSTAT.
Only crops accounting for more than 5% of total cropped area (using FAO-estimates) in 2013 are reported.
[1]The agricultural survey makes a distinction between bananas for cooking, beer or fruit, while FAO makes no such distinction. We reported yields of 'banana for cooking'.

The reasons behind the discrepancy between both estimates are explored in table 2.3. The table directly compares crop-specific yields and cropped area as reported by FAO and the agricultural survey. FAO-based estimates of yields of all crops are greater than those reported by the agricultural survey. This is especially the case for cassava, for which yields differ by a factor of five. Yields of maize and potatoes are also substantially greater according to FAOSTAT, by 33% and 40% respectively. The discrepancy between overall yields estimated by the FAO (2077 kg/ha) or the agricultural survey (1478 kg/ha) is mainly caused by three crops: cassava, maize and sweet potatoes. Comparing cropped area of every crop as share of total cropped area confirms the differences between the two data sources. According to the FAO, for instance, maize was grown on 15% of cropped area, while according to the agricultural survey maize accounted for 5.5% of total

cropped area in season B. Perhaps, the differences in cropped area can partially be attributed to the FAO numbers being yearly estimates, while the reported numbers of the agricultural survey only refer to season B.

2.5 Comparing agricultural yields in Rwanda between datasets and over time

Table 2.4 summarises the main findings of this paper. It shows estimates of 'overall' yields, which take into account all the important crops in Rwanda weighted according to their share of total cropped area. We estimated overall yields from three different data sources: yearly FAOSTAT estimates, household surveys and an agricultural survey.

The evolution of yields since the implementation of the agricultural reforms in 2007 can be assessed using these estimates of overall yields. The increase in yields is very different according to the dataset, albeit always positive. According to the FAO, yields increased by 55% between 2006 and 2011, while household surveys point to a 20% increase over the same period.

Comparing estimates of overall yields between datasets using very different methodologies requires more care, because the different data sources used different methodologies to measure a same underlying concept. Nevertheless, a comparison of overall yields between data sources can provide us with additional insights.

Table 2.4: Overall yields (kg/ha) in Rwanda estimated with different data sources

Year	Yields (kg/ha)		
	FAO	Household surveys	Agricultural survey[1]
2006	1306	1140	
2011	2029	1370	
2013	2077		1478

[1]only season B (March to end of July)

This comparison reveals large discrepancies of overall yields between different datasets. Estimates based on FAO-statistics and household surveys in Rwanda were rather similar in 2006. In 2011, however, overall yields in Rwanda estimated with FAO-statistics and household surveys differed by 50%. The agricultural survey tends to confirm the estimates from the household surveys as more realistic. This suggests that the estimates of the FAO have been too optimistic since 2007, when the agricultural reforms were introduced.

Finally, we examined whether other publicly available statistics about Rwanda can be reconciled with our finding of weaker yield growth than predicted by the

FAO. More specifically, we looked at growth in fertilizer application, trends in food imports and the evolution of poverty rates. None of these statistics in itself can rule out a 66% increase in overall yields as predicted by the FAO. Taken together, however, they do tend to suggest that FAO's prediction is too optimistic.

The Government of Rwanda points towards the strong increase in fertilizer imports since 2007 to explain strong yield growth. According to official statistics, fertilizer imports quadrupled from 8000 tonnes prior to the implementation of CIP to 35 000 tonnes in 2012 (Monitor Group, 2013; GoR, 2014, 2015b). Such a strong increase in fertilizer use could indeed explain the strong yield growth as reported by the official statistics. These figures are, however, contradicted by estimates based on the household surveys and the agricultural survey. The household surveys show that the number of households applying fertilizer increased from 16% in 2006 to 34% in 2011. The agricultural survey tends to confirm these statistics. Detailed results are provided in appendix 2.B. It is striking that the different data sources are internally consistent. If the official figures of increasing fertilizers imports are reliable, yield growth predicted by FAOSTAT seems plausible. If, on the other hand, fertilizer use reported by the household surveys and the agricultural survey are considered the most reliable data sources, FAOSTAT's prediction of yield growth seems misguided. Hence, depending on the data source, statistics on fertilizers use confirm or refute our finding of less impressive yield growth than predicted by the FAO.

A second, indirect, approach to check the reliability of our findings is by examining trends in food imports in Rwanda. One would expect decreasing net food imports with increasing food production. This is not confirmed by FAOSTAT statistics. Imports of cereals, particularly maize, decreased from 133 000 tonnes in 2007 to 77 000 tonnes in 2008, but have recovered rapidly since 2008, reaching 238 000 tonnes in 2011. These figures are difficult to reconcile with the sharp increase in food production as predicted by the FAO. This suggests that FAOSTAT may overestimate total food production, although other factors such as population growth and economic growth may also partially explain growing food imports.

The GoR is often praised by international donors for its sharp reduction in poverty rates. Poverty decreased from 57% in 2006 to 45% in 2011 (Ansoms and Rostagno, 2012; GoR, 2012; Desiere et al., 2015c). These figures were estimated with the household surveys EICV 2 and EICV 3 that were also used in this study (see chapter 4 for more details on living conditions and poverty in rural Rwanda). Although we cannot tell whether these poverty estimates are reliable, we can examine if this reduction can be reconciled with our estimate of yield growth. It is well-established that growth in the agricultural sector significantly reduces poverty. Typically, one finds that a 1% increase in yields decreases poverty by between 0.5% and 2% (Datt and Ravallion, 1998; Irz et al., 2001; Thirtle et al., 2003). Consequently, a modest increase of yields of between 10% and 40% between 2006 and 2011 in Rwanda is already sufficient to decrease poverty by 20%. Hence,

our estimate of yield growth based on the household surveys (+20%) does not rule out a poverty reduction from 57% in 2006 to 45% in 2011 as is claimed by the GoR (GoR, 2012b).

2.6 Discussion

This study does not take a definitive stance on the success or failure of the agricultural reforms in Rwanda. The results are inconclusive, and our findings can easily be criticised by arguing that household surveys are just not well-adapted to measuring yields. This is indeed partly true. Yet, we believe that there is no reason to assume that FAOSTAT statistics are any closer to the 'truth' than household surveys or the agricultural survey. What we intend to show in this study is that different datasets lead to different conclusions, and this raises several questions.

A first pressing question is why the FAO-numbers, which represent the official statistics of the GoR, are probably overestimating yields in Rwanda. There are two possible explanations for this: the challenges related to collecting accurate agricultural statistics and the political economy of agricultural data.

Even with dedicated agricultural surveys in the best performing agricultural statistical offices in Africa, collecting reliable agricultural statistics is still challenging (Jerven, 2013a; Vandercasteelen et al., 2013). The main reason is the predominance of subsistence agriculture in Africa. This limits the need for bookkeeping and explains why estimates of production in surveys often relies on recall by the household head (Beegle et al., 2012). This can cause inaccurate numbers. Deininger et al. (2012), for instance, reports that recall underestimates production by 40% compared to record keeping in diaries. Moreover, mismeasurement may be more pronounced for some crops than for others. It is, for instance, well-known that obtaining reliable production numbers for roots and tubers, such as cassava, is especially difficult. In contrast to high-value crops, cassava is harvested in small quantities over several months because it stores better in the ground. In addition, cassava is often only fully harvested during times of food crisis (Carletto et al., 2015b). Perhaps, these difficulties explain why FAOSTAT-estimates of yields of cassava (15 ton/ha) are five times greater than those of the agricultural survey (3 ton/ha). An additional challenge in Rwanda is the fact that most crops are grown in mixed cropping systems. This makes it extremely difficult to accurately estimate the share of land devoted to each crop (Fermont and Benson, 2011).

Although gathering reliable data is difficult, this does not yet explain why, according to FAO, yields have increased substantially since 2007. As these statistics were provided by the GoR, the sharp increase in yields since 2007 may be explained by a political economy argument (Sandefur and Glassman, 2015). The increase in yields coincides with the implementation of agricultural reforms in Rwanda

and officials may have had an incentive to overestimate agricultural production to demonstrate that their reforms were working. For instance, the increase of yields was the largest for maize (+200% since 2007), which is one of the 'priority' crops of the government. More generally, we found that yields of priority crops have increased more than yields of non-priority crops. Local officials in Rwanda are bound by performance contracts that specify development targets and are set by the national government in line with national policies (Ansoms, 2008a; Ingelaere, 2010). Local officials who do not succeed in achieving their targets miss out on promotions and may even get fired for below average performance (Versailles, 2012). This provides a strong incentive to tweak the numbers.

Yet, this raises a new question: why are yields of the main staple crops reported in the agricultural survey, which was conducted by the National Institute of Statistics of Rwanda in collaboration with the Ministry of Agriculture, much lower and probably more realistic? Why did officials charged with data collection not overestimate yields in this case? The answer may lie in the way the data were collected which determines how easily numbers can be 'negotiated'. As Jerven (2014b) argues, when the empirical evidence is weak, there is ample room for a negotiation about agricultural data. Although we have no proof, FAO's numbers are likely to be based on eye-estimates by local extension officers of cropped area and total harvest, which is common practice in many countries (Carletto et al., 2015b). Eye-estimates are known to be very inaccurate and can, therefore, more easily be tweaked to satisfy political objectives. It is not even necessary that this manipulation occurs consciously, it is already sufficient that officials in charge of data reporting simply believe that the agricultural reforms work and, hence, overstate production numbers. For instance, as the import of fertilisers was widely reported to have surged because of the CIP (by more than 32% per annum by one account), officials may have expected a substantial increase in yields and food production (Druilhe and Barreiro-Hurlé, 2012; Monitor Group, 2013). An agricultural survey, on the other hand, follows a pre-defined, 'scientific' design and is considered to be the gold standard for collecting reliable agricultural data (Fermont and Benson, 2011). As a result, numbers from agricultural surveys may be more 'trusted' and are less susceptible to (unconscious) manipulation. A better understanding into the interplay between 'trust in data quality' and political pressure to use numbers to prove that policies are working is an interesting avenue for further research.

2.7 Conclusion

Our findings have several implications for policymakers in Rwanda and all actors involved in collecting, processing and analysing agricultural data. First, a careful evaluation of the impact of the agricultural reforms in Rwanda on yields remains important because an increase in yields and food production is the main objective of the programme. As these reforms have already been criticised for many

other reasons including their top-down approach, increasing social tensions in local communities and reducing tenure security and food security at household level (Ansoms, 2008b; Ansoms and McKay, 2010; Pritchard, 2013; Dawson et al., 2016), a strong, positive impact on yields and food production is required to justify the implementation of the program and to push the reforms even further. Second, agriculture still accounts for the lion's share of the national economy in Rwanda. Reliable agricultural data are thus a condition sine qua non for accurate national accounts, which, in turn, are important to monitor growth.

The reliability and accuracy of FAO-numbers and agricultural statistics in Africa have already been criticised by many authors (Devarajan, 2013; Jerven, 2013a), including by the FAO and the World Bank themselves (World Bank and United Nations and Food and Agricultural Organization, 2010). This is confirmed by this study, which suggests that FAO numbers in Rwanda are too optimistic and may even be plainly wrong. The danger is that these numbers, rather than those of the household survey or the agricultural survey, get embedded within the FAO's international system of data management and will be taken up over and over again for new analyses (e.g. in cross-country regressions, see Woods (2014)). In any case, we should ensure that statistics describe realities and avoid at all cost them becoming a reality on their own. One factor that augments this risk is the lack of clear documentation which provides all the necessary details about how, for which purpose and by whom the FAO numbers were collected. In this respect, we can only join the call of other researchers concerned about data quality to increase the transparency of the data collection process (Jerven, 2013a). Fortunately, several institutional initiatives are currently already under way to improve the quality of agricultural statistics in developing countries (Addinson et al., 2015; Chen et al., 2013; FAO, 2012b; World Bank and United Nations and Food and Agricultural Organization, 2010).

Appendix

2.A Selection criteria to discard observations from EICV 2 and EICV 3

In section 2.3.1 Data, we mentioned that we discarded many observations from the household surveys, EICV 2 and EICV 3. We did this for various reasons. This process reduced the sample size of EICV 2 from 6900 to 2225, and that of of EICV 3 from 14 308 to 8878. In this section, we carefully outline the reasons for discarding these observations and investigate whether there are substantial differences between excluded and included households in terms of farm and household characteristics and the geographic distribution of households within Rwanda.

We consecutively discarded the following household types: households living in urban areas; households with missing information on land or reporting they cultivate no land; and households without any information on food production or reporting zero production. Table 2.A.1 shows the number of discarded households for these reasons. Around 1% of the observations in EICV 2 were discarded because we had no information on land, while around 7% of the observations in EICV 3 were removed for the same reason. Missing information on total production occurred more frequently in the EICV 2 dataset (2% of observations) than in the EICV 3 (1% of the observations). Finally, we discarded households that reported cultivating a crop, without providing information about the amount harvested. For instance, a household that reported cultivating beans, but did not report how much beans they actually harvested was excluded from the final dataset. This means that we discarded a household if there was at least one missing variable in crop production. This was a strict criterion and the main reason why households were discarded. For this reason, 41% of the households in EICV 2 and 20% of the households in EICV 3 were discarded. We have no explanation why EICV 2 contained more missing production numbers than EICV 3.

Table 2.A.1: Discarding observations from EICV 2 and EICV 3

	EICV 2	EICV 3
Observations in initial sample	6900 (100%)	14308 (100%)
Observations discarded:	4675 (68%)	5430 (38%)
Urban areas	1620 (23%)	1437 (10%)
Land area is zero or missing	68 (1%)	954 (7%)
Total production is zero or missing	154 (2%)	115 (1%)
Production numbers of at least one crop missing	2833 (41%)	2924 (20%)
Observations in final sample	2225 (32%)	8878 (62%)

To check whether the households discarded from the datasets because of missing production numbers differed substantially from the households included in the final dataset, we conducted a missing data analysis. The aim was to see whether or not our final sample was biased. We investigated whether farm and household characteristics of the excluded households differed substantially from the included households. Also, we checked whether the spatial distribution of the households within Rwanda remained unchanged after excluding households.

Table 2.A.2 and 2.A.3 show the results for EICV 2 and EICV 3, respectively. In terms of household characteristics (household size, age of household head, gender of household head) there are no important differences between included and excluded households. With regards to farm size, however, the included households have smaller farms than excluded households. This difference is more pronounced for EICV 3 (difference of 1111 m^2) than for EICV 2 (difference of 424 m^2). A possible explanation is that larger farms are more likely to cultivate more different crops than smaller farms and as such the data are more likely to contain missing variables. It seems unlikely that the difference in land size between included and excluded households biases our estimates of average yields. The inverse-productivity size relationship also holds in Rwanda (Ali and Deininger, 2014; Ansoms et al., 2008), meaning that average yields are lower on larger than smaller farms. Hence, excluding more large farms than small farms is more likely to result in an overestimation of yields, in particular in 2011. The estimation of overall yields with household surveys in the main paper is thus an upper bound and does not alter our conclusions.

Table 2.A.2: Missing data analysis EICV 2

	No missing production data (included households)	Missing production data (excluded households)
Land (m^2)	8288	8712
Number of plots	3.71	3.94
Household size	4.95	5.15
Age household head	43.98	45.54
Female-headed household	28%	28%
Yield, kg/ha (mean)	2195	1770
Yield, kg/ha (median)	1140	825
n	2225	2833

At least one of the production numbers was missing for the excluded households. Yet, most of them reported production number of several other crops. Consequently, we can still estimate average yields on these farms. As expected, yields of the excluded farms were substantially lower than those of included farms as they did not report all their harvest. Including the excluded farms in our final sample would thus reduce overall yields in Rwanda.

Table 2.A.3: Missing data analysis EICV 3

	No missing production data (included households)	Missing production data (excluded households)
Land (m^2)	5857	6968
Number of plots	4.71	4.3
Household size	4.76	4.92
Age household head	46.09	45.08
Female-headed household	28%	27%
Yield (mean)	1874	,435
Yield (median)	1370	876
n	8878	2924

Table 2.A.4: Geographic distribution of households in EICV 2 (included versus excluded households)

Province	No missing production data (included households)	Missing production data (excluded households)	Total per province
Kigali City	69 (97%)	2 (3%)	71
Southern province	791 (56%)	620 (44%)	1411
Western province	536 (39%)	853 (61%)	1389
Northern province	428 (47%)	489 (53%)	917
Eastern Province	401 (32%)	869 (68%)	1270
n	2225 (44%)	2833 (56%)	5058

Table 2.A.5: Geographic distribution of households in EICV 3 (included versus excluded households)

Province	No missing production data (included households)	Missing production data (excluded households)	Total per province
Kigali City	126 (82%)	28 (18%)	154
Southern province	2913 (89%)	364 (11%)	3277
Western province	2409 (80%)	614 (20%)	3023
Northern province	1595 (72%)	632 (28%)	2227
Eastern Province	1835 (59%)	1286 (41%)	3121
n	8878 (75%)	2924 (25%)	11 802

Finally, we examined the geographic distribution of the households among provinces in Rwanda (tables 2.A.4 and 2.A.5). Comparing included and excluded households (because of missing observations for at least one crop), we note that we discarded more households in the Eastern province and included more households from the Southern province in both surveys. We cannot explain this trend.

Overall, we believe that the discarded observations are not substantially different from the included observations. Hence, we believe that our results are not influenced by the selection criteria for the inclusion or exclusion of households from the final dataset.

2.B Fertilizer use in Rwanda

One important and oft-emphasized strategy to increase yields is increasing fertilizers application. This was recognized by the GoR. The Crop Intensification Program aimed to substantially increase fertilizer use to restore soil fertility and increase productivity. To this end, it developed a market-oriented fertilizer distribution system, facilitated the import of fertilizers and subsidized fertilizers to make it accessible to small-scale farmers (GoR, 2014). The National Fertilizer Policy aims to increase fertilizer use to 45 kg/ha in 2017/18, which corresponds to 55 000 metric tonnes for the country(GoR, 2014). According to official government documents, import of fertilizer increased from 8000 metric tonnes prior to the implementation of CIP to 35 000 tonnes in 2012 (Monitor Group, 2013; GoR, 2014, 2015b). The same sources mention an increase in fertilizer use per hectare from 4 kg in 2007 to 30 kg in 2013. In addition, the website CountryStat Rwanda, which gathers agricultural data and is managed by the GoR, reports an increase in DAP imports from 99 tonnes in 2005 to 2910 tonnes in 2009, while the import of NPK increased from 7215 tonnes to 42 289 tonnes over the same period. The GoR frequently refers to the increased use of fertilizers as the main explanation for the strong yield growth.

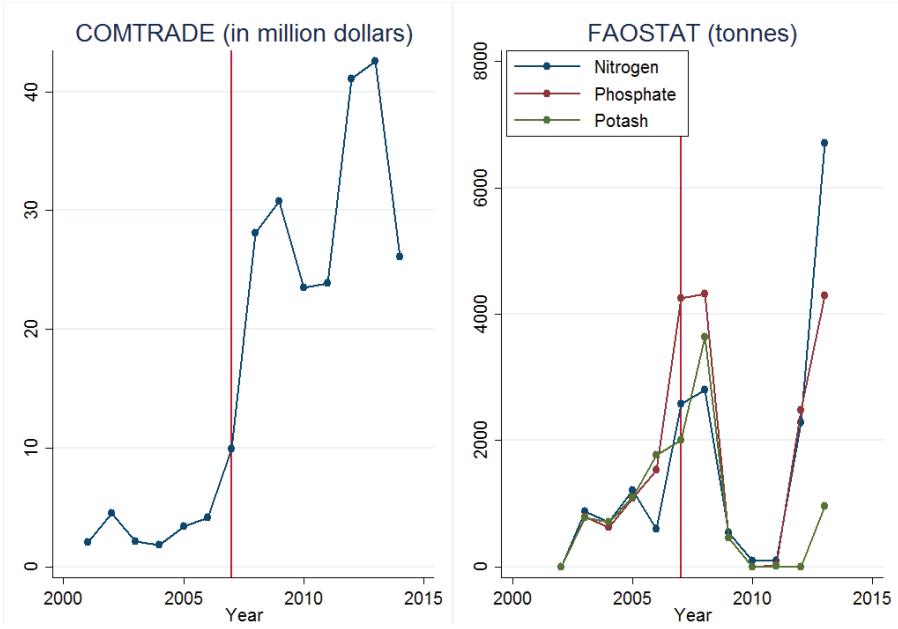
To check these official numbers, we again compared estimates from FAOSTAT (and COMTRADE), the EICV surveys and the agricultural survey. Again, we found some surprising anomalies between the datasets. The increase in fertilizer use is much more pronounced according to official sources (government documents, COMTRADE and, to some extent, FAOSTAT) than according to estimations derived from the household surveys and the agricultural surveys. This mirrors our main finding that official sources find stronger yield growth than estimates based on household and agricultural surveys. Overall, it seems that fertilizers imports increased less rapidly than predicted by the GoR or that the fertilizers did not reach the small-scale farmers included in the household and agricultural surveys.

FAOSTAT/COMTRADE

We relied upon two official sources to investigate macro-trends in fertilizer imports in Rwanda: the COMTRADE database and the FAOSTAT statistics about fertilizer imports. The COMTRADE statistics confirm the official figures: total import value of fertilizers increased from 4 million dollars in 2006 to 28 million dollars in 2008 (figure 2.B.1, left panel). The official FAOSTAT statistics show a very different picture of fertilizer imports (figure 2.B.1, right panel). FAOSTAT disaggregates fertilizers by its nutrient content. These statistics show that fertilizer imports increased from 2006 to 2008, decreased to very low levels from 2008 to 2011 and have increased tremendously since. As Rwanda imports all its fertilizers and exports only very small amounts, fertilizer imports equals fertilizer consumption (results not shown). The contrast between statistics from FAO and

COMTRADE are striking and impossible to reconcile. This is all the more surprising as a note attached to the FAOSTAT statistics states as source of the data 'Official data from questionnaires and/or national sources and/or COMTRADE (reporters)'.

Figure 2.B.1: Fertilizer import according to COMTRADE and FAOSTAT



Household surveys

The household surveys allow us to examine if access of small-scale farmers to fertilizers has improved since 2007. A single question was included in the household surveys that asked respondents to report annual expenditure on fertilizers. The problem here is that it is challenging to convert expenditure data into the amount of fertilizer purchased. First, we only have some price information in 2010/11, but not for 2005/06. According to AMITSA, which monitors fertilizers prices in the East African Community, the price of NPK and Urea varied between 300 and 400 RwF/kg from July 2010 to September 2011, while the price of DAP varied between 400 and 500 RwF/kg (Amitsa, 2011). We used the lowest price (300 RwF/kg) to convert expenditure data into the amount of fertilizer purchased in 2010/11 and thus overestimated fertilizer use. Second, we cannot simply compare expenditure data reported in EICV 2 and EICV 3 because of inflation, but also because the GoR has started to subsidize fertilizer since 2007. It is therefore likely that the real fertilizer price has decreased since 2007.

Yet, we can still investigate if the number of households that applied fertilizers

increased from 2005/06 to 2010/11. The number of households reporting having purchased fertilizers during the last year increased from 16% in 2005/06 to 34% in 2010/11 (table 2.B.1). This suggests that access to fertilizers improved, although the majority of the households does not yet have access to fertilizers

When converting expenditure data in 2010/11 in the quantity of fertilizer purchased, it turns out that average fertilizer application is 32 kg/ha. The distribution of fertilizers application is, however, highly skewed, with some households using almost all the fertilizers. For instance, in 2010/11 64% of the households did not use fertilizers, 11% used less than 14 kg/ha, 15% used between 14 kg/ha and 76 kg/ha and 5% of the households used more than 163 kg/ha. Similarly, only 7% of the households reported purchasing more than 50 kg of fertilizers in 2010/11. The average application of 32 kg/ha is therefore the result of a limited number of farmers using large amounts of fertilizers. In addition, this number is likely to be biased upwards because of large outliers due to measurement and data entry errors.

Although there is some evidence that fertilizer use has increased from 2005/06 to 2010/11, the increase seems too limited and too much benefiting a small number of households to have a substantial effect on overall yield growth in Rwanda.

Table 2.B.1: Fertilizers use at household level

	Rwanda (2006)	Rwanda (2011)
Applied chemical fertilizers (% of HH)	16%	34%
Fertilizers expenditure (RwF and (kg/ha)[1])		
Mean	1227	4354 (32 kg/ha)
Sd	5803	31 376 (118 kg/ha)
75th percentile	0	1750 (14 kg/ha)
90th percentile	1600	10 000 (76 kg/ha)
95th percentile	6000	19 080 (163 kg/ha)

[1]Only expenditures is 2011 ($300 Rwf = 1$ kg of fertilizer) can be converted into fertilizer application per hectare, because we have no price information for the period 2005/2006. See text for more details.

Agricultural survey

The report of the agricultural survey only discusses whether households use fertilizers in the three seasons, but does not report how much fertilizer they used. According to this survey, 17.3% of the households used chemical fertilizers in season B, which is the most important season in terms of production (GoR, 2013, table 78, p. 69). Fertilizer was used by 19.9% and 65.9% of the households in season A and C, respectively (GoR, 2013, season A, table 32, p. 36; season C, table 115, p. 96). We cannot explain why fertilizer use was much more widespread in season C than in the two other seasons.

According to the household survey (EICV 3), 34% of the households purchased

chemical fertilizer in 2010/11. It is difficult, however, to compare estimates from the agricultural survey with those from the household survey, because the former provides seasonal estimates and the latter only annual estimates. Both surveys, however, confirm that most small-scale farmers have no access to chemical fertilizers.

CHAPTER 3

Area measurement in agricultural surveys: GPS or compass and rope?

Abstract: Accurate and precise land area measurement is a critical issue in agricultural surveys. Compass and rope measurement has traditionally been considered as the gold standard. Nowadays, GPS devices are steadily replacing the traditional measurement method because GPS measurement is less time consuming and expensive. Yet, within agricultural statistical offices, there is still an active debate whether GPS devices are sufficiently precise and accurate to measure plots smaller than 0.5 hectares. This study settles this debate. It assessed land area measured with GPS against its gold standard, the compass and rope method, for more than 50 000 plots, most of them smaller than 0.5 hectares. On average, measurement with GPS is not much different from measurement with compass and rope, even for very small plots. The precision of GPS measurement increases, however, rapidly with plot size. Measurement is fairly precise on plots larger than 1000 m^2. When plot size is greater than 1000 m^2 (0.1 ha), more than 90% (80%) of GPS estimates differed by less than 10% (5%) relative to compass and rope estimates. For most practical purpose, GPS measurement is to be preferred over compass and rope measurement.

3.1 Introduction

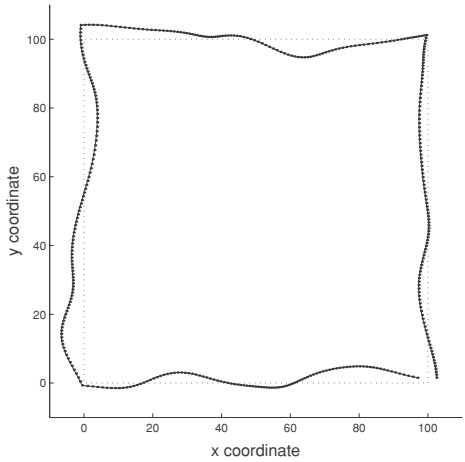
Accurate and precise measurement of land area is one of the key objectives of agricultural surveys. Historically, compass and rope methods were considered the gold standard to obtain precise area measurement (Diskin, 1999; FAO, 1982; Fermont and Benson, 2011). However, modern devices, i.e. Global Positioning Systems (GPS), are gradually replacing compass and rope methods because GPS measurement is easier and more cost-effective than traditional methods.

Although compass and rope measurement is the gold standard for precise and accurate area measurement, it needs to be implemented carefully in the field to avoid random measurement error (Carletto et al., 2014). It requires that enumerators measure the distance between subsequent corners of the plot as well as the angle between the sides of the plot. This is challenging if the plot is irregularly shaped, if the vegetation is dense or if it is difficult to clearly identify the corners of the plots. For these reasons, Carletto et al. (2014) argue that the accuracy and precision of compass and rope measurement can be problematic if the enumerators are not sufficiently trained or experienced in compass and rope measurement. GPS measurement, by contrast, is easier to implement in the field because the enumerator only has to pace the perimeter of the plot, which is also feasible if the plot is irregularly shaped. In addition, GPS measurement takes, on average, three times less time than compass and rope measurement (Carletto et al., 2014).

From a logistic point of view, GPS measurement is clearly preferable over compass and rope measurement. Moreover, previous research has shown that GPS measurement is as precise and accurate as compass an rope methods for plots larger than 0.5 ha (Bogaert et al., 2005; Fermont and Benson, 2011). Whether this is also the case for smaller plots remains an open question. Theoretical work has shown that the precision of GPS measurement decreases with plot size (Bogaert et al., 2005). GPS measurement uses GPS coordinates, estimated by satellites, to define the boundaries of the plot. These coordinates are not perfectly determined, which introduces measurement error. This error is small for every data point, but the errors of subsequent data points are serially correlated. As a consequence, the boundaries of the plot are not perfectly defined. This is illustrated in figure 3.1. Because the measurement error occurs at the plot boundaries, small measurement errors are relatively more important for small plots than larger ones. As a result, the precision of GPS decreases with plot size. As the serial correlation between data points occurs randomly around the 'true' coordinates, there is no reason to expect that GPS measurement is inaccurate, that is, systematically under or overestimates plot size. The question remains, however, below which threshold of plot size compass and rope measurement is too imprecise to measure land area.

Doubts with regards to this threshold haunts statistical offices in developing countries, where average plot size is often considerably smaller than 0.5 ha (Jayne et al.,

Figure 3.1: GPS data points are serially correlated, causing measurement error in plot size



Source: adapted from Bogaert et al. (2005)

2003). As far as we know, few studies have assessed this critical threshold and it seems as if the often cited threshold of 0.5 ha is not based on extensive empirical research. An experiment in Uganda, reported by Schøning (2005), found that GPS underestimates land size and noted a weak correlation between GPS and compass and rope estimates ($R = 12\%$) when the measured area was smaller than 0.5 ha. It is, however, not clear in their study how the threshold of 0.5 ha has been determined. Keita and Carfagna (2009) also reported a tendency of GPS to underestimate plot size, but did not detect a link between accuracy and plot size.

Recently, researchers at the World Bank conducted a study similar to ours that compared area measurement with GPS, compass and rope and self-reporting by the farmer (Carletto et al., 2014, 2015a). As we will discuss, their results are nearly identical to our findings, which confirms the robustness of the results. Our study can, nevertheless, complement their findings as our dataset contains many more observations (over 50 000 versus 1765) and more very small plots (90% of plots were smaller than 1000 m^2 versus 60% in the study of the World Bank).

We contribute to the literature by quantifying measurement error of GPS devices for small plots. To this end, we use exceptional data from an agricultural survey in Burundi that measured more than 50 000 plots of land with both GPS and compass and rope methods. Results show that GPS only slightly underestimates plot size relative to the compass and rope method. Moreover, the precision of area measurement with GPS increases rapidly with plot size. For most practical purposes GPS measurement is to be preferred over compass and rope methods.

3.2 Methods

When assessing the reliability of GPS to measure small areas, we considered measurement with compass and rope as the gold standard. Hence, compass and rope measurement was used as the benchmark against which GPS measurement was assessed. As discussed in the introduction, the compass and rope method does not measure plot size perfectly. Hence, if plot size measured with GPS and tape and compass differ, it is not necessarily the GPS measurement that is wrong. What is important, however, is that it is widely accepted that compass and rope measures plot size accurately and precisely, even for small plots (Fermont and Benson, 2011). Hence, if measurement with GPS is, on average, equal to measurement with GPS, we can conclude that GPS measurement in accurate. Similarly, if the absolute difference between GPS measurement and compass and rope measurement decreases with plot size, we can conclude that the precision of GPS increases with plot size.

We first examined whether measurement with GPS is unbiased. In other words, we examined whether GPS measurement tends to over or underestimate plot size. Therefore, we regressed plot size measured with GPS on the same area measured with compass and rope. Ideally, the constant in this regression is zero, while the correlation between measurement with GPS and measurement with compass and rope equals 1.

We then assessed the precision of GPS measurement and analysed whether precision decreases with plot size. One measure of precision is relative error defined as follows:

$$relative\ error = \frac{Area\ measured\ with\ GPS\ -\ Area\ measured\ with\ compass\ and\ rope}{Area\ measured\ with\ compass\ and\ rope} \tag{3.1}$$

Relative error can be negative (if GPS underestimates plot size) or positive (if GPS overestimates plot size). Note that a small over or underestimation of plot size with GPS causes a large relative error if the plot is small, while a similar over or underestimation only causes a small relative error if the plot is relatively large. Using a graphical approach developed by Bland and Altman (1986, 1995), we then plotted relative error in function of average plot size, obtained as the average of GPS and compass and rope measurement[1]. This approach allows us to examine

[1]The results remained similar when we plotted relative error versus plot size measured with compass and rope (the gold standard) rather than the average plot size. The reason is that, on average, plot size measured with GPS equals plot size measured with compass and rope. Hence, the average plot size equals plot size measurement with compass and rope (Bland and Altman, 1995).

whether the accuracy and/or the precision of GPS measurement is correlated with plot size.

Both over and underestimation of plot size has to be avoided. Hence, the absolute value of the relative error, referred to as absolute error, defines the precision of the measurement. To study the relation between the precision of GPS measurement and plot size, we calculated the number of plots with an absolute error smaller than 2.5%, 5% and 10% as a function of plot size.

3.3 Data

We used data from an agricultural survey conducted in 2011-2012 in Burundi by the Statistical Office of Burundi (République du Burundi, 2013a). This survey was administered during the three agricultural seasons to 2560 households and the size of every parcel was measured. A parcel was defined as a plot of land devoted to a unique crop or mixed cropping system. This explains why 90% of the parcels were smaller than 1000 m^2. One of the key objectives of the agricultural survey was accurately measuring land area. Because GPS measurement was considered insufficiently precise on small plots, enumerators had to measure every parcel with both GPS and compass and rope. Enumerators were trained in the use of GPS and compass and rope methods prior to the implementation of the survey. Once in the field, enumerators first measured plot size with compass and rope and then with GPS. If the closure error, which is a proxy of measurement precision with the compass and rope method, exceeded 5%, the area was remeasured. For GPS measurement, enumerators walked around the plot and wrote down plot size as estimated by GPS.

In total, 52 554 parcels were measured. Given the sheer scale of this survey, some errors in data reporting and entry are unavoidable. To avoid that these errors would bias the results, we discarded 1% of the observations with the greatest absolute errors. This implied that all observations with an absolute error greater than 0.70 were not included in the analysis. In appendix 3.A we show that our main results do not change when all the observations are included in the analysis.

Because compass and rope measurement is burdensome, enumerators could be tempted to only measure the plots with GPS and report a similar estimate for compass and rope measurement. If this has occurred, both measurements would be strongly correlated and we would wrongly conclude that GPS is as reliable as compass and rope measurement. To avoid this temptation, enumerators had to report all the intermediate measurements (i.e. distances between corners and angles between plot sides) required for compass and rope measurement. Moreover, only for 3% of the plots were both measurements exactly equal. This suggests that enumerators measured all plots with both methods

Finally, we checked whether the measurement with GPS and compass and rope follows Benford's law. This law states that in most data series small digits occur proportionally more frequently as leading digits (Miller, 2015). In other words, the digit '1' is expected to be the leading digit for 30% of the measurements, while the digit '9' is expected to be the leading digit for only 4.6% of the measurements. This law has been used to detect data fraud, because fraudulent numbers do not follow Benford's law (Rauch et al., 2011). The results, reported in appendix 3.B, are noteworthy. GPS as well as compass and rope measurement mirror the distribution predicted by Benfords law. Hence, we are fairly confident that all plots have been measured with the two different methods and have not been manipulated to facilitate field work.

3.4 Results

Average plot size was 434 m^2 and 441 m^2 measured with GPS and compass and rope, respectively, indicating that GPS slightly underestimated plot size. It turned out that GPS underestimated plot size on 61% of the plots. Mean and median relative error was -3.5% and -1.4%, respectively. Plot size was thus systematically underestimated with GPS by 1% to 4%. Figure 3.2 shows the correlation between plot size measured with GPS and compass and rope. This correlation was nearly perfect, which was confirmed by regressing plot size measured with GPS on plot size measured with compass and rope (table 3.1). When only including plots in the regressions smaller than 1000 m^2, 500 m^2 and 100 m^2, the model still explained 99%, 97% and 74% of the variance, respectively (results not shown). Hence, we did not find evidence that bias in GPS estimates was greater for small plots relative to large ones.

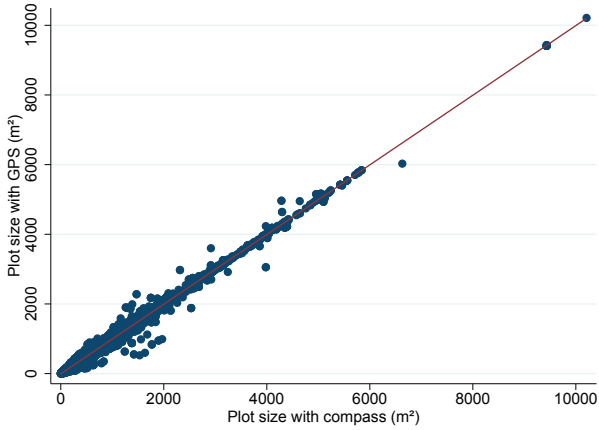
Table 3.1: Regression analysis

	OLS Dependent variable: plot size with GPS (m^2)
Plot size with compass and rope (m^2)[1]	0.99987
Constant[2]	-7.076
n	52 030
R^2	0.995

[1] Coefficient not statistically different from 1 ($p = 0.67$)
[2] Coefficient statistically significant from 0 ($p < 0.0001$)

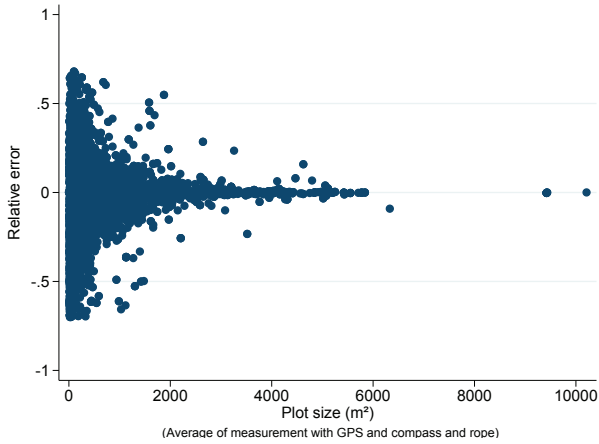
This was confirmed by plotting relative error as function of average plot size (figure 3.3). There is no indication that the GPS measurement is inaccurate since the relative errors are centered around zero. On average, GPS does not over or underestimate plot size. It is, however, also apparent from figure 3.3 that the precision of GPS measurement increases with plot size. For instance, an over or

Figure 3.2: Correlation between measurement of land area with GPS and compass and rope



underestimation of land area by 50% was not uncommon for plots smaller than 500 m^2, but occurred relatively rarely for larger plots. Because the precision of GPS measurement varies with plot size, it did not make sense to calculate 95% confidence intervals for the relative errors as is standard practice in the literature about the validation of new measurement instruments (Bland and Altman, 1995). As figure 3.3 shows, these confidence intervals would vary with plot size.

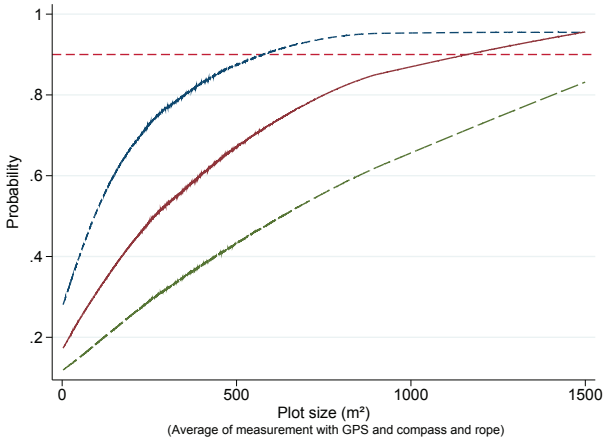
Figure 3.3: Relative error as function of plot size



Instead, we calculated the number of plots for which GPS measurement differed by less than 10%, 5% and 2.5% relative to compass and rope measurement as function of plot size (figure 3.4). This confirmed that precision increases sharply

with plot size. Ninety percent of area measurements (indicated with horizontal, dotted line) were within an absolute error of 10% and 5% when the plot is larger than 550 m^2 and 1100 m^2, respectively. Eight-six percent of the plots between 1500 m^2 and 10 000 m^2 have an absolute error smaller than 2.5%.

Figure 3.4: Share of observations (%) with an absolute error smaller than 10% (blue, dotted curve), 5% (red, solid curve) and 2.5% (green, dashed curve) relative to measurement with compass and rope as function of plot size



3.5 Discussion and conclusion

Whether GPS is an appropriate tool for area measurement in agricultural surveys in regions with many tiny plots of land depends on the purpose of the survey. Nevertheless, our study helps to define some rules of thumb.

First, there is some evidence that GPS slightly, but consistently, underestimates plot size relative to measurement with the compass and rope method. This corroborates previous findings (Keita and Carfagna, 2009; Schøning, 2005). However, this underestimation is so small (on average less than 3% in our sample) that it can be neglected for most practical purposes. Overall, area measurement with GPS is thus unbiased. This is good news if the objective of the survey is to measure total crop area at household or national level: aggregating area measurement of many plots will reduce measurement error. It is also promising for repeated measurement of plot size with GPS in the sense that averaging repeated measurements will reduce measurement error.

Second, measurement precision increases with plot size. This is again excellent news when estimating total crop area as it implies that larger fields, which contribute more to total crop area, are more precisely measured. Equally important

is the finding that the threshold below which compass and rope measurement is much more precise than GPS is well below 0.5 ha, which is often cited in the literature as the critical threshold. When plot size is greater than 1000 m^2 (0.1 ha), more than 90% (80%) of GPS estimates differed by less than 10% (5%) relative to compass and rope estimates. However, if the study requires precise measurement of crop area of very tiny plots, compass and rope methods may still be preferred over GPS. For instance, if one is interested in estimating production per hectare on very small plots, precise measurement of crop area is important. In our study only 60% of the plots between 100 m^2 and 200 m^2 were correctly estimated within a relative error of $\pm$ 10%.

Our findings confirm the findings of Carletto et al. (2014), who also compared GPS with compass and rope measurement. They concluded that GPS measurement is accurate, but that its precision increases with plot size and noted that relative errors of $\pm$ 10% are not uncommon on small plots.

Besides practical implications for area measurement in agricultural surveys, this study has one important implication for academic research using land area measured with GPS as independent variable in a statistical analysis. As we showed, relative measurement error in plot size is negatively correlated with plot size. This means that measurement error in plot size violates one of the conditions of 'classical' measurement error, which assumes random measurement error around the true, unobservable value of the independent variable. Non-classical measurement error biases the coefficient of the mismeasured variable and may even change the sign of the estimated coefficients (Bound and Krueger, 1989; Pischke, 2007). Measurement error in plot size should be considered non-classical if the plot is smaller than 1000 m^2. Consequently, only studies that deal with very small plots should assess the effect of non-classical measurement error on their findings (see chapter 7 for an application).

This study has several limitations, warranting further research. Researchers have argued that factors such as the shape and slope of the plot and the density of the vegetation affect the precision and accuracy of area measurement. Although we lacked the data to test this formally, we believe that these factors do not have a substantial impact on the accuracy and precision of area measurement. Irregularly shaped plots are easier to measure with GPS than with compass and rope because the latter method approximates the area by a polygon, which is less precise if the area is irregularly shaped. The slope of the plot affects both GPS and compass and rope measurement, but this effect is only important for large plots. Ideally, one should set up an experiment to test to which extent these factors influence area measurement and to assess under which conditions compass and rope measurement is more accurate and precise than GPS measurement.

The main limitation of the research is that we could not compare GPS measurement with self-reported are measures. Self-reported measures remain widespread

in household and agricultural surveys, because they are the least time-consuming since the do not require the enumerator to travel to each plot. Carletto et al. (2015a) have shown that self-reported measure are inaccurate: small plots tend to be overestimated, while larger plots are underestimated relative to GPS measurement.

Yet, this study demonstrates convincingly that, if self-reported measures are deemed too unreliable for the purpose of the survey, GPS measurement can replace compass and rope measurement under the condition that most plots are larger than 1000 m^2.

Appendix

3.A Land area measurement with GPS or compass and rope: full results

In the analyses reported in the main text, we discarded all observations with an absolute error greater than 0.7 (1% of the observations). As such, we avoided that errors due to data entry or misreporting biased our results. Some discrepancies between compass and rope measurement and GPS were so large that few would attribute them to imprecise measurement with GPS. For instance, one field was reported to measure 603.9 m^2 when measured with compass and rope and 61 542 m^2 when measured with GPS, causing an absolute error greater than 100, probably due to a data entry error. However, excluding observations with large absolute errors will improve the correlation between measurement with GPS and compass and rope and may thus lead to an overoptimistic assessment of the precision of GPS measurement.

To check the robustness of our results we reconducted the analyses only discarding observations with an absolute error greater than 10 (22 observations). Results of the base model (table 3.A.1) and the robust model (table 3.A.2) were then compared. Average absolute error increased from 7% (limited sample) to 10% (full sample). This increase was mainly caused by an increase of absolute error from 16% to 21% for fields between 0 and 100 m^2 and an increase from 10% to 13% for fields between 100 m^2 and 200 m^2. Consequently, most very large absolute errors occurred on very small plots. Additionally, the robustness check showed that the number of observations with an absolute error smaller than 10%, 5% and 2.50% did not change substantially compared to the base model.

Table 3.A.1: Measurement error as function of plot size (sample restricted to observations with absolute error smaller than 0.7)

Group (m^2)	n	Area measurement (m^2)		Difference	Absolute error	% of observations with absolute error smaller than:		
		Compass	GPS			10%	5%	2.50%
0 – 100	6759	64.9	58.9	−6.0	0.16	45%	28%	17%
100 – 200	10 869	149.7	142.2	−7.5	0.10	62%	38%	22%
200 – 300	9261	246.8	239.4	−7.4	0.07	74%	49%	28%
300 – 400	6240	347.3	340.2	−7.2	0.06	84%	59%	36%
400 – 500	4459	447.8	439.1	−8.7	0.05	91%	68%	42%
500 – 600	3151	545.6	538.8	−6.8	0.04	93%	73%	47%
600 – 700	2422	645.6	638.1	−7.4	0.04	94%	78%	52%
700 – 800	1757	747.6	742.0	−5.6	0.03	96%	81%	54%
800 – 900	1451	848.6	842.5	−6.2	0.03	95%	86%	61%
900 –1000	1058	948.6	945.3	−3.3	0.03	96%	89%	67%
1000 –1500	2690	1202.4	1194.8	−7.7	0.03	95%	89%	71%
>1500	1913	2287.1	2279.2	−7.9	0.02	97%	93%	86%
Average	52 030	442.2	434.1	−7.1	0.07	76%	56%	37%

Table 3.A.2: Measurement error as function of plot size (sample restricted to observations with absolute error smaller than 10)

Group (m^2)	n	Area measurement (m^2)		Difference	Absolute error	% of observations with absolute error smaller than:		
		Compass	GPS			10%	5%	2.50%
0 – 100	6943	64.3	59.8	−4.4	0.21	44%	27%	17%
100 – 200	10 968	149.6	145.1	−4.5	0.13	61%	37%	22%
200 – 300	9303	246.9	241.9	−4.9	0.08	74%	48%	28%
300 – 400	6270	347.3	343	−4.2	0.07	83%	58%	36%
400 – 500	4486	447.8	442.6	−5.2	0.06	90%	67%	41%
500 – 600	3160	545.6	538.4	−7.2	0.05	92%	73%	47%
600 – 700	2428	645.6	636.7	−8.9	0.04	93%	78%	52%
700 – 800	1774	747.6	738.6	−9	0.04	95%	80%	54%
800 – 900	1461	848.6	860.4	11.9	0.06	95%	86%	60%
900 –1000	1062	948.6	942.1	−6.5	0.03	96%	89%	67%
1000 –1500	2735	1201.9	1199.3	−2.6	0.06	93%	87%	70%
>1500	1942	2296.6	2252.1	−44.5	0.03	96%	92%	85%
Average	52 532	441.3	435.2	−6.08	0.1	75%	55%	36%

In sum, the choice of discarding all observations with an absolute error larger than 0.7 does not invalidate our conclusions. Measurement with GPS is unbiased and its precision increases rapidly with plot size.

3.B Benford's Law

Benford's law states that in most data series small digits occur proportionally more as the leading digit. More formally, a set of numbers satisfies Benford's law if the leading digit occurs according to the following probability distribution:

$$P(d) = log_{10}(1 + \frac{1}{d}) \tag{3.2}$$

With $d = (1, 2, 3, 4, 5, 6, 7, 8, 9)$. Hence, if a set of numbers follows Benford's law, the leading digit is '1' for 30.1% of the observations ($log_{10}(2)$), '2' for 17.6% of the observation ($log_{10}(1 + 1/2)$) and so forth. A chi-square test with 8 degrees of freedom can be used to test formally if the distribution follows Benford's law.

Table 3.B.1 shows the frequency distribution of the leading digits for measurement with GPS and compass and rope as well as the expected probabilities according to Benford's law. Both measurements tend to satisfy Benford's law since the difference between observed and expected probabilities is small. A chi-square test rejects that the measurements follow Benford's law. However, this test is not appropriate in this case because there are so many observations and a chi-square test is proportional to the number of observations. With more than 50 000 observations, a chi-square test only accepts the null hypothesis if the set of numbers satisfies Benford's law nearly perfectly (Rauch et al., 2011).

Table 3.B.1: Frequency distribution of leading digits of GPS and compass and rope measurement versus expected frequency according to Benford's law ($n = 52\,030$)

Leading Digit	Compass and rope (% of observations)	GPS (% of observations)	Benfords law (% of observations)
1	28.40	28.53	30.10
2	19.64	19.26	17.61
3	13.33	13.16	12.49
4	10.03	10.47	9.69
5	7.64	7.64	7.92
6	6.52	6.53	6.69
7	5.49	5.25	5.80
8	4.81	4.86	5.12
9	4.14	4.31	4.58

In sum, it is remarkable that both measurements tend to follow Benford's law. As a violation of Benford's law is often taken as suggestive evidence of data manipulation, we can be fairly confident that the plot measurements were not manipulated by the enumerators to facilitate field work.

CHAPTER 4

A validity assessment of the Progress out of Poverty Index (PPI)

Abstract: Development organisations need easy-to-use and quick-to-implement indicators to quantify poverty when requested to measure program impact. In this paper we assess the validity of the Progress out of Poverty Index (PPI)$^{\text{TM}}$, a country-specific indicator based on ten closed questions on directly observable household characteristics, by its compliance to the SMART criteria. Each response receives a pre-determined score, such that the sum of these scores can be converted into the likelihood the household is living below the poverty line. We focus on the PPI scorecard for Rwanda, which was validated using two national household surveys conducted in 2005/06 and 2010/11. The PPI is Specific, Measurable, Available cost effectively, and Timely available. Yet, its Relevance depends on the way it is used. Although it accurately distinguishes poor from non-poor households, making it a useful reporting tool, its limited sensitivity to changes in poverty status restricts its usefulness for evaluating the impact of development projects.

This chapter is published as:
Desiere, S., Vellema, W., D'Haese, M., 2015c. A validity assessment of the progress out of poverty index (PPI$^{\text{TM}}$). Evaluation and Program Planning 49(0), 10–18

4.1 Introduction

Development programs with the objective of poverty alleviation want to target the poorest households, but lack resources, time and expertise to develop their own detailed poverty measures or conduct full-scale household surveys. Consequently, such development programs rely on standardized indicators to measure poverty and evaluate the impact of their program. Ideally, such indicators are designed according to the SMART criteria [1]: *Specific, Measurable, Available cost-effectively, Relevant* and *Timely available* (European Evaluation Network for Rural Development, 2014; Poister, 2008). The Progress out of Poverty Index (PPI^{TM}), introduced by the Grameen Foundation, is promoted as a tool that can quantify the share of program participants living below the poverty line, assess the performance of the intervention among the poor and poorest, and track poverty levels over time[2]. By design, the PPI meets four of the five SMART criteria. It is *Specific, Measurable, Available cost-effectively, and Timely Available.* The *Relevance* criterion, however, requires validation. Assessing this validity is the objective of this paper.

4.1.1 Background

In order to be able to more accurately value the merits and shortcomings of the PPI, we provide a brief overview of alternative ways to measure poverty, paying particular attention to the extent to which these measures comply with the SMART criteria.

Consensus appears to have been reached on the vision that poverty is multidimensional. However, such consensus does not exist on the best way to measure poverty, evidenced by the large and growing number of poverty indicators. The most frequently used poverty indices are income- or expenditure-based (Ravallion et al., 1991).

Of these, perhaps the most well-known are the dollar-a-day extreme poverty line and the more generous two-dollar-a-day poverty line developed by Ravallion et al. (1991) for the 1990 World Development Report. Households are considered poor when their income or total expenditure falls below a certain threshold. A downside of income and expenditure-based poverty indices is that data collection is costly, extremely time-consuming and prone to measurement error (Beegle et al., 2012; Deaton, 1997). For example, the food expenditure part of the 2005/06 Rwanda Household Living Standard Survey counted 75 pages and required enumerators to visit each household 11 times (Schreiner, 2010). Hence, expenditure-based poverty

[1] Several definitions of SMART have been developed. We follow the definition as proposed by the European Evaluation Network for Rural Development of the European Commission (`http://enrd.ec.europa.eu/en`, accessed April 2014)

[2] www.progressoutofpoverty.org

indicators are neither *Available cost effectively* nor *Timely available*, which are key SMART principles for successful implementation by development programs.

Additional shortcomings of income and expenditure based poverty measures are the difficulty of accurately defining a line below which people are poor and above which they are not (i.e. quantifying the poverty lines) and their inherently static nature (Carter and Barrett, 2006). Most income-based measures are static in nature as they measure only if a household (or individual) fails to meet minimum income levels to cross a predefined poverty line. Yet, some households – the so-called transient poor – live from an income below the poverty line at a particular moment in time, but have sufficient productive assets to escape poverty, while structurally poor households lack the resources to move out of poverty over time (Barrett et al., 2006; Baulch and Hoddinott, 2000). Development programs with the aim of alleviating poverty in the long term should be especially concerned about these structurally poor households and thus require a poverty measure that allows their identification (Barrett, 2010). Hence, the *specificity* of expenditure-based indicators can be questioned.

Asset-based approaches to poverty measurement have been proposed to distinguish the structurally poor from the transient poor (Carter and Barrett, 2006). Besides being more cost-efficient and less demanding in terms of data requirements, this approach is robust to small fluctuations in poverty levels and, therefore, might be able to capture the structural component of poverty (Adato et al., 2006). Several asset indices are already extensively used. A distinction can be made between those that use a theoretical and axiomatic framework and those that are primarily data-driven.

Of the indicators using a theoretical and axiomatic framework the Human Development Index (HDI) and the Multidimensional Poverty Index (MPI), both developed by the UNDP, are among the most frequently encountered (UNDP, 2010). The HDI can only be used at macro-level. The MPI is developed for both macro and micro-level and comparable to other asset-based poverty indicators. The MPI is a weighted average of three pre-defined dimensions of poverty - education, health and standard of living - and can be decomposed into poverty headcount and intensity (Alkire and Foster, 2011). For each dimension several sub-indicators are included. A household is considered multidimensionally poor if it suffers deprivation in at least 33% of the weighted sub-indicators (Alkire and Foster, 2011). These indices were specifically developed to compare poverty between countries and over time and perform quite well in that context (Alkire and Santos, 2014). However, they might not be sufficiently sensitive to changes in poverty rates in specific local contexts, which casts doubt on their relevance for use in development programs. For such programs, the indicator developed by Zeller et al. (2006) might be more useful. This indicator, like the MPI and HDI, has a theoretical basis but was developed to assess to which extent a policy or program reaches the poorest. A downside of this indicator is its reliance on principal component

analysis (PCA), which makes it decidedly less easy-to-use and, thus, violates the *cost-effective Availability* criterion.

The second group of poverty indicators does not use theoretical justifications to select assets but selects assets solely on their statistical relationship with poverty. This type of indicators accepts that 'income poverty' is the gold standard of poverty, but is too costly and time-consuming to measure for impact assessment purposes for small-scale development programs. Hence, rather than measuring income, it uses assets which are easier to observe than income and which are nevertheless strongly correlated with income. This philosophy is very different from that of multidimensional poverty indicators such as the MPI which select assets to measure poverty to avoid reducing the complex concept of poverty to income poverty. Because of their data-driven nature, these indicators are country-specific. Two frequently used indicators in this group are the Poverty Assessment Tool (PAT) and the Progress out of Poverty Index (PPI). They do not exist for all countries yet, but are under continuous development. The PAT has been developed by USAID (2014) and its use is mandatory for many USAID funded projects, while the PPI grew out of a microfinance initiative in Bosnia-Herzegovina (Matul and Kline, 2003) and was further developed by Mark Schreiner of Microfinance Risk Management L.L.C and the Grameen foundation. At the time of writing, the PPI was available for 46 countries and used by 176 organisations, mostly in the sphere of microfinance, including Oiko Credit, a financial cooperative supporting microfinance initiatives worldwide, and Dia, an organisation supporting MFIs in India (Grameen Foundation, 2014b,a). Both tools are similar in many respects, although the PAT is slightly more accurate and the PPI more widely available (Schreiner, 2014). Neither of the indicators is frequently encountered in the academic literature, although the PPI has been used to assess program impact (Blauw and Franses, 2011; Larsen and Lilleor, 2013) and as a benchmark indicator (Dinh and Zeller, 2010).

As many organisations are interested in using PPI in program evaluation and impact, an independent evaluation of how SMART it is as an indicator for poverty, and the extent to which it can be used for targeting the poor and monitoring and evaluating development programs is highly relevant for policy makers and development professionals.

The PPI has been developed with the specific aim to measure poverty at household level in a particular country. Moreover, the tool has been designed to provide a cost effective and timely available proxy for poverty. Hence, the SMART principles *Specificity, Measurability, cost-effective Availability and Timely availability* are clearly met. The *Relevance* criterion, however, cannot be accurately assessed without validation. We distinguish two important, but different aspects of the *Relevance* criterion: the ability of the indicator to distinguish poor from non-poor households regardless of where they live and the sensitivity of the PPI to changes in poverty over time. The first point is important for targeting and reporting,

while the second point is crucial for the indicator to be useful for monitoring and evaluation.

In this paper we analyse the PPI based on data from Rwanda, which is particularly well-suited to assess the relevance of the PPI for two reasons. First, the PPI has been calibrated on a household survey conducted in 2005/2006 and a similar household survey has been conducted in 2010/2011. This ensures that we have perfectly comparable data. Second, in the 5-year interval between survey rounds the country has experienced considerable economic growth, creating changes in poverty rates (Ansoms and Rostagno, 2012). This is a necessary condition to assess the sensitivity of the PPI to changes in poverty rates over time.

In the next section we briefly explain how the PPI is constructed. Then, we outline the methodology and describe the data used to examine the validity of the PPI. In the results section we show that poverty estimates based on index scores corresponded well to official poverty rates and that the index was useful for reporting and targeting. Furthermore, we show that its sensitivity to changes in poverty over time depended crucially on a limited set of items; most items were stable, and did not change over time. In the final section the implications of these findings are discussed.

4.2 The progress out of poverty index

PPIs have already been developed for 46 developing countries. Their development is always based on detailed household-level data such as captured by the Living Standards Measurement Surveys of the World Bank or national household surveys and the methodology is standardized (Schreiner, 2010). First, out of the household-level variables in the survey, a pre-selection of 100 indicators in the area of family composition, education, housing, and durable goods is made. Out of these, ten are selected that have a high correlation with poverty measured by the uncertainty coefficient (Goodman and Kruskal, 1979), are inexpensive to collect, easy to answer quickly, simple to verify, and liable to change over time as poverty status changes (Schreiner, 2010). These ten items are given weights using logistic regression, such that final scores on the index range from 0 to 100. A scorecard is produced which allows users to calculate scores on the spot (figure 4.1). Using look-up tables, these scores can subsequently be converted into the likelihood that a household is below any one of a number of poverty lines (appendix 4.B.1). In general tables are provided for 50%, 100% and 150% of the national poverty line, the food poverty line and an international poverty line such as the $1.25 (per person/day) line. Finally, the goodness-of-fit is assessed with out-of-sample calibration and standard errors for the likelihood of living below the poverty line given a PPI-score are obtained with bootstrapping. Country-specific details are provided in documentation available at the website of the Grameen Foundation.

Figure 4.1: Poverty scorecard of Rwanda developed by the Grameen Foundation

Indicator	Value	Points	Score
1. How many household members are 17-years-old or less?	A. Five or more	0	
	B. Four	1	
	C. Three	7	
	D. Two	8	
	E. One	13	
	F. None	20	
2. Have all household members ages 7 to 17 been to school in the last 12 months?	A. No	0	
	B. Yes	2	
	C. No one in age range	3	
3. What is the highest grade that the female head/spouse has successfully completed?	A. Never attended school	0	
	B. Attended and completed none, one, or two years	2	
	C. Years 3 or 4 of primary	3	
	D. Years 5 or 6 of primary	5	
	E. There is no female head/spouse	5	
	F. Anything after 6 years of primary	9	
4. What is the status of the male head/spouse in his main occupation?	A. Agricultural wage worker, or does not work	0	
	B. There is no male head/spouse	3	
	C. Self-employed in agriculture, or unpaid worker (homemaker, apprentice, volunteer, etc.)	4	
	D. Non-agricultural wage worker	5	
	E. Self-employed in non-agriculture	8	
5. What is the main material of the floor?	A. Packed earth	0	
	B. Wood, cement, tiles, bricks, stone, or other	7	
6. How many rooms does the household occupy (do not count bathrooms, water closets, or kitchen)?	A. One	0	
	B. Two or three	5	
	C. Four	7	
	D. Five	9	
	E. Six or more	12	
7. What is the main source of lighting for the household?	A. Burning wood, or other	0	
	B. Home-made kerosene or fuel-oil lamp (*agatadowa*)	8	
	C. Candles, gas lamp, electrical grid, or generator	13	
8. What is the main fuel used for cooking?	A. Firewood, field waste, or other	0	
	B. Charcoal, LPG, electricity, or kerosene	16	
9. Does the household own a radio or radio-cassette player?	A. No	0	
	B. Yes	3	
10. How many ares of agricultural land does the household own or use?	A. 0 to 10	0	
	B. 11 to 35	1	
	C. 36 to 60	2	
	D. 61 to 100	4	
	E. 101 to 150	6	
	F. 151 or more	9	
Microfinance Risk Management, L.L.C., `http://www.microfinance.com`		**Total score:**	

The PPI scorecard for Rwanda was developed according to the standardized methodology described above (Schreiner, 2010). It was calibrated on the national household survey EICV 2 (Enquête Intégrale sur les Conditions de Vie) conducted by the government of Rwanda in 2005/2006, which interviewed 6900 households. This survey was developed to monitor living conditions and covered all provinces of Rwanda. Data are publicly available from the website of the National Bureau of Statistics of Rwanda (GoR, 2006). The final PPI index for Rwanda (figure 4.1) contained two questions on household composition (Q1, Q4), two on education (Q2, Q3), four on housing conditions (Q5-Q8) and two on ownership of durable goods (Q9, Q10). Questions with higher discriminatory power in distinguishing

poor from non-poor households were given a larger weight in the overall score. For instance, a maximum number of points (13) was attributed to households using candles, gas lamps, generators or the electrical grid as their main source of lighting (Q7). On the other hand, some responses only marginally increased the likelihood of living above the poverty line: households that sent all their children to school scored only two points more compared to households where at least one child had not gone to school in the last 12 months (Q2). By summing the points received for each question, a total score is obtained which can be converted into the likelihood the household is living below the poverty line using the provided look-up tables (appendix 4.B.1). For instance, a household with a score between 50 and 55 had a likelihood of 22% of living below the national poverty line.

4.3 Data and methods

In this paper we use data from the 2010/11 EICV 3 survey to validate the 2005/2006 PPI scorecard of Rwanda. The EICV 3 expenditure survey contains 14 308 observations and used similar methodology and questionnaire as the earlier survey round, the EICV 2 (GoR, 2012a), which was used by Schreiner (2010) to develop the PPI for Rwanda. Both survey rounds used the same methodology and assumptions to construct the national poverty line and the food poverty line. These poverty lines were used to construct a poverty variable that indicates whether or not a household lived below the poverty or food poverty line, which is available for each household in the dataset.

In Rwanda, as is common for developing countries, poverty lines were calculated following the basic-needs approach (Ravallion, 2012). This means that a basket of consumption goods corresponding with local dietary patterns that meets the minimum energy-intake requirements was selected and converted to its monetary value. Based on the EICV 2, the food poverty line was set at 45 000 Rwandan Franc (RwF) per adult equivalent in 2001 prices. The poverty line takes into account non-food expenditure on top of food expenditure and was set at 64 000 RwF per adult equivalent in 2001 prices. These national poverty lines were adjusted for household composition, inflation, and difference in price levels between the Rwandan provinces to make them comparable over time and between regions. The methodology is described in detail in GoR (2012a) and by McKay and Greenwell (2007).

Because of the similarities between EICV 2 and EICV 3, the methodology to validate the PPI by comparing both surveys is rather intuitive and straightforward. First, the PPI score was calculated for each household in the EICV 3 sample based on the scorecard calibrated on 2005/06 data (figure 4.1). This allows comparing the correlation between PPI scores and poverty rates and thus assess its *Relevance* as a proxy of poverty: a higher PPI score in 2010/11 needs to correspond with a

lower probability of living below the poverty line and thus with a higher income. This relationship should hold for both rural and urban households of which poverty rates seem to have changed differently. Coding was straightforward and only 13 out of 14 308 observations had to be dropped because of missing variables.

Second, the effectiveness of the PPI as a tool to report and target the poorest households included in a development program was assessed, i.e. assessing its *Relevance* in distinguishing poor from non-poor households. Given a cut-off value of the PPI for inclusion in a pro-poor development program, we determined the number of households that would correctly be identified as poor and the number of households that was non-poor but would nevertheless be included. The better the PPI is able to identify poor households and to exclude non-poor households, the higher its discriminatory power, and the smarter it is as poverty indicator.

Third, we checked the relevance of the PPI in tracking changes in poverty status over time. To this end, we compared the actual poverty rate in 2010/11 with the poverty rate predicted by the PPI converted into poverty rates with the tables calibrated and provided by Schreiner (appendix 4.B.1). From these results, we determined the questions that contributed most to the overall change in the PPI over time.

4.4 Results

4.4.1 *Relevance:* distinguishing poor from non-poor households

To verify whether the PPI scorecard was still able to accurately identify poverty five years after having been developed, PPI scores were compared to actual poverty rates in 2010/11. A PPI score was calculated for all households in the EICV 3 dataset based on the scorecard for Rwanda (figure 4.1). The distribution of the PPI score was nearly normal, but slightly skewed towards the right ($mean = 41.85$, $sd = 13.07$). Almost 80% of households had a PPI score between 25 and 55, while none of the households had a score below 5 or above 95 (table 4.1).

Table 4.1 confirmed the internal validity of the PPI: a higher PPI score reduced the probability of being poor. For instance, a household with a PPI between 35 and 40 had a probability of 50% (23%) to live below the national (food) poverty line, while these probabilities decreased to 38% (15%) for households with PPI scores between 40 and 45. Similarly, the number of households belonging to the highest income quintile increased with higher PPI scores, while the number belonging to the lowest quintile decreased with higher scores.

Table 4.1: Trends of PPI (calculated from 2010/11 data using the 2005/06 scorecard) by indicators of poverty in 2010/11

PPI	HH below poverty line (%)	HH below food poverty line (%)	HH in highest income quintile (%)	HH in lowest income quintile (%)	n
5 – 9	100	92.3	0	92.3	13
9 – 14	93.8	83.3	0	78.1	96
15 – 19	86.2	68	1.3	62.2	225
20 – 24	80.2	57.8	2.2	51.9	592
25 – 29	74.5	45.4	3.4	38.5	1219
30 – 34	61.3	33.3	4.5	28.2	2039
35 – 39	50.3	23.1	9	18.2	2556
40 – 44	37.8	15	13.5	10.9	2448
45 – 49	25	7.7	24.5	5.3	1795
50 – 54	14	3.7	36.4	2.8	1142
55 – 59	7.4	1.4	54.2	0.5	734
60 – 64	2.9	0.8	74.2	0.6	476
65 – 69	0.6	0	88.6	0	350
70 –100	0	0	94.7	0	610

Source: Authors' calculations from the 2010/11 EICV 3
The reported poverty line and food poverty line are the national poverty lines as calculated by the Government of Rwanda, set at 118 000 RwF and 83 000 RwF in 2011 prices, respectively.

4.4.2 *Relevance:* in both urban and rural areas

Ideally, urban and rural households with identical PPI scores would have the same probability of living below the poverty line. Table 4.2 shows the results for Rwanda. For both rural and urban households, poverty rates decrease with increasing PPI. Given their PPI score, poverty rates were lower for urban than for rural households (see last column, table 4.2). The largest difference (of 13 percentage points) occurred for households with scores between 40 and 45. Considering that unbiased poverty estimates require a representative sample of Rwandan households (Schreiner, 2010), the finding that the overestimation of poverty rates is urban regions remains below 15 percentage points is encouraging for development programs which target households that are not completely representative of the country.

The systematic, overestimation of urban poverty might be related to the question on landownership (Q10). More than half of urban households indicated to own less than 10 ares of land, which does not necessarily signal poverty in an urban region. By including land in the index, there is an inherent underestimation of the expenditure of urban household. Hence, the validity and comparability of the PPI measure suffers from a trade-off between a poverty indicator that is valid country-wide and an indicator that is more accurate, but region-specific. Development programs in urban regions could consider excluding question 10 of the PPI and add the average score of Rwandan households (2.64 in 2005/06 and 2.17 in 2010/11)

on this question to the total score of the other 9 questions. This approach would reduce the bias in the PPI for urban households, while the look-up tables provided by Schreiner could still be used. However, as the index for each country includes different questions, this approach cannot be generalized directly. Rather, country-specific research would be needed for all programs targeting specific sub-groups of the population.

Table 4.2: Likelihood of living below the poverty line given a PPI score (calculated from 2010/11 data using the 2005/06 scorecard) for rural and urban households

PPI	Rural		Urban		Percentage point difference between rural and urban HH
	HH below food poverty line (%)	n	HH below food poverty line (%)	n	
5 – 9	100	12	100	1	0.0
9 – 14	93.6	94	100	2	−6.4
15 – 19	85.8	212	92.3	13	−6.5
20 – 24	80.4	552	77.5	40	2.9
25 – 29	74.8	1121	70.4	98	4.4
30 – 34	61.3	1903	61	136	0.3
35 – 39	51	2370	41.9	186	9.0
40 – 44	38.9	2239	26.3	209	12.6
45 – 49	25.7	1626	17.8	169	8.0
50 – 54	14.7	965	10.2	177	4.5
55 – 59	8.8	532	3.5	202	5.4
60 – 64	3.9	231	2	245	1.9
65 – 69	0.8	122	0.4	228	0.4
70 –100	0	169	0	441	0.0

Source: Authors' calculations from EICV 3.

4.4.3 *Relevance:* reporting and targeting of poor households

The Grameen Foundation promotes the PPI as a reporting tool that allows quick identification of poor households included in a development program. In his paper describing the development of the scorecard for Rwanda, Schreiner (2010) promotes it as a targeting tool, useful for selecting households to include in the program. The PPI does not directly identify whether or not a household is below a specific poverty line but reports likelihoods. To use the tool for targeting, a cut-off value has to be selected. Households with a score below the cut-off value are included, while households with a higher score are excluded from the program. Poverty likelihoods of individual households and poverty rates of groups of households can be determined through look-up tables provided with the scorecard (appendix 4.B.1). These tables are constructed based on the same information as the scorecard itself, which in the case of Rwanda is the 2005/06 household survey round on which Schreiner (2010) calibrated the PPI scorecard.

For the sake of conciseness, we present results for a PPI cut-off value of 35; each

household with a PPI below 35 would be included in our hypothetical pro-poor program, households with a PPI of more than 35 would be excluded. The 2005/2006 data (appendix 4.B.2) predicts that 47% of households would be included, and the program would reach almost 70% of the poor Rwandese households (table 4.3). Suppose we start the program in 2011 and we use the same PPI cut-off value of 35 calculated using the same scorecard, but with household data collected in 2010/2011. As living conditions have improved, less households have a PPI score below 35 in 2010/11. The program now targets 29% of the Rwandese households, and includes 50% of the country's poor households. Hence while intending to reach 70% of the poor households (point A on figure 4.2), the program reaches only 50% of them (point B on figure 4.2). Figure 4.2 illustrates that the inclusiveness of poor households is missed by 20 percent points when 'old' scorecards are used on 'new' data. Hence the scorecard seems not to be specific enough to capture poor groups.

Table 4.3: Comparison between estimated and actual targeting effectiveness with a cut-off values of PPI of 35 for a development programmes that starts in 2011

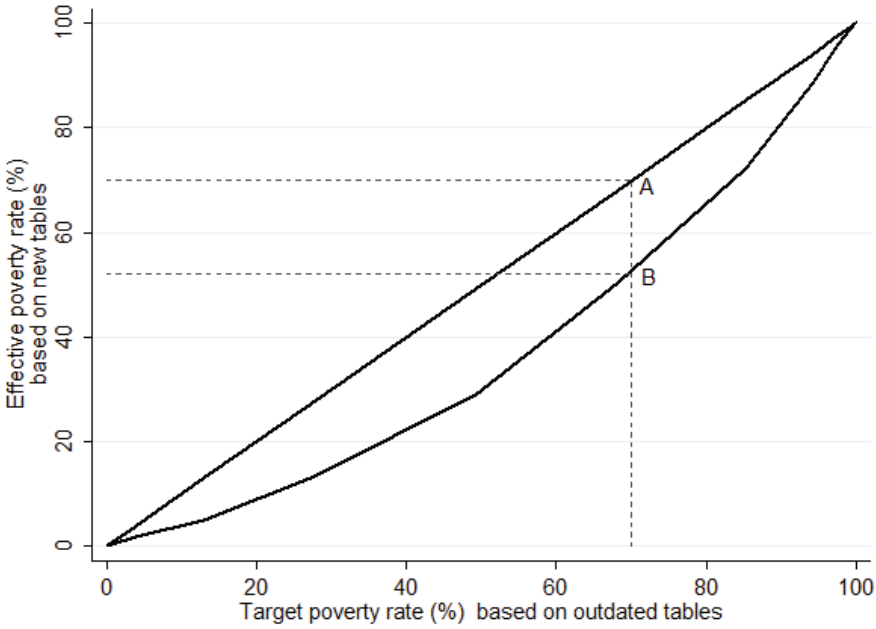
Cut-off value PPI of 35	% of HH targeted	% of poor targeted	% of targeted HH who are poor	Poor HH targeted versus non-poor HH targeted
Estimated targeting effectiveness (calibration with 2005/06 data)	47.4	68.1	77.4	3.4
Actual targeting effectiveness (calibration with 2010/11 data)	29.3	50.3	70.0	2.3

Source: Schreiner (2010, table 12, p75) and own calculations from the 2010/11 EICV 3

Moreover, the severity of mistargeting for different target poverty rates is illustrated in figure 4.2. The horizontal axis shows the proportion of poor households that the project initially aimed to include (which corresponds to different PPI cut-off values), while the vertical axis is the proportion of poor households that were actually included calculated on the 2010/11 data for each PPI cut-off value using the 'old' scorecard (see 4.C.1 for full results). Ideally, for each PPI cut-off value, the proportion of poor households targeted in our program would be equal using either 2005/06 or 2010/11 data (diagonal). Yet, proportions of poor households targeted for each PPI cut-off value calculated with 2010/11 data are substantially lower as shown by the curve below the diagonal which shows the relationship of the proportion of poor households targeted for each PPI cut-off value using 2005/06 data versus 2010/11 data. Projects starting in 2011 that based their cut-off values on estimations on 2005/06 data would consistently reach fewer poor households than intended (the curve never crosses the diagonal).

Targeting effectiveness seems to be highly sensitive to the initial choice of PPI cut-off values. This effect was especially severe for projects choosing relatively low cut-off values, as the effect was most pronounced for cut-off values in the range

Figure 4.2: The number of effectively included poor households (vertical axis) in a development programme is always lower than initially intended (horizontal axis)



Source: Authors' calculation from EICV3 (see 4.C.1 for full results,) and tables provided by Schreiner (2010) (see 4.B.2 for full table)

of 30 to 50. As such, development projects only using the PPI as a reporting device for poverty inclusion would consistently overestimate the number of poor households that were actually reached by their project.

Another relevant question for development aid donors of pro-poor programs is whether the number of households that are non-poor who are nevertheless included in the development program is sensitive to an inaccurate choice of the cut-off value. For instance, based on the 2005/06 table (appendix 4.B.2) a project with a cut-off value of 35 would have estimated that 77% of the targeted households were living below the poverty line. However, based on more recent data (table 4.3), only 70% of targeted households were living below the poverty line. On a positive note, this bias remained below 15 percentage points across all PPI cut-off values.

An alternative approach to illustrate the usefulness of the PPI as a indicator for both reporting and targeting are the so-called ROC curves. ROC curves show the sensitivity and specificity of an indicator. In appendix 4.A, we draw ROC curves

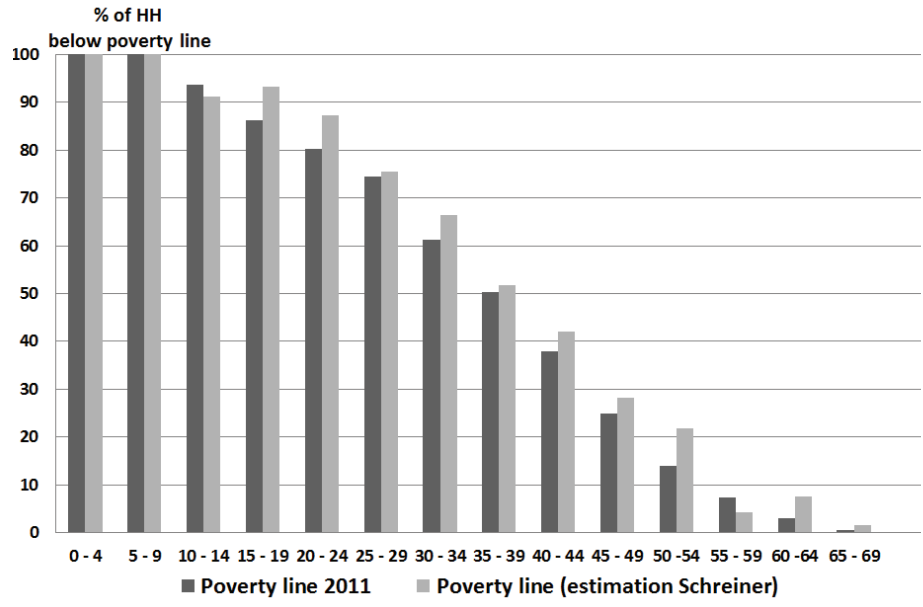
based on the 2005/06 and the 2010/11 data. This confirms that the PPI distinguishes poor from non-poor households in both periods (i.e. the indicator has a high specificity and sensitivity). It also confirms that developing programs implemented in 2010/11 but selecting a cut-off value for inclusion in the program based on the outdated 2005/06 data will systematically include less poor households than they intend to (sensitivity is lower than intended).

4.4.4 *Relevance:* monitoring program impact over time

To be a *Relevant* indicator for impact analysis, the PPI should capture changes in poverty rates over time. Schreiner (2010) states explicitly that some questions such as radio ownership were purposely included in the Rwanda PPI scorecard because they were believed to be sensitive to changes in poverty rates over time. Yet, this was not formally tested by Schreiner (2010).

We test sensitivity to changes in poverty status by comparing estimated poverty rates based on the tables provided by Schreiner (and calibrated on the 2005/06 data) with actual poverty rates observed in 2010/11 based on the national poverty line. Poverty rates estimated with the outdated tables consistently overestimated actual poverty rates in 2010/11 (figure 4.3). However, overestimation was limited and never exceeded 10%. A similar analysis for food poverty lines confirmed these results (results not shown but available upon request).

Figure 4.3: Comparison between actual number of HHs below poverty line in 2011 and the number estimated by Schreiner based on 2005 data



Moreover, the poverty rates at provincial level in 2010/11 were also estimated based on Schreiner's tables (appendix 4.B.2) and compared with the official poverty rates at provincial level (table 4.4). Besides a slight underestimation of poverty in the Southern province, the estimations based on the PPI were in line with government statistics. Hence, the PPI was sufficiently sensitive to monitor the decrease in poverty between 2005 and 2011. Therefore, at first glance it seems that the PPI achieves its objective of capturing structural improvements in poverty rates.

Table 4.4: Poverty rates in 2005/06 and 2010/11 estimated with PPI and compared with official government statistics

	2005/06		2010/11	
	PPI	Official poverty rate	PPI	Official poverty rate
Kigali City	22.0	20.8	16.1	16.8
Eastern Province	56.2	52.1	45.3	42.6
Northern Province	60.4	60.5	45.9	42.8
Southern Province	60.9	66.7	48.5	56.5
Western Province	60.9	60.4	47.1	48.4

Source: Own calculations and government report 'The evolution of poverty in Rwanda from 2000 to 2011: results from the household surveys (EICV)' (GoR, 2012b).

A closer look at the ten indicators that constitute the PPI revealed that the improvement in the overall PPI score from an average of 37.2 in 2005 to 41.9 in 2011 was mainly driven by two indicators, namely main lighting source (Q7) and radio ownership (Q9).Together, these indicators contributed to an increase in the overall PPI score of 3.8 points, which accounted for 80% of overall change. Radio ownership increased from 14% in 2005 to 86% in 2011, contributing 1.5 points to the overall increase. This sharp increase in radio ownership was confirmed by other sources (GoR, 2012). Note that it was also expected to happen by the developers of the scorecard. Similarly, the number of households achieving the maximum score of 13 on question 7, related to their main source of lighting, increased from 9% to 45%, which contributed 2.3 points to the overall change. This sharp increase was primarily explained by the fact that the highest attainable score for this item was attributed to households reporting in 2011 that battery-powered lanterns were their main source of lighting a response category which did not yet exist in 2005. As almost one third of the households reported using battery-powered lanterns, this effect is important. Consider the following: if households with access to battery-powered lanterns were classified in the lowest category, attributing a score of zero to this question, the PPI overestimated poverty by more than 10% for several ranges of PPI scores (results not shown). In this case, PPI would no longer be *Relevant* as an indicator to monitor changes in poverty rates over time.

Most of the other questions included in the PPI are rather insensitive to changes in poverty rates. Items such as household composition or arable land area per

household are unlikely to change quickly as poverty decreases, such that almost all differences in poverty rates have to be picked up by only few items such as radio ownership.

4.5 Discussion and conclusions

The objective of this paper was to assess the validity of the PPI, which was developed as an easy-to-use and quick-to-implement asset-based indicator, for use in development programs. Its validity was assessed using the SMART criteria as defined by the European Commission. By design, the PPI is certainly *Specific* as it serves a well-defined purpose, namely measuring poverty at household level in a specific region; *Measurable*, because the proxy quantifies the probability a is living household below the poverty line; *Available cost effectively*, because the indicator consists of only 10 easy questions; *Timely available*, because collecting the PPI can easily be administered at regular time intervals and the data can be processed quickly. Yet, concerns arise on its *Relevanc*e in distinguishing poor from non-poor household and in capturing changes of poverty over time. Its relevance was tested in this paper, using data from Rwanda.

The relationship with expenditure poverty was analysed using a combination of the household survey round used to develop the PPI, the EICV 2, and its most recent version, the EICV 3. Overall, poverty estimates based on index scores were very close to officially reported poverty rates. The strength of this relationship was consistent between urban and rural areas, showing the robustness of the indicator to distinct living conditions within the country. The PPI was also correctly distinguishing poor from non-poor households, making it a useful targeting tool for development organisations.

Whether the PPI is accurate enough to use for reporting the number of poor households that were included in development programs depends crucially on which poverty thresholds and conversion tables were used. Our results indicate that a project starting in 2011 that would use the original conversion tables, which were developed five years earlier, might reach substantially fewer poor households than it had intended. This effect might be especially severe in the context of a country experiencing stark economic growth and substantial poverty reduction over the interval between scorecard development and field application, such as Rwanda (Ansoms and Rostagno, 2012). Consequently, reports of development projects might systematically overestimate the number of poor households participating in the program.

Whether reaching fewer poor household than initially intended or targeting more non-poor households poses a problem for a development program depends on the specific aim and context of the program. Moreover, it is never advisable to use

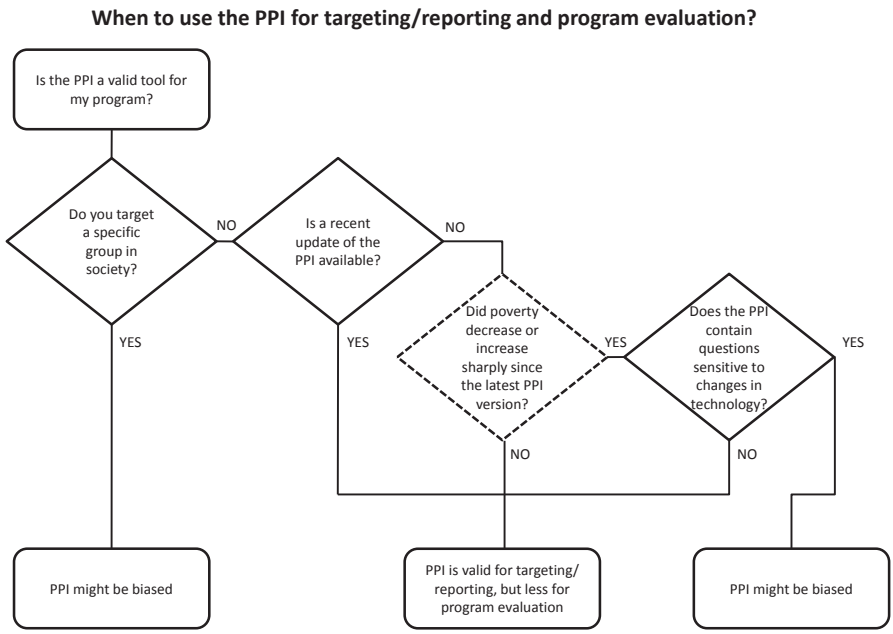
a single index to decide on inclusion or exclusion of a household in a program (Skoufias et al., 2001; Van de Walle, 1998) and it should be emphasized that the advocates of PPI do not advise to do so. Instead, they promote PPI as a reporting device. Although using PPI merely for reporting and not targeting would not result in mistargeting, our results show that such reports would consistently overestimate the number of poor households reached by the program. Hence, the indicator fails on a number of accounts in terms of relevance in measuring poverty.

One of the reasons the PPI was developed was to evaluate project impact. Therefore, care was taken to include items in the scorecard that were likely to change with improved income over time. Although poverty rates were found to be consistently overestimated for almost all ranges of the index, the degree of overestimation did not exceed 10%, indicating the index indeed appears to be reasonably sensitive. However, this sensitivity depended crucially on a limited set of items, with two items being responsible for over 80% of the observed variation and most items showing no significant change over time at all. This drawback is recognized by the Grameen foundation because poverty scorecards are updated as soon as new country-wide expenditure surveys become available. However, in a region with a sharp decrease in poverty rates over a short time-span such as Rwanda, an update of the PPI every 5 years might be insufficient to guarantee effective targeting of the poorest households and sensitivity to poverty status changes to evaluate the impact of development projects. Moreover, as soon as a PPI is updated, the baseline study will probably be outdated because this study will not necessarily include all the items of the new PPI. Consequently, it would no longer be possible or, at least, questionable to measure program impact over time by combining the old and new indicator. Although perhaps the general trend in poverty dynamics could still be assessed through comparing the poverty rates in the baseline year based on the old PPI and estimating poverty in subsequent years based on the most recent PPI, it would be questionable to use this combination to attribute program impact. An additional disadvantage of the need to update the PPI regularly are the costs involved in designing and disseminating the new PPI.

Another concern is that the indicators might be biased upwards by construction. Households that escape poverty might indeed buy a radio or increase the number of rooms in their dwelling, but households confronted with a negative shock which pushes them back into poverty are less likely to sell their radio or decrease the number of rooms in their house. Although this seems intuitive, without panel data with sufficient variation this assertion could not be tested directly. Furthermore, given land tenure systems prevalent in developing countries like Rwanda, it is unlikely that land will be sold. Given that the indicator was developed to be sensitive to changes in poverty, its sensitivity to negative shocks is an important consideration. Such sensitivity to both up- and downward movements in poverty is also crucial to make it a valuable indicator to study poverty dynamics and poverty traps (Carter and Barrett, 2006).

In conclusion, for such a relatively simple and easy-to-use indicator, the PPI does a remarkable job in estimating poverty levels and can be considered a SMART indicator, with some reservation regarding the *Relevance* component within the SMART framework. Figure 4.4 summarizes the external conditions affecting the *Relevance* of the PPI. The PPI is always a *Relevant* tool for reporting and targeting if a recent update is available. However, its accuracy might be compromised when the scorecard is several years old, an effect which is likely to be more pronounced for countries with sharply declining poverty rates. Although questions sensitive to technology change were included in the PPI to improve its sensitivity to changes in poverty rates over times, these questions also reduces its accuracy in targeting/reporting. Moreover, whatever the external conditions, its sensitivity to changes in poverty is rather limited and warrants further study. Hence, some hesitation is required before using the index as a tool to evaluate the impact of development projects.

Figure 4.4: External conditions affecting the *Relevance* of the PPI



Appendix

4.A ROC curves

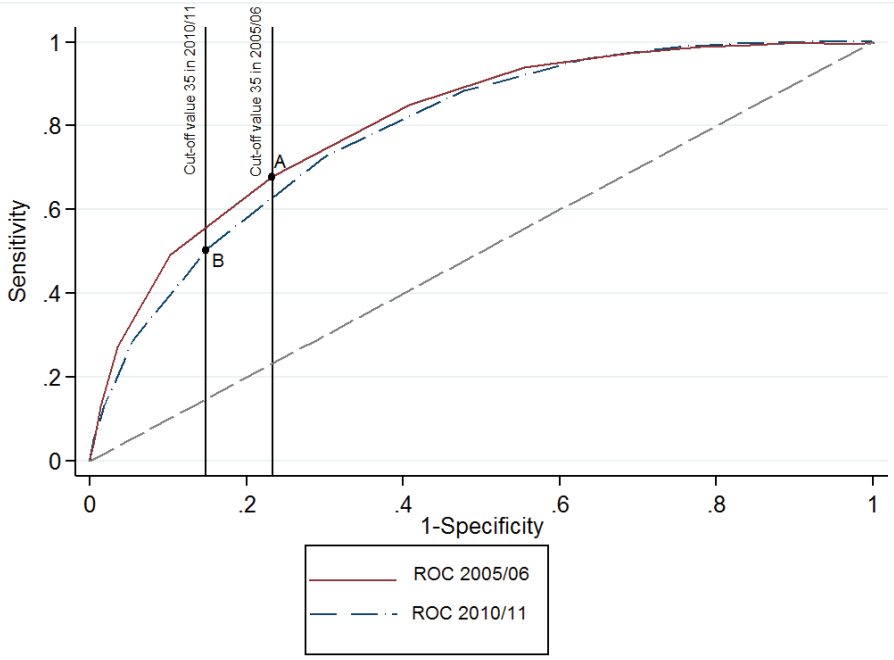
It the main text we showed that the PPI accurately distinguishes poor from non-poor households in 2010/11, although the indicator is calibrated on data from 2005/06 (table 4.1). We then showed that development programs implemented in 2010/11 would consistently reach fewer poor households than they intend to because they select a cut-off value for inclusion in the project based upon outdated data (table 4.3 and figure 4.2).

An alternative approach to illustrate both findings is the Receiver Operator Characteristic (ROC) curve. This curve shows the sensitivity and specificity of the indicator when the cut-off value of the PPI for inclusion in the program varies. The sensitivity, or true positive rate, of a poverty indicator is defined as the proportion of households that are correctly included in the program because they are poor. A good indicator has a high sensitivity since a development program wants to minimize the number of poor households that are wrongly excluded from the program. The specificity, or true negative rate, of a poverty indicator is defined as the proportion of non-poor households that are correctly excluded from the program. A high specificity is required because a development program does not want to target non-poor households. In sum, an indicator with a high sensitivity and specificity is useful for both reporting and targeting purposes. There is always a trade-off between sensitivity and specificity. If a program selects a high cut-off value for inclusion, most households will be included in the program. In this case, the sensitivity is high, but the specificity low. Whether a high sensitivity or specificity is required depends on the objectives of the program. For instance, if the cost of wrongly excluding a household from the program is high, the program will select a low cut-off value which has a high sensitivity, but a low specificity.

Figure 4.A.1 shows the ROC curves for the PPI in 2005/06 and 2010/11. In both periods, the PPI is a reliable indicator to distinguish poor from non-poor households. The PPI is slightly less reliable in distinguishing poor from non-poor households in 2010/11, as the ROC curve for 2010/11 is always below the ROC curve for 2005/06. This was expected since the PPI was calibrated on 2005/06 data and confirms the findings reported in the main text (table 4.1).

A development program set up in 2010/11 uses the information from 2005/06 to select a cut-off value for inclusion in the program. In other words, only the ROC curve for 2005/06 (red, solid curve in figure 4.A.1) is observed in 2010/11, while the correct 2010/11 ROC curve (blue, dotted line in figure 4.A.1) is unobservable. As a consequence, the development program does not observe the specificity and sensitivity of the PPI in 2010/11. Consider, for instance, a development program that selects a cut-off score of 35. Based on the 2005/06 data, the sensitivity and

Figure 4.A.1: ROC curves for the PPI in 2005/06 and 2010/11



specificity of this indicator is 68% and 77%, respectively. In 2010/11, however, the sensitivity and specificity of the indicator with a cut-off value of 35 has changed for two reasons. First, the ROC curve has slightly shifted downwards from 2005/06 to 2010/11, implying that the sensitivity of the indicator has decreased for any given level of specificity. Second, a cut-off value of 35 does not correspond to the same level of specificity in 2010/11 as in 2005/06. This is illustrated in figure 4.A.1 by the two vertical curves, which show the specificity for a cut-off value of 35 in the two periods. This curve has shifted towards the left over time. For both reasons, the sensitivity of the indicator has decreased from 68% in 2005/06 (point A) to 50% in 2010/11 (point B). Note that this trend was also reported in the main text (table 4.3 and figure 4.2). At the same time, the specificity of the indicator has increased from 77% in 2005/06 to 85% in 2010/11. In sum, a development program implemented in 2010/11 excludes more poor households in its program than it intends to (lower sensitivity). As a positive side effect, it also includes less non-poor households in the program (higher specificity).

4.B Original tables provided by Schreiner (2010)

Table 4.B.1: Estimation of likelihood of living in poverty given a PPI score (calibrated on EICV 2, provided by Schreiner (2010))

PPI	HH below poverty line (%)	HH below food poverty line (%)
0 – 9	100	100
9 – 14	91.20	85.40
15 – 19	93.20	81.30
20 – 24	87.30	65.50
25 – 29	75.50	52.40
30 – 34	66.50	42.70
35 – 39	51.80	27.10
40 – 44	42.10	15.30
45 – 49	28.10	5.70
50 – 54	21.80	12.00
55 – 59	4.20	1.50
60 –100	≤ 7.50	≤ 4.30

Source: Schreiner (2010, table 4, p66 and table 4; p77)

Table 4.B.2: Effectiveness of PPI in targeting the poorest household for different cut-off values of PPI in Schreiner (2010)

Cut-off value	% of HH targeted	% of poor who are targeted	% targeted who are poor	Poor targeted versus non-poor targeted
10	0.4	0.7	100	Only poor targeted
15	2.3	3.9	91.9	11.3
20	7.6	13	92	11.6
25	16.3	27.3	90.1	9.1
30	31.3	49.3	84.9	5.6
35	47.4	68.1	77.4	3.4
40	64.7	85.3	71	2.4
45	76.2	94.1	66.5	2
≥ 50	≥ 84.3	≥ 97.7	≤ 62.4	1.7

Source: Schreiner (2010, table 12, p75)

4.C Targeting efficiency: full table

Table 4.C.1: Effectiveness of PPI in targeting the poorest household for different cut-off values of PPI calibrated on 2010/11 data

Cut-off value PPI	% of HH targeted	% of poor who are targeted	% of targeted HH who are poor	Poor HH targeted per non-poor HH targeted
10	0.1	0.2	100	only poor HH targeted
15	0.8	1.8	94.5	17.2
20	2.3	5.1	88.9	8
25	6.5	13.3	83.4	5
30	15	28.9	78.3	3.6
35	29.3	50.3	70	2.3
40	47.1	72.4	62.6	1.7
45	64.3	88.4	56	1.3
50	76.8	96	50.9	1
55	84.8	98.8	47.4	0.9
60	90	99.7	45.1	0.8
≥ 65	≥ 93.3	100	≤ 43.6	≤ 0.8

Source: Authors' calculations from EICV 3, but methodology based on Schreiner (2010, table 12, p75)

CHAPTER 5

Assessing the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale (HFIAS)

Abstract: This study evaluates the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale (HFIAS) for rural households in Burundi. A panel of 314 households was interviewed in 2007 and again in 2012 to collect detailed agricultural production data and to assess households' food security status using the HFIAS. Tobit models showed that the HFIAS is significantly correlated with objective measures of food security such as annual food production, livestock keeping and coffee production in the two periods. This confirms that the HFIAS is cross-sectionally valid. However, while total food production decreased by more than 25% in calorific terms between 2007 and 2012, households reported an improvement in their perceived food security over the same period. This finding may be partly explained through response shifts, in which households assess their own food security status in comparison to that of their peers. This evidence suggests that HFIAS may not be inter-temporally valid and should not be used as a single indicator to study temporal trends in food security.

This chapter is published as:
Desiere, S., D'Haese, M., Niragira, S., 2015a. Assessing the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale (HFIAS) in Burundi. Public Health Nutrition pp. 1–11

5.1 Introduction

Measuring food security is challenging but important, as hundreds of millions of people around the world still lack access to sufficient food (Barrett, 2010). The World Food Summit of 1996 defined food security as 'a situation that exists when all people, at all times, have physical, social and economic access to sufficient, safe and nutritious food that meets their dietary needs and food preferences for an active and healthy life' (WHO, 2014). Defining and measuring indicators for these concepts is difficult, since food security is multidimensional and has both objective and subjective dimensions. Measuring food security has been the object of ongoing debate in both the academic world and within the development arena.

Objective indicators of food security at household level include the well-known calorie deprivation index, monetary indicators that can be used to construct a food poverty line, anthropometric measures and, more recently, dietary diversity scores that consider both micro- and macro-nutrients (Carletto et al., 2013a; Headey and Ecker, 2012, 2013; Jones et al., 2013; Maxwell et al., 2013; Webb et al., 2006; Cafiero et al., 2014). Supporters of subjective approaches argue that objective indicators do not take into account important, intangible, aspects of food insecurity such as constant worries about the possibility of food deprivation or limited dietary variations. Several indicators have been developed that include these aspects (Marques et al., 2014; Pérez-Escamilla, 2012). The FAO, for instance, has recently launched the 'Voice of the Hungry Project' (Cafiero et al., 2014; Ballard et al., 2013), which developed an experience- based food security indicator called the Food Insecurity Experience Scale (FIES). Since 2014, this indicator is included in the Gallup World Poll, which yearly conducts nationally representative surveys in more than 150 countries[1]. This indicator will be used to monitor food security over time and compare the food security situation between countries. A predecessor of the FIES is the Household Food Insecurity Access Scale (HFIAS). This nine-item scale captures people's perceptions about food insecurity using a range of indicators such as anxiety about food supply, limited dietary variety and quality and insufficient food availability (Coates et al., 2007). Such an index is easy to use and can be implemented at low cost, thus making it ideally suited for governments or NGOs to use to monitor and evaluate program impacts[2].

However, the simplicity of the HFIAS raises questions about its reliability. One can ask whether, in a cross-sectional survey, it is able to effectively discriminate between food secure and food insecure households. Previous studies showed that the HFIAS is correlated with objective food intake-based measures of food security and is thus cross-sectionally valid (Becquey et al., 2010; Deitchler et al., 2010;

[1] http://www.fao.org/economic/ess/ess-fs/voices/en/

[2] The distinction between objective and subjective indicators is more pragmatic than semantic. One may argue that several questions of HFIAS (e.g. Q6-Q9) are answered objectively and do not probe into perceptions.

Headey and Ecker, 2012; Knueppel et al., 2010; Regassa and Stoecker, 2012; Sahyoun et al., 2014; Toledo Vianna et al., 2012). In a longitudinal survey that follows the same households over time, the main concern is the inter-temporal validity of the index. This means that households experiencing an objective decline in food access over time should report feeling more food insecure than before. To date, few studies have investigated the inter-temporal validity of subjective food security indicators. A notable exception is a study in urban Burkina Faso which concluded that the HFIAS is able to capture the impact of high food prices on households' food security (Martin-Prevel et al., 2012). However, a similar study in urban Ethiopia showed that female volunteer AIDS caregivers reported feeling more food secure during three subsequent survey rounds in 2008, despite higher food prices and the loss of food aid (Maes et al., 2009). This latter finding calls into question the inter-temporal validity of the HFIAS.

This study contributes to this literature by evaluating cross-sectional and inter-temporal validity of the HFIAS over a time span of five years for a representative sample of rural farmers in the north of Burundi.

5.2 Methods

5.2.1 Sampling and study design

Household surveys were conducted in Ngozi, a rural province in the north of Burundi, from mid-June until the end of July in 2007 and then in the same period in 2012. The surveys were carried out by an experienced team from the University of Burundi in collaboration with researchers from the Universities of Antwerp and Ghent (Belgium). Four of the ten enumerators and the team leader participated in both survey rounds. The interview period coincides with the dry season, when agricultural production is low (15% of the annual total) (Cochet, 1998; USAID, 2009; République du Burundi, 2013b). There were also practical reasons for choosing this period: most villages in this region are only accessible during the dry season, and farmers have lighter workloads in this period, allowing them time to spend on interviews. The questionnaire which was drafted in French, was administered by a trained interviewer in approximately one hour in the local language, Kirundi. The enumerators were bilingual and a test-phase sought to ensure that all enumerators translated and interpreted the questions similarly.

Ngozi is administratively divided into nine 'communes', which are further divided into villages, known as 'collines' (République du Burundi, 2006). Within each of the nine administrative units, the surveys randomly selected ten villages, and four households from within each village, to participate in the study. Hence, a total of 360 households were interviewed.

In 2012, 340 out of the 360 households that had participated in the first round in 2007 were re-interviewed. However, 26 observations had to be disregarded due to missing variables or large outliers. Thus, the final dataset contains 314 valid observations.

Participants were informed about the study and provided their verbal consent prior to being interviewed. No sensitive personal data were sought. Because of the approach used and the questionnaire content, no formal ethical approval was sought prior to performing this study.

The Household Food Insecurity Access Scale

The HFIAS was developed by a team of researchers at Tufts University as part of the Food and Nutrition Technical Assistance (FANTA) project funded by USAID (Coates et al., 2007; Deitchler et al., 2010). The method assumes that food insecurity causes predictable reactions that are the same across countries and can be captured and quantified through a survey. Based on the eighteen questions of the US Household Food Security Survey Module (HFSSM), but adapted to the specific context of developing countries, the scale contains nine questions (overview in table 5.2). Together these questions cover a broad spectrum of experiences related to food security. The first asks about anxiety over food availability, the next three are related to food quality and the last five to the quantity of food intake. Each time a question elicits a "yes" response, it is followed by a frequency-of-occurrence question with three options: "rarely", "sometimes", "often". Responses of no to the initial question are coded as zero, whereas the answers "rarely", "sometimes" and "often" are coded as 1, 2 and 3, respectively. Subsequently, the scores on the nine questions are summed to calculate the index. This results in a continuous food insecurity indicator that ranges from 0 (food secure) to 27 (severely food insecure) (Coates et al., 2007).

Food production and consumption

The survey collected data on the total annual harvest of the main crops in Burundi (bananas, beans, cassava, coffee, maize, peanuts, peas, potatoes, rice, sorghum, sweet potatoes, soy, and taro[3]) based on a one-year recall by the household head. The twelve selected crops account for more than 90% of the energy intake of a household in Rwanda, a neighboring country with a similar dietary pattern (GoR, 2010).

Total yearly food production was used as a proxy for food consumption and two different indicators for food consumption were constructed[4]. The first indicator

[3]Bananas, sweet potatoes and beans are the main staple crops and accounted for respectively 46%, 21% and 11% of annual food production in 2007 and 26%, 25% and 24% respectively in 2012.

[4]We also expressed the monetary value of total aggregated agricultural production, based on self-reported prices, both including and excluding bananas. The correlation between aggregate

aggregated total annual production in terms of its energy content. The second indicator also expressed the total annual harvest in these terms, but excluded bananas. There were two reasons for the construction of this second indicator. Firstly, banana is a semi-cash crop that is both consumed in the household and sold on the market as the main ingredient for beer. Therefore, an increase in banana production does not necessarily directly entail an improvement in food security of the household because the additional revenues might be used to cover expenses not related to food consumption. Furthermore, there was a large drop in banana crop production between the survey years due to a bacterial disease. For both reasons, it was necessary to investigate whether bananas had a different impact on food security compared to other crops. Finally, both proxies for food consumption were expressed per capita and per day to make them more tangible.

Statistical analyses

To test the cross-sectional and inter-temporal validity of the HFIAS, we estimated the correlation between the HFIAS and household and farm characteristics that are expected to contribute to food security. The models assumed that a household, i, rationally evaluates its own food insecurity status based on its underlying household-specific characteristics in each period, t. However, not all household characteristics are directly observable and this requires making a distinction between observable household characteristics (X_i) such as food consumption and unobservable household characteristics (u_i), such as household-specific strategies to cope with stress in times of food shortages. Hence, the following model was estimated:

$$HFIAS_{it} = \alpha + \beta X_{it} + u_i + \gamma_t + \epsilon_{it} \tag{5.1}$$

Observable household characteristics are food consumption, coffee production, livestock ownership, off-farm work and household size. The production of banana bunches is included in a second set of analyses.

To test the cross-sectional validity of the HFIAS, we estimated equation 5.1 without taking into account the longitudinal nature of the data. Hence, equation 5.1 was estimated separately for both the 2007 and the 2012 samples excluding the year-fixed effects, γ_t, and household fixed effects, u_i, as independent variables. This has the advantage that we do not assume the same correlation between household characteristics and food insecurity in both periods, but the drawback is that we cannot control for unobservable household characteristics. Equation 5.1 was estimated with a Tobit model, which yields unbiased estimates even when the dependent variable is truncated in nature, which is the case for the HFIAS (Hsiao, 2003). Moreover, error terms were clustered at village level to avoid bias due to unobservable village characteristics[5].

production in monetary terms and that in terms of energetic value was higher than 75% in both periods.

[5]The errors were assumed to be clustered as follows: $\epsilon_{it} = v_g + \mu_{it}$, with v_g the error component specific to village g, and μ_{it} a normally and independently distributed error term.

To test the inter-temporal validity of the HFIAS, the longitudinal nature of our data was exploited. Equation 5.1 was estimated with a random-effect Tobit model, which allowed us to control for unobservable household characteristics and to take into account the truncated nature of the data (Hsiao, 2003). Inter-temporal validity was accepted if the year fixed effect γ_t was not significantly different from zero, because this condition is sufficient to ensure that all the variation over time of the HFIAS is explained through observable and unobservable household characteristics. Equation 5.1 was also estimated with a difference-in-difference approach. This means that the change in the HFIAS between 2012 and 2007 is regressed on changes in household characteristics[6]. This has the advantage that the estimations will not be biased by factors which might influence food security (such as education, soil quality or the household's assets), which did not change between 2007 and 2012.

Several sensitivity analyses were performed[7]. First, we checked for the possibility of enumerators interpreting the HFIAS questions differently (despite training prior to the survey), with some possibly consistently over or underestimating households' food security status. To control for this enumerator-specific effects were included in the models. The models were also re-estimated for a sub-sample restricted to the enumerators that participated in both rounds and for a sub-sample restricted to households that were interviewed by the same enumerator in both rounds. All analyses were performed with STATA 11.0 SE.

5.3 Results

5.3.1 Descriptive statistics.

Household size (5.8 on average, $p = 0.88$), farm size (around 1 ha, $p = 0.49$) and the number of households keeping livestock (around 20%, p=0.18) hardly changed between 2007 and 2012 (table 5.1). The proportion of households with at least one member engaged in off-farm activities decreased significantly, from 38% to 18% ($p < 0.01$). The proportion of households growing coffee also decreased somewhat, from 63% to 55% ($p < 0.01$). Households that did cultivate coffee harvested 441 kg in 2007 and 279 kg in 2012 on average. These figures may appear to indicate a sharp decline in coffee production, but coffee production in Burundi has a biannual harvest cycle in which an excellent harvest in one year is followed by a bad harvest in the next year (International Coffee Organization, 2013). Hence, it was no surprise that production in 2012 (a bad year) was lower than in 2007 (a good year).

[6]More formally, the following equation was estimated: $HFIAS_{i2012} - HFIAS_{i2007} = \gamma + \alpha(X_{i2012} - X_{i2007}) + \epsilon_i$ Inter-temporal validity is rejected if γ is significantly different from zero.

[7]For conciseness, we do not report on all the sensitivity analyses. All the models were also re-estimated with Generalized Estimating Equations (GEE) and count models.

However, there was marked decrease in total aggregated food production. Average production per day and capita equaled 2424 kCal in 2007, but decreased by 30% to 1762 kCal in 2012 ($p < 0.01$). This decrease was mainly driven by an even sharper decrease in banana production, which fell from 139 bunches per household in 2007 to only 50 bunches per household in 2012 ($p < 0.01$). This drop had a large effect on the total aggregated production as bananas are one of the main components of the Burundian diet (République du Burundi, 2013b), have a high energetic value and were cultivated by more than 95% of households in the sample. When we excluded bananas from total aggregate production figures, overall mean production did not change significantly between 2007 and 2012. The significant decrease in banana production was caused by the disease Xanthomonas Wilt, which has infected many banana trees in the region and is threatening the livelihoods of many households in eastern and central Africa. Agricultural research has not yet found an effective prevention or treatment of the disease (Tripathi et al., 2009).

Table 5.1: Sample descriptive statistics of small-scale farmers in Ngozi, Burundi, by round of data collection

	2007	2012
Household characteristics		
Household size	5.76	5.74
Age of household head	41***	45***
Farm size (ha)	0.84	0.89
Cattle ownership (% households)	19	24
Working off-farm (% households)	38***	18***
Farm characteristics		
Food production (kCal/d)	2430***	1770***
Food production excluding banana (kCal/d)	1250	1300
Coffee production (% households)	63**	55**
Coffee production (kg)	441***	258***
Banana production (% households)	95**	98**
Banana production (bunches/year)	139***	50***

Values are means or % of households; $n = 314$. Symbols indicate significant differences between rounds: *** ≤ 0.01, ** ≤ 0.05, *≤ 0.10

P-values were obtained with t-test and chi-square tests for means and percentages respectively

Banana production is expressed in harvested bunches: estimated average weight 15 kg/bunch

Despite these downward changes, the responses to all nine questions of the HFIAS suggested an improvement in the food security situation between 2007 and 2012 (table 5.2). For instance, in 2007, 80% of the households claimed to have eaten a smaller meal than they needed at least once in the previous two weeks, compared to 70% in 2012. When the frequency of occurrence questions were taken into account (table 5.2, columns 3 and 4), the HFIAS decreased significantly from a mean score of 13.9 in 2007 to 10.8 in 2012 ($p < 0.01$). Thus, households reported

feeling more food secure in 2012 than in 2007, despite the decrease in food, banana and coffee production.

The internal consistency of the responses to the questions was assessed using Cronbach's alpha. All questions related positively, and Cronbach's alpha was 0.93 in 2007 and 0.95 in 2012. Principal component analysis revealed two components, food quality and food intake (results available upon request). The first component was positively associated with the first five questions, while the second component loaded positively on the last three questions. These two components accounted for 76% and 86% of total variance in 2007 and 2012, respectively, indicating that the HFIAS is internally valid.

Table 5.2: Responses to the nine question of the HFIAS, by round of data collection

	Affirmative responses[1]		Mean score	
	2007	2012	2007	2012
1. Did you worry that your household would not have enough food?	75*	68*	1.61***	1.32***
2. Were you or any household member not able to eat the kinds of foods you preferred?	88***	74***	2.04***	1.57***
3. Did you or any household member have to eat a limited variety of foods?	88***	75***	2.11***	1.66***
4. Did you or any household member have to eat some foods that you really did not want to eat?	89***	75***	2.05***	1.61***
5. Did you or any household member have to eat a smaller meal than you felt you needed?	80***	70***	1.85	1.49***
6. Did you or any other household member have to eat fewer meals in a day?	76***	62***	1.74	1.32***
7. Was there ever no food to eat of any kind in your household?	44**	35**	0.83	0.75
8. Did you or any household member go to sleep at night hungry?	29	28	0.49	0.53
9. Did you or any household member go a whole day and night without eating anything?	59***	28***	1.20***	0.55***
Average HFIAS (0: food secure - 27: severely food insecure)			13.9***	10.8***

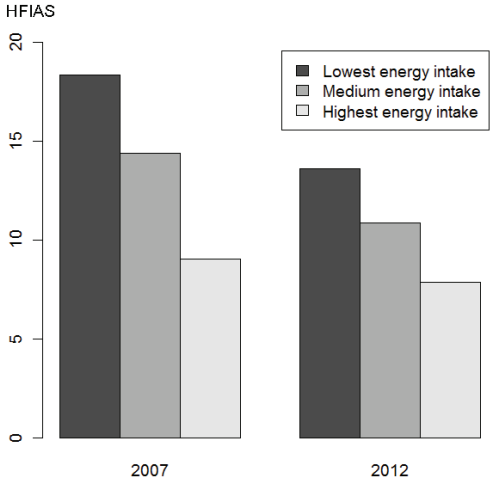
[1]% of households. $n = 314$. P-values obtained with chi-square tests. Symbols indicate significant differences between periods at: ***≤ 0.01, **≤ 0.05, *≤ 0.10

5.3.2 Cross-sectional validity

Before turning to the regression analysis, the association between annual food production and the HFIAS was graphically examined (figure 5.1). Annual food production and the HFIAS were clearly related in the two periods. A classification of households by tertiles based on annual food production, showed that food insecurity decreased significantly as food production increased ($p < 0.01$). The difference in average HFIAS between the first (lowest food production) and the third

tertile (highest food production) was more than nine points in 2007 and just below six points in 2012 ($p < 0.01$). This suggests that the HFIAS is a cross-sectionally valid indicator of food security.

Figure 5.1: The HFIAS in relation to annual food production (in kCal) per capita in 2007 and 20121



Tertiles of daily food production per capita correspond with cut-offs at 1170 kCal and 2690 kCal in 2007 and 1020 kCal and 1960 kCal in 2012. Each group contains 104 or 105 observations.

More formal regression analyses confirmed this finding (table 5.3, models 1 and 4). Food production, cattle ownership and coffee production were positively and significantly correlated with food security in both periods. An increase of food production of 625 kCal/per day per person in 2007 is associated with a decrease of one point in the HFIAS, which is not a negligible effect, given that the mean HFIAS was 13.9 in 2007.

Keeping livestock was strongly and positively associated with food security. This correlation is consistent with cattle being a source of wealth and an important vehicle for saving in an environment characterized by imperfect credit markets (Bundervoet, 2010). Hence, only richer households owned cattle. In addition, their manure is an important fertilizer in an environment where only a few households have access to chemical fertilizers, and soil erosion poses a serious threat to agricultural productivity (Cochet, 2004). The positive association between cattle ownership and food security was more pronounced in 2012 than in 2007, although there is no obvious explanation for this finding. In both surveys, households that were not involved in the cultivation of coffee scored on average 1 point higher on the HFIAS than households that produced the average amount of coffee.

Table 5.3: Analyses with Tobit models of correlation between the HFIAS and farm characteristics in 2007 and 2012

	(1) HFIAS 2007	(2) HFIAS 2007	(3) HFIAS 2007	(4) HFIAS 2012	(5) HFIAS 2012	(6) HFIAS 2012
Constant	18.38***	18.70***	19.80***	14.60***	15.12***	14.70***
Cattle ownership	-4.35***	-3.35***	-4.89***	-7.28***	-6.94***	-6.20***
Working off-farm	2.24***	2.22***	1.38*	1.55	1.53	0.24
Food production (kCal/day)	-0.0016***		-0.0014***	-0.0015***		-0.0025***
Coffee production (kg)	-0.0028***	-0.0023***	-0.0030***	-0.0052***	-0.0050**	-0.0054***
Food production excluding banana (kCal/day)		-0.0021***			-0.0018***	
Banana production (bunches)		-0.0141***			-0.0196	
Enumerator 2			2.11*			7.37***
Enumerator 3			-1.57			7.39***
Enumerator 4			6.67***			7.40**
Enumerator 5			-2.78*			5.61***
Enumerator 6			0.55			-6.87***
Enumerator 7			-5.53**			-1.91
Enumerator 8			-8.07***			1.45
Enumerator 9			-0.78			-7.91***
Enumerator 10			-0.94			-2.55
n	314	314	314	314	314	314

Symbols indicate significant differences at: ***≤ 0.01, **≤0.05, *≤0.10
Regression models were estimated with Tobit models with errors clustered at village level.
Banana production is expressed in harvested bunches: estimated average weight 15 kg/bunch.

Engagement in off-farm activities correlated positively with food insecurity in both years, but the coefficient was only statistically significant in 2007. This correlation is consistent with the assumption that it was mainly very poor, nearly landless, families who worked as paid farm workers; a justifiable assumption as self-reported off-farm wages were very low (around €0.57/day), even by local standards. Off-farm activities should thus be interpreted as a coping strategy for food insecure households, a strategy which has also been documented in Rwanda (Ansoms and McKay, 2010; Rizzo, 2011).

Household size correlated negatively with food insecurity in both periods, but the association was stronger in 2012 than in 2007. Given that food production per capita was included in the regression, this negative associations may indicate that larger households have a higher income from off- and non-farm jobs and are therefore less food insecure.

In order to single out the effect of banana production on food security, banana production was included in the regression analyses as a separate independent variable (table 5.3, models 2 and 5). The correlation between food security and banana production was positive and had a similar magnitude as the correlation between food security and food production[8].

The essential finding of these cross-sectional models is that the correlation between the HFIAS and household and farm characteristics was of similar magnitude in both periods. The only important difference between the models was the constant term. The constant in the 2012 model is about 4 points less than the constant in the 2007 model. Hence, a household with exactly the same production characteristics in 2007 and in 2012 reported feeling more food secure in 2012 than in 2007, despite the decrease in overall food production in the region.

5.3.3 Inter-temporal validity

The preceding analysis has already provided some evidence that households reported feeling more food secure while their total food production decreased. A random-effect Tobit model confirmed this finding (table 5.4, model 1)[9]. The correlation between the HFIAS and the dependent variables is quantitatively similar to the base models (table 5.3, models 1 and 4). The model also confirmed that, after controlling for the covariates, the average HFIAS was more than 4 points

[8] A bunch of bananas was estimated to weigh 15kg, which equals around 16500 kCal. This corresponds roughly to an increase in average daily production of 7.85 kCal per person, given an average family size of 5.76. Given an estimated negative association of -0.0021 (table 5.3, model 2) between an increase with 1 kCal/d/person and the HFIAS, an increase with 7.85 kCal/d/person corresponds to a decrease by 0.016 points of the HFIAS. This is similar to the effect of an additional bunch of bananas which resulted in a decrease in the HFIAS by 0.014 points.

[9] This model also showed that unobservable household characteristics affected the HFIAS. A formal likelihood ratio test rejected that $u_i = 0$ ($p < 0.01$). Hence, a random effect model is preferable to a model that pools all the data and neglects its longitudinal nature.

lower in 2012 than in 2007 ($p < 0.01$). Hence this model rejects the inter-temporal validity of the HFIAS.

Table 5.4: Longitudinal models analyzing correlation between the HFIAS and farm characteristics for different subsamples

	Original sample	Original sample	Sample restricted to households interviewed by enumerators that participated in both rounds	Sample restricted to household interviewed by the same enumerator in both rounds
Constant	18.59***	18.45***	17.59***	21.62***
Cattle ownership	-5.60***	-5.61***	-5.09***	-8.15***
Working off-farm	1.90**	0.98	1.31*	-0.94
Food production	-0.0015***	-0.0015***	-0.0013***	-0.0032***
Coffee production	-0.0032***	-0.0032***	-0.0031***	-0.0039
Year: 2012	-4.53***	-3.64***	-3.43***	-4.04**
Enumerator fixed effects	No	Yes	Yes	Yes
n	314	314	159/148	31

Symbols indicate significant differences at: ***≤ 0.01, **≤0.05, *≤0.10.
Regression models with random-effect Tobit models.
159 and 148 households were interviewed by one of the enumerators that participated in both survey round in 2007 and 2012, respectively.

A difference-in-difference model also rejected inter-temporal validity (table 5.5, models 1 and 2). A household with the same farm characteristics reported a HFIAS score that was on average 4.5 points lower in 2012 than in 2007 ($p < 0.01$). An increase in food or banana production, engaging in coffee farming or acquiring cattle between 2007 and 2012 were all significantly positively associated with an increase in households' food security status.

Households with less members in 2012 than 2007 reported feeling significantly more food insecure, while households with more members did not report any significant change. A decrease in household size from 2007 to 2012 is likely to occur if an adult child leaves the household, limiting the potential of the household to earn an off- and non-farm income and perhaps reducing landholdings if adults sons inherit already part of the land (Van Leeuwen, 2010), leading to increased food insecurity of the household.

Changes in off-farm work, no longer owning cattle or no longer cultivating coffee were not associated with changes in food security status. However, the positive correlations between changes in food security status and changes in production characteristics were too weak to explain many of the changes in the HFIAS between 2007 and 2012. For instance, a household that increased daily production by 1000 kCal/person between 2007 and 2012 reported a decrease in the HFIAS of just

one point less than a household that did not increase its production. Hence, the upward shift in the perception of food security of the households cannot be attributed to changes in food production, cattle ownership, off-farm work, coffee production or household composition.

Table 5.5: Analyses with difference-in-difference model of correlation between changes in the HFIAS and changes in farm characteristics between 2007 and 2012

	(1)	(2)
Constant	-3.27***	-3.26***
Continuous variables		
Change in food production (kCal/d)	-0.00093***	
Change in food production (excluding bananas) (kCal/d)		-0.00012***
Change in banana production (bunches/year)		-0.0076**
Categorical variables ($yes = 1, no = 0$)		
Stopped growing coffee	1.00	1.13
Started growing coffee	-2.59*	-2.38
Acquired cattle	-2.89*	2.70*
No longer owns cattle	2.79	2.61
No longer engaged in off-farm work	-0.90	-0.88
Engaged in off-farm work	0.40	0.49
R^2	8.4%	8.0%

Symbols indicate significant differences at: ***$\leq$ 0.01, **$\leq$0.05, *$\leq$0.10. $n = 314$.
Errors clustered at household level.
Banana production is expressed in harvested bunches: estimated average weight 15 kg/bunch.

5.3.4 Sensitivity analyses

Enumerator bias

The inclusion of enumerator-specific dummies revealed that some enumerators consistently over or underestimated households' food security in both periods (table 5.3, models 3 and 6). For instance, the HFIAS of households interviewed by enumerator 3 in 2012 was on average 7.4 points higher than the average score of households interviewed by enumerator 1. Estimates for this categorical variable cannot be compared between models 3 and 6 because not all enumerators participated in both rounds. It should also be noted that these additional dummies did not capture location-specific effects, because enumerators interviewed different households within the same village. The inclusion of enumerator-specific dummies did not considerably affect the estimates of the main independent variables, but only improved the explanatory power of the models. In addition, constants in both models were quite similar to the base models. It seems that differences between enumerators mattered, but that all enumerators were able to discriminate between food secure and food insecure households. However, these regressions showed that scores on a subjective measure could be severely biased as a result of enumerators' divergent interpretations of the questions. This aspect of subjectivity also has

to be considered when choosing between using the HFIAS and other measures of food insecurity.

The differences between enumerators in assessing households' food insecurity status are unlikely to explain the important finding of the lack of inter-temporal validity of the HFIAS (table 5.4, models 2 to 4). The fixed-year effect remained negative and highly significant in the random-effect Tobit models that included enumerator-fixed effects (model 2), restricted the sample to enumerators that participated in both rounds (model 3) or to households that were interviewed by the same enumerator in both rounds (model 4). Only 31 households were interviewed by the same enumerator in both survey rounds. This sample size was too small to estimate a meaningful difference-in-difference model as an additional robustness check.

Measurement error in food production[10]

In the previous analyses, the subjective indicator of food insecurity, the HFIAS, was assessed using total food production as the gold standard of food security. The HFIAS is more prone to a subjective interpretation by enumerators or respondents than 'objective' data such as total yearly food production, which can, theoretically, be determined exactly. However, objective data are more susceptible to measurement error than subjective data, which is one of the main reasons why simple food security indicators are so popular. For this reason, 'objective' indicators can also be considered as 'subjective' because they may be subject to reporting bias.

Measurement error is especially a concern for the variable 'food production' because this variable is based on one-year recall by the household head and is therefore likely to be measured with substantial error (Beegle et al., 2012). Even more worryingly is the possibility that measurement error is not random: production of households with a higher food production might be underestimated, whereas households producing less food are probably more likely to remember exactly how much they harvested in the previous year. Systematic measurement error would bias the coefficients of the regressions and may cause spurious results. For the same reason, the amount of coffee harvested may also be susceptible to measurement error.

To correct for bias due to measurement error, we applied an instrumental variable approach. Total, annual food production of a household was instrumented by its total land holdings and the share of land located on the hills (and not in the marsh lands, which are more fertile). Every plot of cultivated land was measured with GPS and we are therefore rather confident that this variable has been correctly measured and is not influenced by the enumerator who conducted the fieldwork. Instead of total coffee production, we included a dummy that equalled one if

[10]The following two sections are not included in the published version of this paper

the households had cultivated coffee in the previous year. Again, this variable is unlikely to be wrongly reported.

Results of the (instrumented) Tobit models in 2007 and 2012 and an instrumented random-effect model are shown in table 5.6. This shows that, for all models, the negative association between food production and food insecurity still holds, even after correcting for measurement error. In contrast to our expectations, the negative association between food production and HFIAS was even more pronounced if production was instrumented compared to the base models. It might be that the instruments also capture coffee production as the dummy for coffee production was no longer significant in the IV-models, which would explain the stronger negative association between food production and HFIAS than in the base models. Another explanation is that annual food production is imprecisely measured. It is well-known that random measurement error in an explanatory variable causes attenuation bias, that is, a bias of the estimated coefficients towards zero (Carroll et al., 2012). By instrumenting the noisily measured variable 'annual food production', attenuation bias is reduced, which may explain the stronger correlation between the HFIAS and total food production. Importantly, the year-fixed effect remained highly significant in the IV random-effect model (table 5.6, model 6). Hence, inter-temporal validity is also rejected by the IV-model.

In conclusion, the additional IV-analyses show that the results were not caused by measurement error in the explanatory variables. The subjective nature of the HFIAS remains therefore the main suspect to explain the lack of inter-temporal validity of the HFIAS.

Table 5.6: Correcting for measurement error in the variable 'total, annual food production' with an IV-approach

	(1) Tobit model 2007	(2) IV Tobit model 2007	(3) Tobit model 2012	(4) IV Tobit model 2012	(5) Random-effect model	(6) IV random-effect model
Constant	46.37***	93.31***	37.10***	65.20***	39.92***	72.65***
Cattle ownership ($yes = 1, no = 0$)	-4.03***	-1.33	-7.03***	-5.58***	-4.23***	-2.68***
Working off-farm	2.45***	1.29	1.55	1.52	1.80***	1.47**
Log of food production (kCal/d)^2	-4.24***	-10.75***	-3.42***	-7.41***	-3.31***	-7.81***
Coffee farmers ($yes = 1, no = 0$)	-1.86**	0.73	-2.34**	-1.84	-1.89***	-0.78
$Year = 2012$					-3.55***	-4.83***
n	314	311	314	312	314	623

Symbols indicate significant differences at: ***$\leq$ 0.01, **$\leq$0.05, *$\leq$0.10.
We took the natural logarithm of food production/capita and land because this reduced the impact of outliers and resulted in a better performance of the IV in the first stage.
Instruments for log of calorific value of production: log of land, log of land squared, share land located on the hill.

Food aid

An increase in food aid between 2007 and 2012 could explain the lack of intertemporal validity of the HFIAS. If food aid increased substantially during the period, households could be less food insecure, while at the same time producing less food. Unfortunately, the survey did not include a question on food aid.

We obtained data on food aid at communal level in both 2007 and 2012 from the World Food Program in Burundi[11]. The WFP also coordinates food aid programs of partner organizations such as CARE. Total food aid and the number of individuals reached remained more or less constant between 2007 and 2012 (table 5.7). Given an estimated population of 660 717 people in Ngozi (most recent census), an individual received on average 4.16 kg in 2007 and 3.89 kg in 2012 in food aid. Consequently, we believe that food aid cannot have played an important role in the food security status for most households in the sample.

Table 5.7: Food aid provided by WFP and partners in Ngozi, Burundi in 2007 and 2012

	2007	2012
Food aid (in ton)	2750	2570
Beneficiaries	447 000	423 000
Kg per capita/beneficiary	6.15	6.07
Kg per inhabitant of Ngozi	4.16	3.89

In 2012, 70% of the individuals who received food aid were children in primary and secondary education as part of a school feeding program and 23% of the individuals were pregnant women and children under five years old. Hence, there is no evidence that the most food insecure households were targeted by food aid programs. The data for 2007 were less detailed, and we could not determine exactly if food aid was targeted to a specific group.

This data, combined with population data, allowed us to construct a new variable at the communal level to capture average food aid per capita (summary statistics not shown). As a robustness check, this variable was included in the Tobit and random-effects models (table 5.8). Food aid was significantly, positively associated with food insecurity in 2007, but was insignificant in 2012. This might indicate that in 2007 the most food insecure 'communes' were targeted, while all school children had access to school feeding programs in 2012, independent of the food security situation at the communal level. To control for a different correlation between food aid and the HFIAS in 2007 and 2012, we included an interaction term between food aid and the year-effect in the panel model (table 5.8, column 3). Even after taking food aid into account, households still reported feeling more food secure in 2012 than 2007.

[11] Personal communication, February 2014, data is not publicly available

Table 5.8: Correlation between the HFIAS and food aid

	(1) Tobit model 2007	(2) Tobit model 2012	(3) Random-effect Tobit model
Constant	18.44***	14.61***	18.62***
Cattle ownership	-4.07***	-7.35***	-5.48***
Working off-farm	2.26***	1.49	1.89**
Food production (kCal/d)	-0.0017***	-0.0015***	-0.0016***
Coffee production (kg/year)	-0.0036***	-0.0056**	-0.0039***
Food aid (kg/capita)	0.76**	-0.06	0.70**
Food aid x $year = 2012$			-0.72**
$Year = 2012$			-4.49***
n	314	314	314

Symbols indicate significant differences at: ***≤ 0.01, **≤0.05, *≤0.10.
The variable 'food aid' was centered around its mean.

Internal validity of the HFIAS

The previous analyses discussed the external validity of the HFIAS. In other words, they examined whether the HFIAS is correlated with another indicator of food security, that is, total food production. Internal validity is a second important criterion that every indicator needs to meet. It requires that the nine questions of the HFIAS measure a same, underlying construct. This can be tested with principal component analysis (PCA) or with psychometric models such as Rasch modelling (see chapter 6 for an application of this technique). The results are briefly discussed in the next section. For the sake of conciseness, we do not report the results of these analyses here. They were published as online supplementary material alongside the published version of this paper (Desiere et al., 2015a).

5.4 Discussion

This study shows that the HFIAS is a cross-sectionally valid indicator of food security. This is in line with the literature. However, its inter-temporal validity can be questioned, because the self-reported food security status of households increased despite food production decreasing between the two surveys. This finding has not been often reported before in the literature on food insecurity indicators. Hence, we closely examine the factors that might invalidate this conclusion.

An important assumption in this study is that food production is strongly correlated with food consumption: we did not collect detailed food expenditure data or food intake data. Several studies indicate that Burundian farmers mainly produce for subsistence purposes (Detry, 2008). This is confirmed by our data. For instance, 35% of the households sold sweet potatoes in 2012, and these households

sold on average less than 30% of their harvest. Only coffee and bananas were extensively marketed. Moreover, food production is the main source of wealth in rural Burundi (e.g. it is strongly correlated with assets such as land) and is thus expected to be significantly correlated with food security.

A second, closely related assumption concerns the timing of food consumption. The HFIAS only probes into households food security status over the last four weeks. As total food consumption in the four weeks before the interview is approximated by total, yearly food production, the lack of inter-temporal validity would also be observed if most households consumed considerably more in the four weeks before the interviews were conducted in 2012 than in 2007. This would explain the lack of inter-temporal validity, although it would not contradict a general decrease in food production over time. The fact that interviews were conducted in the same month in both rounds of data collection can only partially mitigate this concern. However, the finding that yearly food production is strongly associated with the HFIAS in both periods suggests that this possibility is probably not driving the results[12]. Future studies would ideally make use of detailed food intake data to avoid this caveat.

The main finding of lack of inter-temporal consistency of the HFIAS hinges on the observation that food production decreased or, at least, did not increase between 2007 and 2012. This decline is confirmed by secondary datasets. The Food Balance Sheets published by the FAO showed that food supply per person per day in Burundi decreased from 1656 kCal in 2007 to 1604 kCal in 2009. Similarly, an aggregation of the total food production based on the main staple crops (published by FAO) shows that the total production in Burundi did not increase between 2007 and 2011, while the population grew considerably. Simple calculations based on these figures showed that food production per person per day decreased from 2295 kCal in 2007 to 2127 kCal in 2012. A website recently launched by the government of Burundi (in close collaboration with the FAO) provides agricultural statistics at the provincial level (CountryStat-Burundi, 2013). The reported trends of food production in Ngozi corroborate our findings. They found a 60% decrease in banana production and an 80% decrease in the production of sweet potatoes between 2007 and 2012, while production of the other main crops remained more or less stable. It should, however, be mentioned that the reliability of these figures is difficult to check. Nevertheless, we are fairly confident that the decrease in food production and, hence, total income and food consumption is a region-wide phenomenon.

Another competing explanation for the improvement in the perceived food security status of the households between 2007 and 2012 is an increase in food aid. However, food aid in Ngozi provided by the World Food Program (WFP) and its partners decreased, from 2750 tonnes in 2007 to 2570 tonnes in 2012, which

[12] This problem would still hold if we had conducted the interviews on exactly the same day in 2012 as in 2007, instead of only during the same month.

corresponds to 4.16 and 3.89 kg per capita, respectively[13]. Moreover, there is no indication that food aid programs were better targeted at food insecure households in 2012 than in 2007. In 2012 more than 70% of resources were devoted to school feeding programs, which provide all primary schoolchildren with a daily, free meal, independent of their food security status. A final possibility is an increase in remittances between 2007 and 2012. However, the importance of this livelihood source is likely to be small as only five households in the sample claimed to receive remittances.

This study did not assess the internal consistency of the HFIAS, as this aspect of the indicator has already been validated in previous studies (Marques et al., 2014). In other words, we did not conduct a thorough psychometric evaluation of the HFIAS to evaluate whether the nine questions included in the HFIAS measure a one dimensional latent trait. If these questions do not reliably measure food security, this could explain the lack of inter-temporal validity. For instance, our findings could be biased if a question is interpreted differently by the respondents in 2007 than 2012. A careful analysis revealed that question 9 may have been interpreted differently in 2012 than 2007, with more households reporting 'going a whole day and night without eating' in 2007 than expected (results of this analysis are available upon request). This question contributed significantly to the overall decrease in the HFIAS from 2007 to 2012 (see table 5.2). However, even if we would exclude this question from the indicator, we would still observe an increase in perceived food security from 2007 to 2012. This suggests that lack of stability of the questions may partially explain the lack of inter-temporal validity, but it is unlikely to be the only reason. Additional psychometric studies are nevertheless required to examine this point in more detail.

A study in Ethiopia (Maes et al., 2009), which found a similar inconsistency over time, stresses the possibility of 'observation bias' and 'response shifts'. The former might occur if respondents pretend to be more food insecure than they really are in the first round of a survey because they expect that less food secure households will receive food aid. In the second round, households would respond more honestly, reporting their 'true' food security situation. We believe that this bias is likely to be limited in our study because respondents were well informed on the research aim at the start of the interview. Moreover, a very limited number of international NGOs are active in the area, and therefore respondents do not expect any food aid.

Finally, response shifts might arise if respondents shift their internal standards as their living conditions change over time. This theory predicts that individuals assess their well-being not only by comparing their current situation with the past but also by gauging their relative position within their community (Günther and Maier, 2014; Sprangers and Schwartz, 1999). This lack of a common reference

[13]Personal communication with the head of the WFP in Burundi, February 2014.

frame, both over time and between poor and rich households, is a general limitation of subjective indicators (Headey and Ecker, 2013). The study in Ethiopia (Maes et al., 2009) pointed to this phenomenon to explain why volunteer HIV/Aids caregivers, frequently faced with individuals even worse off than themselves, reported feeling more food secure even though their food security situation (measured objectively) deteriorated. Objectively, these caregivers had indeed become more food insecure over time, but they were less affected than the households that they visited regularly and therefore felt more food secure. A similar effect may be at play in our study area. Agricultural production decreased in the entire region (primarily caused by the loss of banana trees), but, given the limited migration rates and lack of communication infrastructure, only a few households had access to information about living standards in other provinces in Burundi or in the capital. It is therefore likely that respondents compared their food security situation with that of their neighbors and evaluated their position within the local community instead of comparing their current situation with the past. This would simultaneously explain the cross-sectional validity and lack of inter-temporal validity of the HFIAS. Similar patterns have been found in research on happiness (D'Ambrosio and Frick, 2012; Luttmer, 2005; Fafchamps and Shilpi, 2008). Recently, new measures of poverty have been proposed that explicitly take into account this reference-dependent utility (Günther and Maier, 2014).

5.5 Conclusion

The development of the Household Food Insecurity Access Scale (HFIAS) is an attempt to construct an indicator of food insecurity that is internally, cross-culturally, cross-sectionally and inter-temporally valid and that captures all aspects of food insecurity. Moreover, this indicator needs to be user-friendly so that food insecurity in rural areas can be easily monitored by NGOs and governments. Harmonizing these ambitious, and sometimes contrasting, objectives is a major challenge.

Results from this study in the north of Burundi confirmed the cross-sectional validity of the HFIAS, as it is significantly correlated with annual food production, livestock keeping, off-farm work, coffee production and household size. However, we are less convinced about the inter-temporal validity of the index, as perceived food security increased while total production declined over the same time period. As this is one of the first studies investigating the inter-temporal validity of this indicator of food insecurity over a long time period, additional studies are needed to confirm (or refute) our results in different settings. In particular, follow-up studies should use detailed food intake data, rather than data on annual food production, to assess the inter-temporal validity of experience-based indicators. The main shortcoming of this study is indeed the assumption of a (strong) correlation between food consumption and production. In other words, it was assumed that

total annual food production is a good proxy for the 'true' food security status of a household. The validity of this assumption can be questioned. Future research thus requires panel datasets (or large cross-sectional datasets) that include food intake data. In the near future the dataset of the World Gallup Poll (Ballard et al., 2013), which includes an experience-based food security indicator since 2014, could be used to test the inter-temporally validity of subjective indicators more rigorously.

The findings reported in this study suggest that detailed production and consumption data will remain indispensable in the examination of the dynamics of food security. Consequently, studies which assume the inter-temporal validity of subjective indicators should be interpreted carefully, as this assumption is questionable (Headey, 2013; Verpoorten et al., 2013). Finally, the results raise the question of what the HFIAS actually measures and how households assess their own food security situation. Part of the answer might lie in 'response shifts' in which respondents reassess their internal standards over time due to a general decrease of the living standards within their community. This is an interesting avenue for further research.

CHAPTER 6

Verifying validity of the Household Dietary Diversity Score: an application of Rasch modelling

Abstract: The Household Dietary Diversity Score (HDDS) was developed to measure household food access, one of the levels of food security. Previous research has shown that dietary diversity is related to food security. However, the validity of the HDDS in the form developed by the FANTA project - twelve food groups, 24-hour recall - has never been verified. Using data from 1015 households in Colombia and Ecuador, the internal validity of the HDDS was assessed with Rasch models. In other words, it was evaluated whether the twelve food groups consistently measured the same latent trait, that is, food access. The different dietary patterns between Colombia and Ecuador and two cultural groups within Ecuador required data to be split into three subgroups. This shows that the HDDS cannot be used to compare food security between culturally different groups. Even within the homogenous groups, there was only a limited fit between the food groups and the underlying latent trait. Some food groups were even negatively correlated with the latent trait, implying that the probability of consuming certain food groups decreased with the probability of consuming the other groups. The findings warrant against using the HDDS as sole indicator of food access.

This chapter will be published as:
Vellema, W., Desiere, S., D'Haese, M., 2016. Verifying validity of the household dietary diversity score: an application of rasch modelling. Food and Nutrition Bulletin.

6.1 Introduction

While the definition of food security formed at the 1996 world food summit (FAO, 1996) is widely adopted, disagreement remains on the indicators that assess, quantify and qualify food security and on how to operationalize these indicators at national, household or individual level (Webb et al., 2006; Pinstrup-Andersen, 2009; Jones et al., 2013; Leroy et al., 2015; Cafiero et al., 2014). Food security is measured in different ways. For example, anthropometric measures are used to monitor growth of children under five (Pinstrup-Andersen, 2009); recalls of food consumed in the past 24 hours or over a longer reference period are recorded to measure intake of macro- and micronutrients (Kennedy et al., 2010); and data on food expenditure is used to define food poverty lines (Rose and Charlton, 2002); while experience-based responses such as the Household Food Insecurity Access Score (HFIAS) elicit perceived consequences of not having enough food (Jones et al., 2013). Research institutions and development organizations alike apply such indicators to identify food insecure households or analyse effects of interventions on food security (Jones et al., 2013).

The Household Dietary Diversity Score (HDDS) is a frequently used indicator of food security. It was developed as a quick-to-implement and easy-to-use survey-based indicator to measure the impact on household food access of programs with improvements in food security as their core objective (Swindale and Bilinsky, 2006). The second version of the accompanying guide mentions that "An increase in the average number of different food groups consumed provides a quantifiable measure of improved household food access. In general, any increase in household dietary diversity reflects an improvement in the household's diet" (Swindale and Bilinsky, 2006, p6), which suggests the HDDS might be used as a household-level indicator of food security - indeed, it is frequently used as such (Leroy et al., 2015; Cafiero et al., 2014). However, the validity of the HDDS has never been verified, making it impossible to substantiate claims that it is a useful indicator of food security. The objective of this paper is to fill this glaring gap.

Rasch models were used to verify the internal or construct validity of the HDDS. These models were specifically developed to test whether an additive scale consisting of several items measuring a single underlying construct meets the criteria required for interval scale measurement (Rasch, 1960). This approach differs from most other statistical techniques in that it starts from a mathematical model which meets the required criteria and tests the extent to which the data fits the model. When the data does not fit the model, it is not the model but the data which is considered wrong. By assessing the deviations of the HDDS from the criteria, specific shortcomings of the indicator can be highlighted. In effect, Rasch analysis provides the lens through which we look at the internal functioning of the indicator. Applying this methodology to analyse the construct validity of the HDDS is the main contribution of this paper to the literature.

6.2 Household Dietary Diversity Scores

Dietary diversity refers to the variety of foods consumed by individuals or households (Ruel, 2003; Jones et al., 2013). An indicator of dietary diversity is a particularly interesting way to measure food security, because it is simple to implement, can be administered at household and individual level, and is a useful outcome in itself (Hoddinott and Yohannes, 2002). There is a shortage of validity studies of survey-based dietary diversity indicators, especially regarding the way questions are posed and how these are handled and interpreted (Ruel, 2003; Leroy et al., 2015). Particularly pressing issues are the responsiveness of food security indicators to improved food security, their discriminatory power in distinguishing food secure from food insecure households, and their validity across different cultural settings.

When measured at an individual level, dietary diversity scores are generally found to be a good proxy for micronutrient adequacy (Hatloy et al., 1999; Arimond and Ruel, 2004; Steyn et al., 2006; Kennedy et al., 2007; Moursi et al., 2008; Arimond et al., 2010). Dietary diversity might not only be linked to dietary quality, but also imply dietary quantity. According to Bennett's Law, as people become wealthier they switch from starch-dominated diets to more varied diets including vegetables, fruit, dairy products, and meat (Bennett, 1941). Although calorie intake might not increase above a certain level of wealth, Jensen and Miller (2010) suggest people quickly shift to improving the taste of their food bundle when their incomes increase. Their findings are in line with classic theories of demand (Maslow, 1943). In other words, households with sufficiently diverse diets can be assumed to at least consume enough food not to be hungry. Studies confirm a positive relationship between household dietary diversity and household food security (Hoddinott and Yohannes, 2002; Faber et al., 2009; Kennedy et al., 2010; Headey and Ecker, 2013). However, these studies were based on indicators differing in regard to their inclusion of individual foods versus food groups, number of food groups, weights, and recall period, making it hard to establish a definitive link. In fact, some authors even question what it is that is being measured by these indicators (Ruel, 2003; Headey and Ecker, 2013; Cafiero et al., 2014).

In particular, only two research papers are named on which the conclusion that "an increase in dietary diversity is associated with socio-economic status and household food security" is based (FAO, 2012b). Of these papers, Hatloy et al. (1999), in a case study in a southern county of Mali, indeed find such an association for socio-economic status. For nutritional status, the association was only found in urban areas. Furthermore, their index for dietary diversity is based on ten food groups, not the suggested twelve. Perhaps the most extensive work on this topic is by Hoddinott and Yohannes (2002), who study the relationship between dietary diversity and a range of food security measures using datasets covering both rural and urban households from 10 poor or middle-income countries. The authors find

a robust positive relationship independent of whether individual foods or food groups are used to measure dietary diversity which holds over urban and rural areas, seasons, and recall period. However, in neither of these studies is the HDDS indicator used in the form promoted in the guidelines.

Dietary diversity is measured by counting the number of foods or food groups consumed over a certain reference period. These groups can be simply counted or a weight can be attached to them based on their nutritional value. Some indicators also take into account the frequency at which the foods were consumed, or specify a minimum portion size required for a food to be counted in the index (see Ruel (2003) and Leroy et al. (2015) for a review of different indicators). Of the food-group indicators, the HDDS analysed in this paper is probably the most widely used by development organizations. It was developed by the Food and Nutrition Technical Assistance (FANTA) and actively promoted by USAID. Moreover, this index is the basis for the recent FAO "Guidelines on measuring household and individual dietary diversity" (FAO, 2012b).

The HDDS was developed to measure household food access and designed to be an easy-to-use and quick-to-implement index, making it ideal for impact evaluations of development programs (Swindale and Bilinsky, 2006). It measures dietary diversity by counting the number of food groups that were consumed by the household over the last 24 hours. The indicator consists of twelve food groups: cereals; roots and tubers; vegetables; fruits; meat, poultry, and offal; eggs; fish and seafood; pulses, legumes, and nuts; dairy products; oils and fats; sugar and honey; and miscellaneous, such as condiments. These twelve food groups are based on the groups used to construct the FAO's food balance sheets (Swindale and Ohri-Vachaspati, 2005). The value of the HDDS equals the number of food groups consumed in the last 24 hours. A higher score should reflect higher dietary diversity and hence better household food access (Swindale and Bilinsky, 2006).

This paper is the first to evaluate the validity of the HDDS in the form promoted in the FANTA guidelines. We limit ourselves in scope to evaluating the construct validity of the indicator, i.e. whether the different food groups contribute to a single underlying construct in such a way that the overall score on the indicator can be interpreted as an interval scale measure at household level. We do not analyse whether the scale indeed measures household food access but follow the indicator guidelines in assuming that it does. In other words, we do not study what is measured by the HDDS, but verify how it measures.

6.3 Data

The construct validity of the HDDS was tested using data obtained from the baseline from a cross-border agricultural development project in Colombia and

Ecuador. Basing our study on such 'real' data, rather than on data collected primarily for the validation of food security indicators makes the results of our study more realistic. It also explains why our dataset did not include other indicators of food security such as food intake data or anthropometric data that could have been used as benchmarks against which the HDDS could have been assessed.

Colombia and Ecuador are culturally close and economically similar. Both countries are considered upper-middle income countries according to the World Bank classification, yet have high inequality and poverty rates. Data was collected in the Ecuadorian amazon basin and the southern mountain range in Colombia, which are among the poorest parts of the countries. In the Amazon basin 59.7% of the population lives below the national poverty line (INEC, 2006); in Colombia's southern Andes, 50.6% of the population lives below the national poverty line (DANE, 2011).

Data was collected in April and May 2012 through structured questionnaires, with interviews conducted by local enumerators which were trained and supervised by permanent staff of the International Center for Tropical Agriculture (CIAT). Interviews were conducted with pen and paper, and data entry and cleaning took place at CIAT headquarters in Palmira, Colombia.

All interviewed households were small-scale farmers, depending on agricultural production for most of their income. Respondents were either the head of the household or the person most closely related to the head of the household, like a spouse. Since the data was collected for the baseline of a development project that aimed to increase food security and household income, the number of households in treatment and control group was selected to detect a 'modest' impact of the program on the outcome variable. It is fair to say that the power calculations were rather imprecise because key information such as the variance of household income was lacking for the region under study (see Vellema et al. (2015) for details of the power calculation). In Ecuador, sampling of project beneficiaries was done by stratification at cantonal level based on a list of inhabitants obtained from the national institutes of statistics (INEC). In Colombia, stratification was done at municipal level, which corresponds to the cantonal level in Ecuador, i.e. the administrative level below province (which are called departments in Colombia). The stratification was based on member lists of the national federation of coffee producers (FEDECAFE). In total, 510 households were interviewed in Colombia, and 514 in Ecuador. After removing observations for non-response, the full dataset contained 509 Colombian and 506 Ecuadorian households.

Interviews were conducted according to a detailed standardized protocol; enumerators received two weeks of training including field trials before starting data collection. Data was collected on family composition, including ethnicity of household members, and income. Agricultural production destined for own consumption was valued at farm-gate prices. The used HDDS surveys were made more specific

Table 6.1: Descriptive statistics

	Colombia	Ecuador	
		Kichwa	Immigrants
Family size	4.12	6.26	4.84
Income (USD)	5939	1331	2196
HDDS	8.06	5.26	6.8
n	509	209	297

Mean of selected variables.
Values for Colombia converted from Colombian Pesos using exchange rate of 31 May 2012.

for each country by adding commonly consumed foods to the specification of the food groups. For example, food group 1, cereals, was specified for the Ecuador survey as 'In the last 24 hours, did you consume any kind of cereal like rice, maize, or wheat, or any product made from cereals, such as bread, cookies, humitas, etc?'. For Colombia, this question was specified as 'In the last 24 hours, did you consume any kind of cereal like rice, maize, or wheat, or any product made from cereals, such as bread, arepas, envueltos de choclo, noodles, puff pastries, toast, cakes, or any other food made from millet, sorghum, maize, rice, wheat, barley, oats, etc.?' Descriptive statistics are shown in table 6.1. For the analysis, the data from Ecuador had to be split into two cultural groups, Kichwa and migrant households, as will be explained in the results section. For legibility, these groups are represented separately in the table.

6.4 Methodology

Rasch models were developed by Rasch (1960) to measure an individual's level of a latent trait. The models assume that the probability of an individual's response to a question depends only on item difficulty and individual ability. In this study, the latent trait is assumed to be household food access, as suggested the developers of the HDDS (Swindale and Bilinsky, 2006). The food groups making up the indicator are the items. Rasch models do not depend on a priori assumptions about item difficulty. Rather, item difficulty is an outcome of the analysis. Rasch models are most frequently applied in education and psychology, but commonly used in other human sciences (Bond and Fox, 2001), and increasingly applied to medical research.

Rasch models have been used to study food security indicators before. They have been applied to test experience-based indicators, such as the core food security module (CFSM) developed by the US Department of Agriculture (Derrickson et al., 2000; Opsomer et al., 2003), Latin American Household Food Security Mea-

surement Scale (ELCSA) (Toledo Vianna et al., 2012), Household Food Insecurity Access Scale (HFIAS) (Deitchler et al., 2010), and most recently, the Arab Family Food Security Scale (Sahyoun et al., 2014). Rasch models allow evaluating whether items are equally difficult in different cultural settings because estimated item parameters are not sample specific (Salzberger et al., 1999; Casillas et al., 2006).

Rash analysis assumes hierarchical ordering of items. In the context of the HDDS, this implies that households consuming the most difficult item i.e. the food group eaten only by those households with high food access - should also consume easier items. Although there is an extensive literature on dietary patterns which concludes that households shift to more expensive foods when their income increases (Thorne-Lyman et al., 2010), implying some hierarchy between food groups, it is not clear to which extent this hierarchy is accurately captured by the food groups as defined in the HDDS. The hierarchical ordering of items is essential for the applicability of Rasch modelling, but cannot be tested directly. Not meeting this key assumption has several consequences, which are explained in the discussion section.

Two other conditions an indicator of food access should meet in order to be a valid and reliable proxy of the latent trait, household food access, could be tested directly by using Rasch analysis. First, the indicator needs to be robust to cultural differences. Hence, conditional on the latent trait, item difficulty should be consistent between countries, cultures, and food habits. Second, the probability of an affirmative response to an item (food group) needs to be stable over the latent trait, such that each food group contributes positively and significantly to the overall score on the indicator. These conditions are necessary for the indicator to reliably distinguish households with high food access from households with low food access and to allow cross-cultural and inter-temporal comparison of households based on the HDDS.

Its most simple form, the 1PL Rasch model (equation 6.1), is based on the assumption that the probability of an affirmative answer to item i (e.g. consumption of a food group) by person p is determined by the difference between the person's ability θ_p (e.g. its food access status) and the difficulty of the item, β_i. In other words, the higher a person's food access status and the less 'difficult' a particular food group is, the more likely it is that this person is consuming that particular food group. Formally, the 1PL model is specified as follows:

$$ln\frac{P_{pi}}{1-P_{pi}} = \theta_p - \beta_i \tag{6.1}$$

This formula states that the log odds of the probability of an affirmative response of person p to item i is a linear function of the ability of person p (θ_p) and the difficulty of question i (β_i).

A poor item fit might indicate that the item does not measure the same latent trait as the other items, but it might also indicate that the item is not as strongly correlated with the latent trait as the other items. A simple 1PL Rasch model assumes all food items are equally informative of a household's ability. The more flexible parameterization of the 2PL model allows testing the correlation of item i with the latent trait, by adding an interaction term, α_i:

$$ln\frac{P_{pi}}{1-P_{pi}} = \alpha_i\theta_p - \beta_i \tag{6.2}$$

The additional parameter, α_i, determines the discriminatory power of the items, i.e. it measures the extent to which an item helps to distinguish high from low performers. The larger is α_i, the more a small increase in θ increases the probability of an affirmative response to item i.

For interval scale measurement, each item should contribute positively to the latent trait, such that food access status increases with the consumption of each food group. In terms of the model, this implies $\alpha_i > 0$. If α_i is not significantly different from zero, the probability of an affirmative response is no longer a function of θ. This implies that an individual with a highly diversified diet could not be distinguished from a household with a less diversified diet. More worrying are items (food groups) with a negative α_i. Such items showed an inverse relationship with the latent trait, implying that the probability of consuming food group i decreased with increasing food access. As the HDDS score equals the number of consumed food groups, food groups with an inverse relation with dietary diversity will bias HDDS downwards. Clearly, such items should not be included in a valid indicator.

A necessary pre-condition for any scale is that item response (food group consumption) should only depend on ability, not on any other individual or household-specific characteristic. This pre-condition was checked using Differential Item Functioning (DIF) tests, which allows testing whether individuals with the same latent trait but different consumption preferences respond differently to items (Pallant and Tennant, 2007; Tennant and Conaghan, 2007). Consumption preferences are likely to differ between cultures and regions. For example, fish consumption might be common in coastal areas, but is linked to a highly diversified diet in rural areas. To verify this condition, prior knowledge of dietary patterns in the region was required.

For each subgroup of households in the sample, a refined indicator was constructed based on the relationship between individual items and overall score on the indicator. In a first step, food groups consumed by nearly all or none of the households were removed. Such items did not add value in distinguishing households with high food access from households with low food access. Furthermore, items with less than ten observations per binary choice alternative might cause estimates to become unstable (Linacre, 2002) and hence were removed. Second, the relationship

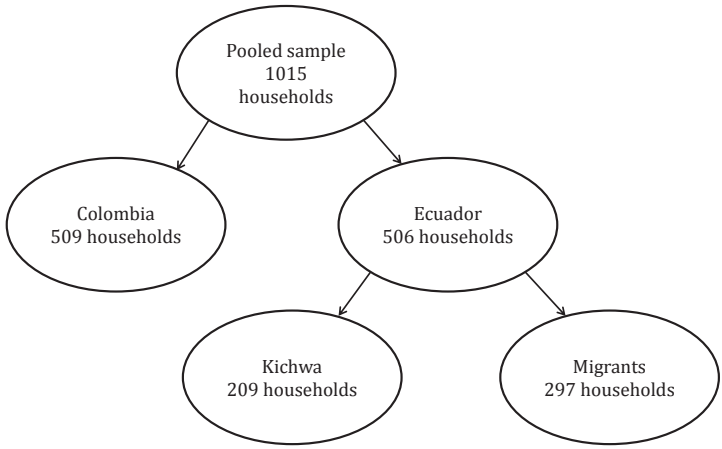
of the remaining items with the total score on the indicator was visually verified with Item Response Functions (IRFs). Well-functioning items should have a significant positive relationship with the overall score, indicating households consuming the food group had a higher probability of having higher food access. Badly functioning items were removed from the refined indicator. Item fit was further examined using item characteristic curves (ICCs), which show expected and observed probabilities for each item in a single graph (Bond and Fox, 2001).

The resulting refined indicators were tested for robustness and local independence. Robustness was checked by removing observations with low person-fit and verifying whether the ranking of items differed between the reduced sample and the full sample. Local independence was tested for by the significance of the correlation between response pairs (Ponocny, 2001; Tennant and Conaghan, 2007). All equations were estimated using Rasch analyses performed using R version 2.12.1, with packages irtoys and eRm (Mair and Hatzinger, 2007; Partchev et al., 2009).

6.5 Results

Consumption patterns of Colombian and Ecuadorian households were completely different, as is evident from tables 6.2 and 6.3. Hence, separate Rasch analyses were performed for each country. Differential Item Functioning showed the existence of distinct dietary patterns for Kichwa and immigrant households in Ecuador, requiring separate analyses for these two subgroups. Such a difference was not found in the Colombian sample. Therefore, three distinct analyses had to be performed, as shown in figure 6.1. The consequently large amount of analyses performed implies that not all results could be reported in the main text. The full results of the estimation of the 2PL models is presented in appendix 6.A. In the next section (6.5.1), results of the analysis for Colombia will be discussed, followed by those for the DIF analysis in Ecuador (section 6.5.2) and the HDDS verifications for Kichwa (section 6.5.3) and migrant households (section 6.5.4).

Figure 6.1: Division of the sample in three groups



6.5.1 Colombia

Food groups consumed by nearly all or very few households reduce the variation of the HDDS indicator and hence its efficiency. In the Colombian sample, this lack of variation was cause for concern: 99% of households consumed the food groups 1 (cereals), 2 (roots and tubers), 11 (sugar/honey) and 12 (other) during the 24 hours before the survey (table 6.2). The nearly uniform consumption of these food groups meant they did not add explanatory power in differentiation between households with high and low food access. Therefore, their removal did not make the overall indicator less precise but was necessary to ensure stability of the estimates of the model (Linacre, 2002). The relationship between individual

Table 6.2: Food group consumption in the Colombian sample

	Food group	% of households
1	Cereals	99
2	Roots and tubers	99
3	Vegetables	49
4	Fruits	50
5	Meat	67
6	Eggs	66
7	Fish	6
8	Legumes	62
9	Milk/diary	23
10	Oils/fat	86
11	Sugar/honey	99
12	Other	99

Food groups that were excluded from further analysis are underlined.

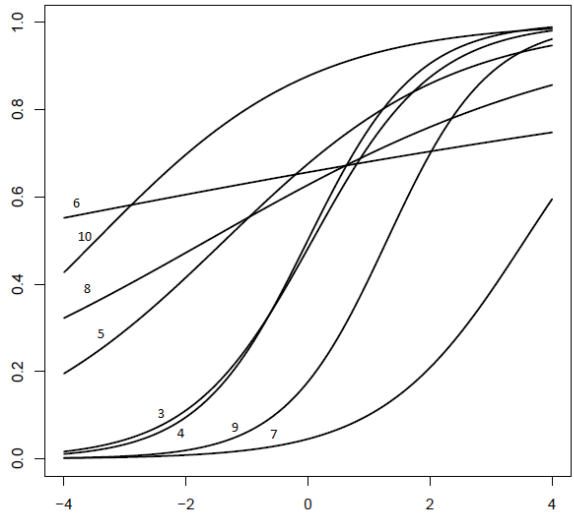
items and the overall score was evaluated with Item Response Functions (IRFs) of an estimated 2PL Rasch model (figure 6.2). IRFs showed the probability of an affirmative response for each item as a function of the latent trait, household food access. The higher was food access (on the horizontal axis), the higher should be the probability of consuming the food group (on the vertical axis). The numbers on the different curves correspond to the items (food groups) provided in table 6.2. All food groups appeared to behave as expected: all curves show an upward slope.

If two items had similar discriminatory power, α, but differed with respect to their difficulty, β, the curve of the most difficult item (higher β) would be plotted towards the right-hand side of the figure. For instance, food group 3 (vegetables) and 7 (fish) had similar discriminatory power (α equaled 1.006 and 0.858 respectively), but vegetables ($\beta = 0.07$) was a considerably easier item than fish ($\beta = 3.55$). Hence, the IRFs of fish and vegetables were almost parallel, but the curve of vegetables was located to the left of the curve of fish.

The α's determine the slope of the IRFs: items with high discriminatory power have steeper slopes. For instance, food group 5 (meat) and food group 8 (legumes) had similar β's, but the slope of the IRF of meat was steeper than the slope of the IRF of legumes, because the latter had a smaller α. In other words, the food group meat had more power in differentiating between households with high and low food access.

The IRF of food group 6 (eggs) was rather flat, which indicated the probability of consuming eggs might be independent of the latent trait. A test confirmed that the discriminatory power of food group 6 was not significantly different from zero ($p = 0.22$), so the item was removed from the refined scale. Eggs might not explain

Figure 6.2: Item Response Function (Colombia)



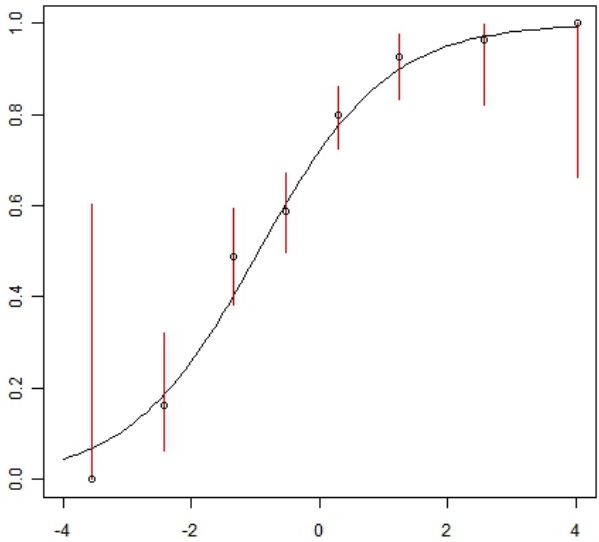
household food access because they are an important component of the daily diet in Colombia, independent of the socio-economic status of the household (Dufour et al., 1997). Most households might eat eggs frequently but not daily. In our sample, eggs were consumed by two-thirds of the interviewed households (table 6.2). All seven remaining food groups had a positive and significant relationship with the latent trait, and were therefore included in the refined scale.

Item fit was verified by visual inspection of the Item Characteristic Curves (ICCs) for each of the seven remaining items. ICCs are similar to IRFs and show the probability of consuming the food group (vertical axis) as a function of the household's food access (horizontal axis). ICCs also show the predicted probability of an affirmative response with its 95% confidence interval represented by vertical lines and the actual observed probability of an affirmative response represented by a dot. Item fit is high when predicted probabilities are close to expected probabilities. For example, for the food group meat (figure 6.3), predicted probabilities corresponded well to actual observations. Results for other food groups were similar.

Results of the robustness check supported the model. Although removing the 6% of observations with low person-fit ($p < 0.02$) did affect the size of the coefficients, it did not affect their difficulty rankings vis-à-vis one another. Local independence held. The nonparametric RM model test showed inter-item correlations between two out of 21 item-pairs, or roughly 10%. Based on the null hypothesis of independence this is no cause for serious concern. Further testing to find the source of

dependencies based on principal component analysis resulted in a maximum eigenvalue of < 1.3, with remaining eigenvalues slowly decreasing in size. Eigenvalues below 1.5 are generally considered as insignificant, confirming local independence (Kahler and Strong, 2006).

Figure 6.3: ICC of food group 5 (meat)



6.5.2 Ecuador: Differential Item Functioning

The amazon basin where the Ecuadorian data was collected had two ethnic groups with distinct dietary patterns. Originally the region was inhabited by the indigenous tribe of the Kichwa, but since the oil boom of the 1970s large groups of mestizo migrants have settled in the region and currently make up almost half the population (Lobao and Brown, 1998; Witt et al., 1999). A glance at the summary statistics for food groups consumption shows marked differences in diet between these groups (table 6.3). Milk and dairy products were, for instance, consumed by only 7% of Kichwa households, while 27% of migrant households reported having consumed this food group in the previous day. This suggested that the pooling the data from Ecuador might cause validity problems.

A formal test confirmed the occurrence of Differential Item Functioning between the ethnic groups ($p < 0.001$), implying that a single index for the Ecuadorian case did not meet condition 3 of cultural robustness. When the items showing the strongest DIF were removed one by one until they no longer showed any DIF ($p = 0.352$), only five food groups were left in the final model: 1, 3, 8, 9 and 11. Such a small number of groups is not very meaningful, as the resulting indicator can take only five values and is probably relatively insensitive to changes in food

access. By not pooling the data, valuable within-group information on specific diets was preserved. Hence, the subsequent analysis was performed separately for each of the two cultural groups[1].

Table 6.3: Food group consumption by Ecuadorian households across different ethnic groups

	Food group	% of Kichwa HHs ($n = 209$)	% of migrant HHs ($n = 297$)
1	Cereals	80	95
2	Roots and tubers	87	81
3	Vegetables	15	37
4	Fruits	26	40
5	Meat	52	66
6	Eggs	46	50
7	Fish	49	29
8	Legumes	18	56
9	Milk/diary	7	27
10	Oils/fat	40	38
11	Sugar/honey	52	77
12	Other	54	86

6.5.3 Kichwa households

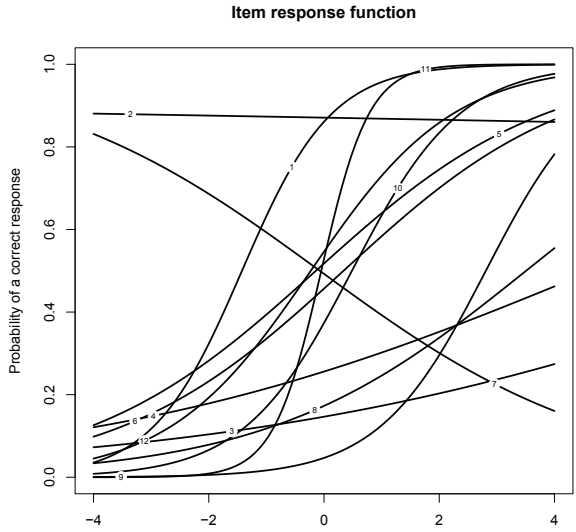
None of the food groups was consumed by so few or so many households to require removal from the indicator. The least frequently consumed food group, milk, was consumed by 15 households (table 6.3). Item Response Functions for all food groups based on the 2PL model are shown in figure 6.4. In order for a food group to usefully contribute to the additive HDDS, the likelihood of its consumption needed to increase with an increase in the latent trait, reflected by a positive and significant slope. Food groups 2 (roots and tubers) and 7 (fish) both appeared to violate this condition.

The IRF of food group 2, roots and tubers, was a flat line. The item had low discriminatory power ($\alpha = 0.04$) and extremely low item difficulty ($\beta = -48.01$). The food group was consumed by 87% of Kichwa households, but their consumption was practically independent of their food access situation, meaning the group added no explanatory power to the overall indicator. It is likely that this food group was consumed by all households on a regular but not daily basis and therefore its consumption had no power in explaining household food access.

The negative slope on food group 7 (fish) indicated the likelihood of consuming

[1]These samples could be considered on the small side for 2PL Rasch analysis, which might lead to biased estimates (De Ayala, 2013). However, they are not problematically small for the purpose of this paper, since we do not rely on precise estimates of α's and β's to draw our conclusions. Furthermore, model tests show only small differences with 1PL models, for which a sample size of 100 is already considered informative

Figure 6.4: Item Response Function (Ecuador, Kichwa HH)



fish decreased with increasing food access. The predicted likelihood of consuming fish decreased from 80% for households with little dietary diversity to less than 20% for households with a highly diversified diet. Previous research found fish to be an important part of the diet in Kichwa communities and consequently its consumption was common, although more so in rural communities than in towns (Webb et al., 2004). No sources were found mentioning an inverse relationship between income and fish consumption, although a possible explanation for the observed effect could be a development project of the provincial government of Napo which donated fish ponds to indigenous households in the region. Such a project was mentioned by respondents in a second survey round conducted in summer 2013[2]. If only food insecure households were eligible for this programme, it would explain the observed inverse relationship of fish consumption with overall dietary diversity. Another potential explanation is that fish is a Giffen good. A good is a Giffen good if it has a positive price elasticity, that is, if its consumption increases when its price increases, violating the law of demand. This occurs if the substitution effect is offset by an income effect. Although it has been shown empirically that some staple crops such as wheat and rice sometimes behave as Giffen goods in developing countries (Jensen and Miller, 2008), it is unlikely that fish is a Giffen good. Fish is not a staple crop and households are unlikely to consume less fish if their income increases. Hence, it makes more sense to assume that the HDDS does not measure food security adequately than to assume that

[2]We were not able to identify the project. Respondents were most likely referring to the "Piscicultura Sostenible para la Amazonía" project executed by the Centro Lianas (www.centrolianas.org).

fish is a Giffen good to explain the negative correlation between the latent trait and fish consumption.

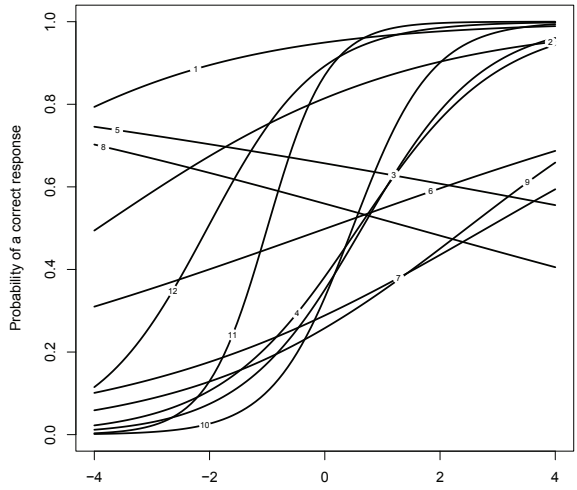
Removing observations with low person-fit to verify robustness resulted in dropping 6% of observations. Differences between the full and reduced sample were negligible. Ranking was unaffected, and coefficient size hardly changed. Local independence tests based on inter-item correlation showed six out of 45 tested pairs, or 13%, showed significant correlation ($p < 0.05$). Further analysis of the source of the variation indicated sampling variation rather than structural variation. The highest eigenvalue was 1.58; other eigenvalues were only slightly lower.

6.5.4 Migrant households

No food groups required removal from the refined indicator for migrant households because of too high or too low consumption frequency (table 6.3). The most frequently consumed food group was cereals, which was consumed by 95% of the population. Only 15 households did not report its consumption. Because this exceeded the critical threshold of ten observations per dichotomous choice alternative (Linacre, 2002), the food group was not removed.

Food groups 5 (meat) and 8 (legumes) appeared to have negative slopes (figure 6.5), warranting their exclusion. Inspection of the coefficients of the 2PL model indeed showed that the slope of food groups 5 and 8 was negative ($\alpha = -0.11$ and $\alpha = -0.16$, respectively), but testing revealed that these slope were not significantly different from zero at the 5% confidence level. These food groups were removed from the refined indicator.

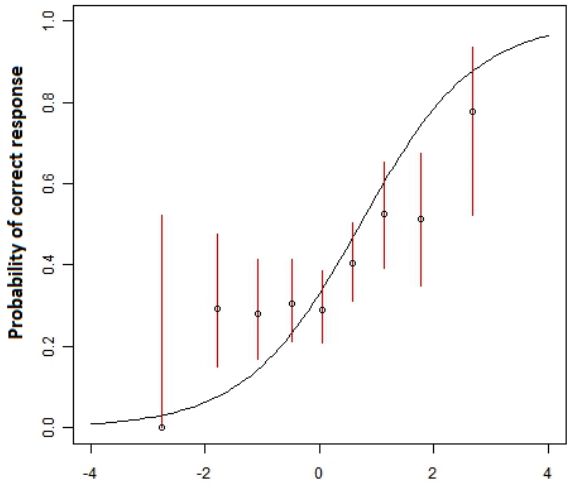
Figure 6.5: Item Response Function (Ecuador, migrant HH)



Upon inspection of the ICC curves for migrant households, food group 7 (fish)

was found to have low item fit. Many more households than predicted consumed fish at the lower tail of the distribution, meaning its consumption did not increase considerably with higher levels of food access (figure 6.6). The Chi-squared test for item fit confirmed this conjecture; the null of good item fit was rejected ($p = 0.013$). Therefore, food group 7 was removed from the refined scale. Re-testing showed the remaining items to have good fit.

Figure 6.6: ICC of food group 7 (fish) for migrant HHs



The resulting scale was checked for robustness by removing the 5% of observations with low person-fit ($p < 0.02$). This removal had a minimal effect on coefficient sizes and did not affect their ranking. Local independence did offer some cause for concern. Out of 36 item pairs tested for inter-item correlations, six were found to be significant (17%). Further testing of the source of the variation gave a maximum eigenvalue of 1.52. In other words, the observed local dependence was likely caused by sampling variation.

6.6 Discussion

In this paper the household dietary diversity score (HDDS) developed by the Food and Nutrition Technical Assistance (FANTA) project was analysed using Rasch models. In particular, it was verified whether the indicator met the criteria required for interval scale measurement. Meeting these criteria would imply the HDDS can be used as a household level indicator of food access. Such validity at household level is relevant for both development and research organizations, as it would allow attribution of project impact to specific outcomes. Rasch models allow differentiation between the discriminatory power and difficulty of items, revealing

the relative importance of individual food groups in differentiating between levels of food access. In our data, this importance differed markedly between countries and between groups within Ecuador. Therefore, in its current form the HDDS did not meet the criteria.

For most applications of Rasch modelling, the assumption of hierarchical ordering of the items is essential. In our application, this would imply that the food groups can be ranked ex-ante based on their difficulty. It also implies that a household that consumes the most difficult item should have consumed the other, easier, items. For dietary diversity, such a strict hierarchical ordering is difficult to establish, as it depends on locally prevailing market conditions (availability, price) and cultural preferences. Yet, in poor and food insecure regions, it is reasonable to assume that more food secure households consume more and less accessible food groups. It is hard to believe that households switch completely from one diet to another or no longer consume certain food groups as they grow richer. In this sense, a hierarchical ordering of food groups is likely. On the other hand, changing dietary patterns may not completely be captured by the HDDS. For instance, food and food insecure households may both eat meat, but more food secure households may switch from chicken to beef. The HDDS is insensitive to such changes. A second argument supporting the applicability of Rasch analysis is the main conclusion - that some food groups are not correlated with the overall HDDS score or with the consumption of other food groups - could be reproduced using 'simple' descriptive and comparative statistics. Therefore, even if the key assumption of hierarchical ordering was not met, in the context of this paper the consequences appear to be limited. Importantly, in this paper Rasch analysis was not used to calibrate the indicator, which would require precise estimates of item parameters and hence would be more sensitive to the consequences of invalidating the assumption.

Assuming hierarchical ordering of the food groups, Rasch models can be used to test two conditions which a valid indicator of food access should meet: (i) cross-cultural validity and (ii) an increasing probability of an affirmative answer with increasing food access. The pooled data, combining data from Colombia and Ecuador, did not meet the condition of cross-cultural validity. More worryingly, even within the sample of Ecuadorian households, significant differences in dietary patterns between Kichwa and migrant households were detected. Data had to be split into three different groups Colombian, Kichwa, and migrant households which were analysed separately. For each of these groups, items (food groups) not meeting the second condition were removed from the scale until a 'refined' HDDS was found that did meet all conditions. An overview of the three resulting scales is given in table 6.4. It contains for each food group either the reason it was omitted from the scale or its difficulty ranking in the overall scale. The most difficult food groups were most likely to be eaten by households with the highest food access score.

Table 6.4: Reason for removal and final difficulty ranking of food groups for refined HDDSs

	Food group	Colombia ($n = 509$)	Kichwa group (n=209)	Migrant group ($n = 297$)
1	Cereals	*targeting*	1	1
2	Roots and tubers	*targeting*	$\alpha = 0$	2
3	Vegetables	5	10	8
4	Fruits	4	9	7
5	Meat	3	3	$\alpha = 0$
6	Eggs	$\alpha = 0$	5	5
7	Fish	7	$\alpha < 0$	*low item fit*
8	Legumes	2	8	$\alpha = 0$
9	Milk/diary	6	7	9
10	Oils/fat	1	6	6
11	Sugar/honey	*targeting*	4	4
12	Other	*targeting*	2	3

Numbers in columns indicate food group difficulty ranking (1 indicates the easiest food group); targeting indicates the food group was excluded because 99% of households consumed it; $\alpha = 0$ indicates the food group does not explain food access; $\alpha < 0$ indicates the food group has a negative relationship with food access, such that its consumption is associated with lower food access; low item fit indicates a significant difference between expected and predicted responses.

There are large differences between the three refined indicators in the number of food groups they contain and in the difficulty ranking of these food groups. In the Colombian data, seven food groups made up the refined indicator: vegetables, fruits, meat, fish, legumes, diary, and oils. These results correspond well with the literature as the refined index mainly contains foods with high nutritional values such as fruits, vegetables, and animal source products. The results for the Ecuadorian subgroups were less convincing. For Kichwa households, the food groups roots and tubers, and fish were excluded from the final index and for migrant households the groups meat, fish, and legumes did not meet the conditions. Especially the non-inclusion of meat and fish in the overall index for both groups is cause for concern, as animal source foods are of crucial importance for macro and micro nutrient intake in developing countries (Murphy and Allen, 2003). Moreover, as there appears to be a direct link between consumption of animal source foods and dietary diversity (Brown et al. (2002), as cited in Ruel (2003)), the exclusion calls into question what the HDDS really measures.

There were substantial differences in the importance of each food group in the overall index between countries and even within a country. This holds even though two culturally similar neighbouring countries were studied. In its current form, the HDDS has no cross-cultural validity, a problem previously mentioned but not tested by Ruel (2003). DIF-analysis showed that the indicator is not even necessarily valid within a country, as in Ecuador dietary patterns differed between

groups with a different cultural background. This lack of cross-cultural validity is problematic as it prevents direct interpretation of the value of the overall indicator. Before interpreting this value, it is essential to have a thorough understanding of local dietary patterns, even when a survey or project concerns only a small area within a single country. Clearly, requiring extensive knowledge before being able to interpret a simple, easy-to-use indicator limits its usefulness for deployment in the rapid assessments required by development projects.

A potential cause of the limited accuracy of the HDDS at household level might be its focus on only the foods consumed in the last 24 hours before the survey (Swindale and Ohri-Vachaspati, 2005). In that case, a straightforward way to overcome this inaccuracy is to increase the recall period. In a study using a 15 day recall period for dietary diversity, Drewnowski et al. (1997) noted diversity increased steeply over the first three days of recall, after which further increases became small. In other words, 24h recall might significantly underestimate true diversity when measuring dietary diversity at an individual or household level. Specifically, it might reduce the inaccuracy stemming from food groups that are eaten frequently, but not daily.

Other factors that might increase the construct validity of the indicator are re-defining the included food groups, adding weights, consumption frequency, and establishing minimum portion sizes. Food groups could be re-defined based on nutritional values, as is already being suggested specifically for iron deficiency (FAO, 2012b) and is common in studies in the field of nutrition (Ruel, 2003). Weights could be added to account for the distinct nutritional value of food groups, as is already done by the Food Consumption Score used by the World Food Programme (WFP, 2008). The frequency of consumption might also be considered, which is particularly important in the presence of habit formation. Then, households might prefer those foods consumed as a child even when alternative food baskets become affordable (Atkin, 2013). Finally, minimum portion sizes should be considered. Ruel (2003) gives an example from Ghana, where fish consumption appeared high until it was found out fish meal was added in small amounts to porridge, obviously limiting its nutritional contribution. Different indicators take one or several of these factors into account, but knowledge of the contribution of each factor to the overall accuracy of the indicator is lacking. Further research is needed to specify and quantify the trade-offs involved.

6.7 Conclusion

The HDDS was developed as an easy-to-use and quick-to-implement survey-based assessment tool to allow measuring the impact on household food access of programs with improvements in food security as their core objective. Our results show the indicator should be cautiously interpreted. The HDDS does not allow

comparing food access between different countries. Moreover, even in a small region within a single country, the indicator should not be used without sufficient knowledge of local dietary patterns. When dietary patterns differ between groups within a region, scores should not be aggregated for the region as a whole. Even within these relatively homogenous groups, there is a limited fit between included food groups and the underlying latent trait, such that the components of the indicator do not form a reliable way of measuring the variable of interest: food access.

Several problems were encountered with regard to the food groups making up the indicator. The gravest problem encountered was the inclusion of a food group with a negative relationship with the latent trait, implying that households were more likely to consume the food group when they had lower food access. Such items should never be included in an additive scale. In each of the three groups studied, there was at least one item which had no relationship with the latent trait, reducing the indicator's accuracy. Such items cause incorrect classification of households into food security states. Both problems might be avoided by redefining the included food groups, adding weights, consumption frequency, and establishing minimum portion sizes. Until these issues are satisfactorily resolved, the HDDS should not be used as an indicator of the food access status of individual households.

Appendix

6.A Output tables of 2PL models

Table 6.A.1: 2PL model Colombia including eggs

	α	β	se α	se β	t-values α
Vegetables	1.005	0.069	0.250	0.108	4.023
Fruits	1.128	−0.007	0.287	0.099	3.937
Meat	0.538	−1.368	0.185	0.455	2.910
Eggs	0.110	−5.901	0.146	7.829	0.754
Fish	0.858	3.552	0.306	1.045	2.806
Legumes	0.316	−1.651	0.148	0.794	2.139
Milk/diary	1.192	1.292	0.331	0.257	3.601
Oils/fat	0.564	−3.480	0.225	1.259	2.509

Table 6.A.2: 2PL model Kichwa households (Ecudaor) including roots/tubers and fish

	α	β	se α	se β	t-values α
Cereals	1.281	−1.424	0.433	0.343	2.955
Roots and tubers	-0.022	84.830	0.293	1102.814	−0.077
Vegetables	0.197	8.949	0.286	12.821	0.689
Fruits	0.229	4.662	0.215	4.325	1.067
Meat	0.502	−0.141	0.215	0.297	2.337
Eggs	0.511	0.339	0.213	0.314	2.395
Fish	-0.406	−0.073	0.204	0.356	−1.995
Legumes	0.447	3.507	0.262	1.947	1.703
Milk/diary	1.073	2.808	0.481	0.956	2.229
Oils/fat	1.066	0.485	0.297	0.186	3.594
Sugar/honey	2.410	−0.048	1.062	0.107	2.268
Other	0.809	−0.229	0.253	0.204	3.197

Table 6.A.3: 2PL model migrant households (Ecudaor) including meat and legumes

	α	β	se α	se β	t-values α
Cereals	0.397	−7.390	0.348	6.211	1.140
Roots and tubers	0.376	−3.938	0.221	2.218	1.704
Vegetables	0.953	0.645	0.277	0.201	3.444
Fruits	0.825	0.58	0.243	0.211	3.392
Meat	-0.106	6.107	0.176	10.097	−0.605
Eggs	0.198	0.034	0.166	0.591	1.193
Fish	0.321	2.812	0.192	1.658	1.670
Legumes	-0.155	1.534	0.164	1.774	−0.945
Milk/diary	0.429	2.465	0.201	1.112	2.137
Oils/fat	1.453	0.481	0.467	0.142	3.111
Sugar/honey	1.903	−0.999	0.629	0.183	3.026
Other	1.040	−2.037	0.321	0.484	3.243

CHAPTER 7

The inverse productivity-size relationship: can it be explained by rounding of self-reported production

Abstract: The inverse productivity-size relationship is one of the oldest puzzles in agricultural economics. Many hypotheses have already been tested to explain the negative association between plot size and land productivity, but none of them are fully satisfactory. In this paper, we propose a new explanation: reporting production as a 'round' number. We show that households tend to report production as multiples of 5 or 10kg. In combination with small average plot sizes this can substantially inflate estimated yields. Small rounding errors in production numbers do indeed cause large overestimation of plot-specific yields particularly when production figures are multiplied by a small plot size. The overestimation of yields will be greater on small plots than on large ones, causing a spurious inverse productivity plot-size relationship. We test this hypothesis with data from an agricultural survey in Burundi. Our results show that rounding of production numbers does bias yields upwards and reinforces the inverse productivity-size relationship, although only to a limited extent. Besides offering a new explanation for the inverse productivity-size relationship, this paper illustrates how 'rounding' errors in self-reported numbers can affect statistical inference.

Paper in preparation:
Desiere, S., D'Haese, M. The inverse productivity-size relationship: can it be explained by rounding of self-reported production.

7.1 Introduction

The inverse productivity farm-size relationship (IR) in developing countries is one of most intriguing issues in agricultural economics. As first noted by Chayanov (1926/1986) in Russia and rediscovered by Sen (1962) in India, it states that production per hectare decreases with increasing farm size. This finding has an important and oft-emphasized policy implication, namely that the redistribution of land from large-scale to small-scale farmers not only improves equity, but also efficiency. Few economists would, however, recommend land reform to promote efficiency gains (Collier and Dercon, 2014) and numerous explanations for the IR have been offered in the literature that question the finding that small farms are more efficient than large farms.

The most commonly accepted explanation is related to missing markets. If labor and land markets are non-existent or imperfect, households with little land will not find sufficient wage work and will therefore apply labor (and other inputs) more intensively to their own fields than is efficient because of the low opportunity cost of their time (Carter and Wiebe, 1990). Similarly, Barrett (1996) argues that the absence of insurance markets pushes small farmers who are net buyers of food to oversupply labor to insure against high prices, while large farmers (who are net sellers) under supply labor to avoid losses if prices are low. Feder (1985) argues that larger farms need to hire wage workers, who are likely to shirk more than workers who cultivate their own land. He argues that higher supervision costs on larger farms explain the IR. These three different explanations all focus on household behavior. Hence, if missing markets explain the IR, there should be no difference in productivity between fields cultivated by a same household, but only between fields cultivated by different households.However, several studies based on plot-level data showed that, even within a single household, small plots are more productive than large plots (Assunção and Braido, 2007; Barrett, 2010). Hence, there exists not only an inverse productivity *farm-size* relationship, but also an inverse productivity *plot-size* relationship. The existence of the inverse productivity *plot-size* relationship rules out market imperfections or higher supervision costs at larger farms as main reason behind the inverse productivity *farm-size* relationship (Assunção and Braido, 2007)

A second strand of the literature relates the existence of the inverse productivity *plot-size* and *farm-size* relationship to unobserved differences in land quality (Bhalla and Roy, 1988; Benjamin, 1995). If smaller plots have generally better soil characteristics than larger plots, then omitting soil quality as an explanatory variable would bias the estimated coefficients and generate an inverse relation. This explanation was, however, convincingly rejected by Barrett (2010) who had access to excellent soil quality data and showed that differences in soil characteristics contributed only marginally to explaining the IR in Madagascar. These authors suggested that measurement error is one of the few remaining potential

explanations for the IR. Measurement error in self-reported land size is indeed a third explanation that has been suggested by Lamb (2003) to explain the inverse productivity *farm-size* relationship. Whether measurement error in plot size strengthens or weakens the inverse productivity size-relationship is, however, unpredictable. The reason is that measurement error in plot size introduces measurement error in both the explanatory variable (land) and the dependent variable (yields, or output per hectare). Measurement error in the explanatory variable reduces the strength of the IR, but measurement error in the dependent variable (yields) may strengthen it. This ambiguous effect has also empirically been observed. Recent studies have shown that measurement error in plot size weakened the IR in Uganda (Carletto et al., 2013b), but strengthened it in several other African countries (Carletto et al., 2015a)[1].

In this paper, we propose a new explanation for the *plot-size* inverse productivity relationship: one that is also due to measurement error. We explore whether measurement error in self-reported production contributes to the negative association between yields and plot size. We argue that the rounding of self-reported production to 'easy numbers' such as 5 or 10 kg can cause this effect. Rounding of self-reported production is an often observed phenomenon in agricultural surveys (Beegle et al., 2012). Plot-specific yields are then calculated by dividing rounded production numbers by plot size. Consequently, a small over or underestimation in production at plot level will lead to a large bias in estimated yields if the cultivated area is small[2]. For instance, consider a farmer with a plot of 80 m^2 of beans and a harvest of 8 kg of beans. The 'true' yield here is 1000 kg/ha. But, if the farmer reports a production of 10 kg, his yield would be estimated at 1250 kg/ha, an overestimate of 25%. The combination of small rounding errors in self-reported production and a small average plot size could thus lead to a large bias in estimated yields. Given that many households in developing countries have quite a number of small plots instead of a few large ones (Jayne et al., 2003), this can contribute to significant over or underestimations of yields at plot level. This theoretical explanation can be empirically tested, which is what we do in this paper, using data from an agricultural survey conducted in Burundi.

This paper examines the *plot-size* inverse productivity relationship, i.e. whether

[1] Barrett (2010) state that measurement error in plot size strengthens the IR *"For example, if survey respondents with smaller plots and farms systematically over-report the size of their farm or plots (perhaps because land is a measure of prestige), one is likely to find a spurious inverse relationship between size and productivity"* This statement is contradicted by Carletto et al. (2013b) who state that: *"For the IR to be partially or fully explained by errors in land measurement, smaller farmers would have to systematically under-report land area with respect to larger farmers, thus resulting in artificially inflated yields at the bottom part of the distribution."* This statement was refined in a more recent article about measurement error in self-reported land size and concluded that measurement error in plot size has an ambiguous effect on the strength of the IR (Carletto et al., 2015a).

[2] This explanation hinges on the assumption that farmer estimates are used to estimate yields. If crop cuts are used instead, rounding may be less of a concern. We will come back to this point in the discussion.

yields are higher on small plots than large plots within a household. Consequently, we did not examine the *farm-size* inverse productivity relationship, i.e. whether yields are greater on small farms than large farms. While the existence of *farm-size* inverse productivity is perhaps a more policy-relevant research question than the *plot-size* inverse productivity relationship (Collier and Dercon, 2014), there are at least three important reasons for studying the effect of rounding on the *plot-size* inverse productivity relationship. First, many academic studies use plot-level data and study the *plot-size* inverse productivity relation instead of the *farm-size* inverse productivity relationship (among others: (Kimhi, 2006; Assunção and Braido, 2007; Barrett, 2010)). Using plot-level data has the important advantage that plot-specific soil characteristics and household fixed effects can be included in the econometric specification. However, some authors fail to discuss the important difference between the *plot-size* and the *farm-size* inverse productivity relationship. Second, several studies reject the hypothesis of imperfect markets as the main explanation for the *farm-size* inverse productivity relationship, because this hypothesis can only explain differences in productivity between households, but not between plots within the same household (Assunção and Braido, 2007). This argument requires that the *plot-size* inverse productivity relationship is correctly estimated. Third, it is possible that rounding also biases the *farm-size* inverse productivity relationship. This bias is likely to be more limited than the *plot-size* inverse productivity relationship because the bias in observed, plot-specific yields is mainly caused by rounding of self-reported production on small plots. If yields are calculated at the household level, aggregated self-reported production will be divided by aggregated cultivated land area. Consequently, small rounding errors will not be considerably inflated by small plot sizes. However, in some studies, households' landholdings might be sufficiently small for the results to still be partially explained by rounding errors.

Besides offering a new explanation for the *plot-size* inverse productivity-size relationship, this paper contributes to the small literature about systematic measurement error in household surveys (Chesher and Schluter, 2002). In particular, it studies 'rounding' or 'heaping' error, which is known to be a nuisance in survey data, though only rarely studied (Wang and Heitjan, 2008; Carletto et al., 2013a). Notable exceptions are Gibson's studies about systematic measurement error in self-reported household consumption (Gibson and Kim, 2007; Gibson et al., 2013). Rounding in self-reported quantities has also been observed in surveys about smoking behaviour (Wang and Heitjan, 2008) and in surveys recording events retrospectively (Bar and Lillard, 2012). Furthermore, there exists a literature showing that prices tend to cluster on round numbers in the financial and exchange markets (Harris, 1991). In contrast to random measurement error, systematic measurement error is correlated with the 'true' value of the mismeasured variable or with other plot or household characteristics. This is the case in this paper because measurement error in yields is negatively correlated with plot size. Because of such correlations, systematic measurement error can cause spurious relations in

the data.

The best and simplest approach to deal with measurement error is comparing the mismeasured variables with its gold standard, that is, its 'true' value. In practice, however, a gold standard only rarely exists or is too cumbersome to be included in large-scale household surveys. A second best solution is re-measuring the mismeasured variable using a different method and then using this second measurement as an instrument in a regression analysis (Carroll et al., 2012). Yet, measuring the same variable twice in a household survey is also exceptional. Hence, more creative solutions are required to study the effect of systematic measurement error on statistical interference, which circumvent the identification problem. This was also necessary in this paper since 'true' yields at plot levels were also unobserved. To this end, we developed a simple simulation model. Next, we identified rounding in our datasets by assuming that production reported as a multiple of 5 kg or 10 kg was potentially rounded.

The remainder of the paper is structured as follows. First, we develop a simple model to clarify how rounding error can cause the IR, test it with a simulation model and derive an econometric specification that tests for bias due to rounding. This approach is applied to a unique cross-sectional dataset from Burundi, where agriculture is characterized by many tiny plots of land. The main implications and limitations of our approach are discussed in the conclusions.

7.2 Empirical framework

7.2.1 A simple model

We develop a simple model to show how rounding errors can systematically bias the estimation of average yields at plot level. Moreover, we show that this bias is more pronounced for small plots than larger ones, which strengthens the inverse productivity-size relationship.

Assume that farmers and enumerators have a preference for round numbers. Hence, harvests are likely to be reported as a multiple of 5 kg, 10 kg or even 25 kg or 50 kg. Self-reported production at plot i, y_i, is equal to 'true production', y_i^*, plus a rounding term, α_i, which is also plot-specific:

$$y_i = y_i^* + \alpha_i \tag{7.1}$$

A statistician only observes self-reported production and the size of each plot (A_i) and uses this information to estimate plot-specific yields:

$$yield_i = \frac{y_i}{A_i} = yield_i^* + \frac{\alpha_i}{A_i} \tag{7.2}$$

If we assume that α_i is on average positive, then yields at plot level would systematically be biased upwards because of rounding errors. Even more importantly, this bias is more pronounced for small plots of land, because small rounding errors are heavily inflated when production is divided by a small plot size. Consequently, even if the IR does not hold, measurement error causes a negative correlation between yields and plot size. This can easily be illustrated as follows. Assume that the inverse productivity-size relationship does not hold. This means that there is no negative correlation between plot size and 'true' yields, i.e. $corr(yield_i^*, A) = 0$, but the correlation between self-reported yield and plot size will still be negative:

$$corr(yield^*, A) = 0 \Rightarrow corr(yield, A) = corr(yield^* + \frac{\alpha}{A}) = corr(\frac{\alpha}{A}, A) \leq 0, \ if \ \alpha > 0$$

Hence, the IR would not be rejected, although the relation is solely caused by rounding errors and thus a statistical aberration. Similarly, even if the IR holds, the strength of this relationship would be overestimated because of rounding errors.

The bias in the IR introduced by rounding depends essentially on the psychological process behind rounding, which is not yet well understood. Two conditions about this rounding process have to be satisfied in order for it to strengthen the IR. First, that households tend to round their production upwards rather than downwards. If they were just as likely to round downwards as upwards this would not strengthen the IR. In this case, rounding would only increase the variance of the distribution of production because observations would be clustered around round numbers (Schneeweiss et al., 2010). However, if it is more likely that farmers round upwards than downwards, then yields from smaller plots would be overestimated relative to those from larger plots, which would artificially generate (or accentuate) the inverse relationship. The condition of 'upwards' rounding is key to our model and can easily be criticized as 'wishful thinking'. It is indeed hard to explain why farmers would be more likely to round upwards than downwards. A potential explanation is that reporting higher production numbers are perhaps more prestigious than lower numbers. It is not possible to directly test whether the assumption of upwards rounding holds in our data. We could only verify that yields were higher if production was reported as a round numbers. This provides suggestive evidence that farmers indeed round production numbers upwards. This test is not completely satisfactory as it does not solve the identification problem. We come back to this point in the conclusion. The second condition is that the rounding (upwards) of production should occur on both small and large plots. If rounding occurs rarely on small plots, the yields from these small plots will not systematically be overestimated and the IR will not be influenced by rounding. This is something that can easily be checked in the data. Moreover, this condition is also important to avoid reversed causality when estimating the IR. This point will be discussed when deriving the econometric specification. These two assumptions - that households round production upwards and rounding also occurs on smaller plots - are central to our model. Descriptive statistics (see section 7.4.1) suggest that both of these assumptions do hold.

7.2.2 A simulation model

To verify if plausible assumptions on rounding behaviour can generate the inverse productive size relationship, we developed a simulation model. This simulation model is partly calibrated on our data from Burundi. This data as well as descriptive statistics are discussed in depth in the next sections. The simulation model incorporates the distribution of plot size in Burundi. This is important because we want to test if the combination of the small, average plot size in Burundi and upwards rounding of self-reported production generates the inverse relationship.

The simulation model consists of five steps. In a first step, plot-specific 'true' yields, y_i^*, are drawn from a normal distribution with a mean of 751 kg/ha and a standard deviation of 100[3]. In the second step, 'true' food production at every plot is calculated by multiplying the 'true' yields with plot size. The third step is the critical one as it simulates the rounding behavior of the farmers. In this step, it is assumed that farmers prefer to report their production as a multiple of 5 kg, 10 kg or even 25 kg. The rounding behavior is calibrated upon the dataset from Burundi which showed that around 55% of the production numbers were reported as multiples of 5 kg (see descriptive statistics in section 7.4.1). If rounding had not occurred, one would expect that multiples of 5kg would only occur on 20% of the plots (i.e. production numbers with zero or five as the last digit). Given that we observed that 55% of production numbers were reported as multiples of 5 kg, we assumed that 35% of the production numbers were rounded.

We simulated three different scenarios. In the first scenario, we assumed that farmers round production upwards to the nearest multiple of 5 kg on 35% of the plots. In the second scenario, production on 25% of the plots was rounded upwards to the nearest multiple of 5 kg, while production was rounded upwards on 10% of the plots to the nearest multiple of 10 kg. In the third scenario, households were assumed to round to multiples of 25 kg on 5% of the plots and to multiples of 5 kg and 10 kg on 20% and 10% of the plots, respectively. The second scenario is our preferred one as it replicates best the empirical distribution of rounding in the datasets. Production numbers with '0' as the last digit (and thus potentially rounded to 5 kg or 10 kg) occurred twice as frequently in the dataset as production numbers with '5' as the last digits (and thus potentially rounded to 5 kg).

In the fourth step of the simulation model, the 'observed' yields were obtained by dividing rounded production by plot size. In the final step, we estimated the

[3] An average yields of 751 kg/ha was chosen because this is the average yields on larger plots in Burundi (see descriptive statistics), which are (according to our model) accurately estimated. We re-simulated the model with 626 kg/ha as the average yields, which corresponds to average yields on the largest plots, excluding rounded observations. Results are similar and are available upon request. What is, however, important is that average yields are not high. For instance, setting average yields at 1500 kg/ha would substantially reduce bias due to rounding errors. The choice of the standard deviation did not matter much. Results remained nearly identical when simulating the model with standard deviations of 50 and 150.

inverse productivity-size relationship. Traditionally, the inverse productivity-size relationship at plot level is estimated as follows:

$$log(yield_{ij}) = \alpha + \beta log(A_{ij}) + u_j + \epsilon_{ij} \tag{7.3}$$

Yields on plot i of household j are thus regressed on plot size A_{ij}. Household fixed effects, u_j, are included to account for differences between households related to, for instance, soil quality or the farmers farming ability. In the simulation model, it is not necessary to include household fixed effect because, by construction, there are no differences between households. Hence, equation 7.3 can be estimated with OLS. It is, however, well-known that fixed effects may strengthen bias due to measurement error (Baltagi, 2008). Lamb (2003) even argued that the inverse productivity-size relationship is induced by the combination of measurement error in plot size and including household fixed effects in the estimations. To test this, we estimated equation 7.3 with both OLS and fixed effect models.

In addition to rounding error, we also incorporated measurement error in plot size in the simulation model. In our dataset, GPS devices were used to measure plot size. GPS may, however, measure plot size imprecisely, particularly on small plots. This may strengthen the IR, particularly when including household fixed effects in the regressions. We considered two types of measurement errors in plot size. The first type assumes that measurement error is independent from plot size and is uniformly distributed in the range of -10% to 10%. In other words, 'true' and measured plot size differ maximally by 10%. The second type of measurement error in plot size assumes that maximum measurement error decreases with plot size towards an asymptote of 5%. Measurement error was calibrated in such a way that plots of 500 m^2 were measured with a relative error of maximum 16% (i.e. measured land size between 420 m^2 and 580 m^2), while plots of 1000 m^2 were measured with a relative error of maximum 10% (i.e. measured land size between 900 m^2 and 1000 m^2). There is indeed some evidence in the literature that relative errors in area measurement with GPS decrease with plot size (De Groote and Traoré, 2005; Carletto et al., 2015b) (see also chapter 3). Including measurement error in plot size in the simulation models allows examining whether the combination of rounding error and measurement error in plot size can generate the inverse productivity-size relationship.

Table 7.1 reports the results of the simulation model under the form of the mean value of β (see equation 7.3). The results confirm that rounding error can generate the inverse productivity-size relationship. The strength of the IR varies between 3% and 8%, depending on the degree of rounding error, the type of measurement error in land and the estimation strategy. This is lower than generally found in the literature (Larson et al., 2014). Consequently, rounding error can, at best, only partially explain the IR. The strength of the IR increases with the severity of the rounding error. In other words, assuming upwards rounding to the nearest multiples of 25 kg on 5% of the plots (scenario 3) has a more profound effect on the IR than only assuming rounding to the nearest multiple of 5 kg (scenario 1).

Uniform measurement error in plot size does not strengthen the IR (type 1), while the assumption of measurement error that decreases with plot size does strengthen the IR (type 2). The effect is more pronounced when including household fixed effects in the regressions. It is also interesting to note that non-uniform measurement error generates a statically significant IR even without assuming any rounding errors. It seems, however, that rounding of self-reported production can strengthen the IR at least as much as measurement error in plot size.

Table 7.1: Simulation results of rounding errors generating the IR

	Baseline No rounding	**Scenario 1** Rounding to 5kg: 35%	**Scenario 2** Rounding to 5kg: 25% Rounding to 10kg: 10%	**Scenario 3** Rounding to 5kg: 20% Rounding to 10kg: 10% Rounding to 25kg: 5%
Baseline: No measurement error in land				
OLS	0.001	-0.031	-0.035	-0.046
FE	0.001	-0.033	-0.038	-0.049
Type 1: Measurement error in land is uniformly distributed				
OLS	-0.002	-0.033	-0.038	-0.049
FE	-0.003	-0.037	-0.042	-0.053
Type 2: Measurement error in land decreases with plot size				
OLS	-0.024	-0.055	-0.060	-0.070
FE	-0.034	-0.067	-0.072	-0.083

Results based on 500 replications. n =18 754

7.2.3 Econometric specification

The results of the simulation model suggest one strategy to empirically observe the effect of rounding errors on the IR. By including dummy variables for rounding and interacting them with plot size in equation 7.3, we can both determine whether rounding biases yields and strengthens the IR. The problem here is that in our dataset we only observed reported – and not 'true' – production numbers. We do not know if production was rounded (nor by how much) or was indeed exactly a multiple of a round number. Therefore, we consider a production number rounded if it is a multiple of 5 kg. This set of observations still consist of two distinct subsets: production numbers that have '5' as last digit (which are thus potentially rounded to the nearest 5 kg) and observations with '0' as last digit (which are potentially rounded to the nearest 5 kg or nearest 10 kg). As the simulation model showed, rounding to the nearest multiple of 10 kg will have a more profound effect on the IR than rounding to the nearest multiple of 5 kg. Therefore, two proxies for rounding were included when estimating the IR: D_5 which equals one if the last digit of the self-reported production is 5 and D_{10} which equals 1 if the last digit is 0. For simplicity, we will refer to D_5 as 'rounded to a multiple of 5 kg' and to D_{10} as 'rounded to a multiple of 10 kg', although the latter group of observations also includes observations that have been rounded to a multiple of 5 kg. To estimate

the effect of rounding error on the IR, the following equation was estimated:

$$log(yield_{ij}) = \alpha+\gamma_1 D_5+\gamma_2 D_{10}+\beta log(A_{ij})+\gamma_3 D_5 log(A_{ij})+\gamma_4 D_{10} log(A_{ij})+u_j+\epsilon_{ij} \quad (7.4)$$

If yields are systematically overestimated because of rounding, the estimated coefficients γ_1 and γ_2 should be positive and significant. Moreover, we expect a more pronounced overestimation for rounding to multiples of 10 kg than multiples of 5 kg. Thus, $\gamma_2 > \gamma_1$. In addition, we expect that the overestimation of yields decreases with plot size, which strengthens the inverse productivity-size relationship. This implies that $0 > \gamma_3 > \gamma_4$.

Reversed causality is a concern in our econometric specification. It is reasonable to expect that rounding occurs more frequently for higher production numbers. For instance, households and enumerators are more likely to report a harvest of 103 kg as 105 kg than to report a harvest of 3 kg as 5 kg. Higher yields, which imply more production, are thus more likely to occur on plots with rounded production numbers. Rounding may thus signal higher yields. This mechanism introduces reversed causality because our model predicts that yields are higher because of rounding and does not predict that if yields are high, production is rounded. Although we do believe that rounding is more likely for higher production numbers, we argue that this will not cause reversed causality in our setting. In our dataset, fewer than 10% of the production numbers are larger than 100 kg. As a result, rounding because production numbers are large only occurs on 'large plots with high yields'. It can then easily be checked whether rounding also occurs for other reasons such as a preference for 'round' numbers by examining if rounding also occurred on small plots. For instance, on some tiny plots in our dataset total production can never exceed 50 kg, even if yields were above 5000 kg/ha. If production is nevertheless also rounded on these plots, this would provide some evidence that all production numbers are likely to be rounded.

7.3 Data

We used data from a national representative agricultural survey of 2560 households conducted in 2011/2012 by the Statistical Office of Burundi and the Ministry of Agriculture, which was financially supported by the Belgian Development Agency and the World Bank. This was the first-nationally representative agricultural survey in Burundi since the 1970s. Its main objective was to update agricultural statistics and to provide reliable production numbers at the provincial level.

A two-stage stratified design was adopted to randomly select the households. First, 20 sectors[4] were randomly selected with a probability proportional to population

[4]The sectors, known as Zone Dénombrement (ZD), are administrative units that cover a small geographical area, usually just including several villages. ZDs in predominately urban areas were

size within each of Burundis 16 rural provinces[5] . All the households within each sector were enumerated and 8 households were randomly selected to participate in the survey. Details of the sampling procedure can be found in a government report about the agricultural survey (République du Burundi, 2013a).

The survey contained 14 sections related to agriculture and the socio-economic status of the household. Detailed plot-level data on agricultural production and land characteristics were collected. All plots under cultivation were visited by the enumerator and the plots of land were measured with GPS. Most fields were intercropped. Because intercropping makes it notoriously difficult to determine the share of the plot devoted to a single crop, only total plot size was measured with GPS. As will be explained below, this explains why we could not calculate crop-specific yields and had to aggregate crop production at plot level. All households were at least visited once during each of the three agricultural seasons (September 2011 to August 2012). The survey focused on staple crops and excluded the cash crops coffee and tea. The household head reported the production of all food crops by plot. Most households reported production in kg and not in local units such as sacks. Hence, it was not necessary to convert local units into kilogram, which would complicate the definition of rounding. The only important exception were 'bananas', which were reported in 'large', 'medium' and 'small' bunches. Enumerators estimated the weight of the 'bunch' and used this estimate to convert the banana production in kilograms in the field. These estimated weights were plot-specific. Importantly, most weights were not equal to a 'round' number such as 5 kg or 10 kg. This ensures that the number of rounded observations was not artificially inflated due to a conversion from local units to kilograms.

Some observations were discarded from the original dataset. First, 6 households with more than 30 ha were discarded. Second, fields that were reported in a different location between seasons were removed from the dataset. Third, fields with the lowest and highest 5% of observed yields were excluded from the final dataset to ensure that results were not driven by data entry errors or outliers. After this screening 2543 household cultivating 18 754 fields remained in the dataset.

Plot characteristics included the amount of fertilizer applied to the plot, the cost of wage labor hired to work on specific plots, the location of the plot and whether any anti-erosion measures had been carried out. Measuring the input of family labor at plot level was considered too burdensome and prone to error. Hence, family labor was only reported at farm level.

Monocropped fields are a rarity in the agricultural system of Burundi, where intercropping is a very common phenomenon (Cochet, 2004). Hence, we aggregated total production per field using the calorific content of each crop as weight. To

excluded from the survey.

[5]The province of Bujumbura Mairie was excluded because it is dominated by the capital Bujumbura and was therefore considered an urban region.

make aggregated production more tangible and comparable, we divided aggregated production, expressed in terms of its energy content, by the calorific content of beans, which is one of the main staple crops. Our dependent variable in all the models is thus observed yields expressed in kilograms per hectare. For inter-cropped plots production was considered to be rounded if this was the case for at least one of the crops in the field.

7.4 Results

7.4.1 Descriptive statistics

Most households in Burundi have less than one hectare of land and cultivate 4 to 6 plots. This explains why average plot size in Burundi is exceptionally small. Figure 7.1 shows the distribution of plot size in Burundi. The distribution is highly skewed to the right as most plots are very small. The mean and median plot size is 0.075 ha and 0.042 ha respectively and only 1.5% of the plots are larger than 0.5 ha. Although this is a common finding in densely populated developing countries, it plays a vital role in our analyses. The fact that there are so many tiny plots implies that a small overestimation of production at plot level will lead to a large overestimation of yields per hectare.

Figure 7.1: Distribution of plot size

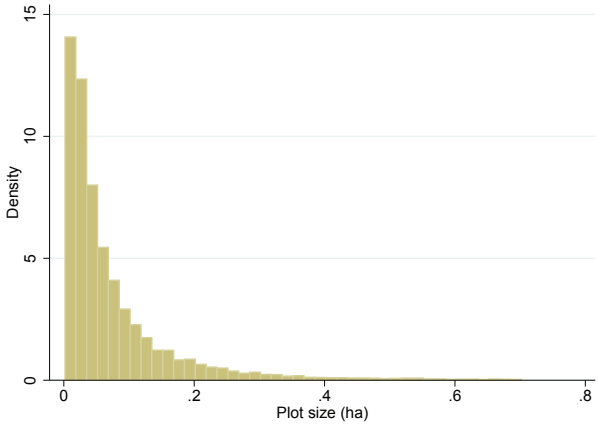


Table 7.2 provides plot characteristics for successive quartiles of plot size. The inverse productivity-size relationship seems to hold in our data: yields equal 1674 kg/ha for the smallest plots and decrease to 751 kg/ha for the largest plots. This may be because fertilizer and hired labor are much more intensively applied on small than large plots. For instance, application of fertilizer decreases from 29 kg/ha for the smallest plots to less than 3 kg/ha on the largest plots. Additionally,

there is clear evidence that small plots are more likely to be located in the fertile marshlands, supporting the hypothesis that difference in soil characteristics can contribute to explaining the IR.

Table 7.2: Simulation results of rounding errors generating the IR

Quantiles of plot size[1]	1 (smallest plots)	2	3	4 (largest plots)
Plot size (ha)	0.011	0.029	0.062	0.199
Yield (kg/ha)	1674	1060	865	751
Applied fertilizer (% of plots)	14.6	15.6	13.8	12.1
Fertilizer (kg/ha)[2]	29.3	11.6	5.9	2.7
Hired labor (% of plots)	12.6	19.9	24.8	28.3
Hired labor (1000 FBU/ha)[2,3]	62.5	41.3	28.9	14.3
% field in marshland	23.1	11.9	6.4	2.6

[1] 4672, 4704, 4689 and 4689 observation in quartile 1 to 4 respectively.
[2]Zero values included.
[3]FBU: Burundian francs: $1000 FBU = \$0.65$

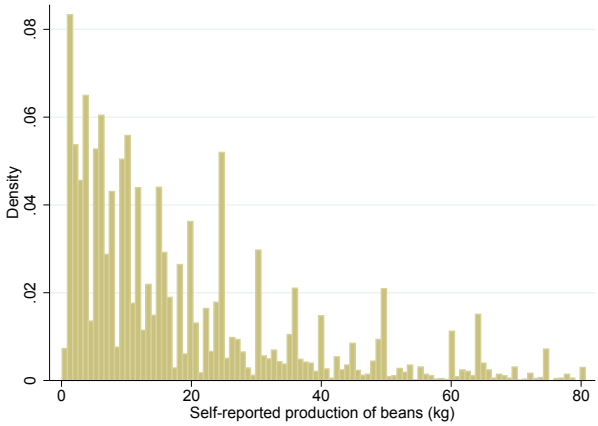
Evidence of rounding errors in self-reported production was apparent for all crops. For instance, figure 7.2 shows self-reported production for beans. There are clearly peaks at 'round' numbers such as 5 kg, 10 kg, 25 kg or 50 kg, indicating rounding errors. A systematic classification of all plots shows that production of at least one crop is a multiple of 5 kg or 10 kg in slightly more than 50% of the observations (table 7.2). Rounding to multiples of 10 kg is twice as likely as rounding to multiples of 5 kg (33% versus 18% of the observations, respectively). This confirms that multiples of 10 kg include both observations that were rounded to the nearest 5kg and the nearest 10 kg. Plot-specific yields are significantly higher when production was reported as a round number. Yields increase from 898 kg/ha if production was not rounded to 1316 kg/ha if it was reported as a multiple of 10 kg. This shows that rounding may lead to a significant overestimation of plot-specific yields. Moreover, it suggests that rounding upwards is more likely than rounding downwards and, thus, supports the first key assumption of the model.

Table 7.3: Distribution of rounded observations

Production reported as:	Number of observations	Yield (kg/ha)
Not rounded	9288	898
Multiple of 5kg	3308	1190
Multiple of 10kg	6158	1316

Table 7.4 shows the number of observations that are reported as a round number for successive quartiles of plot size. For instance, production on 18% and 23% of the smallest plots (first quartile) were reported as 5 kg and 10 kg, respectively. While the number of observations reported as a multiple of 5 kg remains constant with increasing plot size at 18%, the number of observations reported as a multiple

Figure 7.2: Distribution of self-reported production of beans



of 10 kg increases with plot size. This shows that rounding occurs more frequently on larger plots than smaller ones ($\chi^2 = 588, p < 0.001$). It nevertheless also confirms that rounding also occurs on the smallest plots, which is the second key assumption of the model. These findings are discussed in more detail in appendix 7.B where the determinants of rounding behaviour are examined by estimating probit models. This confirms that rounding occurs on large and small plots. Interestingly, we also observed that rounding is more common for some crops such as cassava, potatoes and sweet potatoes. To check if the results are not driven by differences between crops picked up by the 'rounding dummies, the main models were re-estimated on a subsample restricted to plots monocropped with beans (see appendix 7.A).

Table 7.4: Rounded observations (%) by successive quartiles of plot size

Quartiles of plot size	Production reported as:	
	Multiple of 5kg	Multiple of 10kg
1 (smallest plots)	18%	23%
2	18%	29%
3	17%	34%
4 (largest plots)	18%	44%

7.4.2 Econometric results

Given that the two key assumptions of the model – households tend to round upwards and rounding occurs on small and large plots – seem to hold, we estimated the inverse productivity-size relationship with and without controlling for rounding (eq. 7.3 and eq. 7.4). Table 7.5 shows the results. We first estimated the IR

without controlling for rounding with OLS and fixed effects (column 1 and 2), which is the standard approach to estimate the IR. We then attempted to correct for rounding error (column 3 and 4).

The standard approach reveals that there exists a strong inverse relationship between yields and plot size: yields decrease by 40% to 46% if plot size doubles. The IR in Burundi is thus stronger that what is generally found in the literature. Larson et al. (2014), estimating an IR for several African countries, consistently report elasticities ranging from 20% to 35%. In line with expectations, the IR is stronger if household fixed effects are included in the regressions.

Table 7.5: The inverse productivity-size relationship (dependent variable: log of yields)

	Baseline		Correcting for rounding error	
	OLS	Fixed effects	OLS	Fixed effects
Land	-0.407***	-0.463***	-0.444***	−0.510***
Production reported as:				
Multiple of 5kg			0.608***	0.522***
Multiple of 10kg			1.095***	1.013***
Interaction between plot size and rounding:				
Multiple of 5kg x plot size			-0.024	−0.007
Multiple of 10kg x plot size			-0.052***	−0.038**
Household fixed effects	No	Yes	No	Yes
n	18 754	18 754	18 754	18 754
R^2	0.13	0.11	0.20	0.19

Symbols indicate significance levels at: ***≤ 0.01, **≤0.05, *≤0.10

We then attempted to take into account rounding errors (column 3 and 4). Including dummies to control for rounding errors shows that yields are substantially higher if production was reported as a multiple of 5 kg or 10 kg. At average plot size, predicted yields are nearly 60% and 110% higher if production was reported as a multiple of 5 kg or 10 kg, respectively. This confirms that systematic overestimation of yields due to rounding may be a serious concern.

We then tested whether rounding also strengthens the IR by including interactions between plot size and rounding. The strength of the IR did not decrease when including controls for rounding error. The estimated coefficient on plot size even decreases slightly when controls for rounding are included in the estimations. For instance, this coefficient decreased from -0.407 (column 1) to -0.444 (column 3). The interaction terms between plot size and rounding (column 3 and 4) are, however, negative, indicating that rounding strengthens the IR. The magnitude of the interactions is small and only the interaction between plot size and rounding to a multiple of 10 kg is significant. These findings are in line with the simulations which also predicted a small effect of rounding on the IR (ranging from 3% to 8%) and a more pronounced effect for rounding to multiples of 10 kg than to multiples

of 5 kg.

Several robustness checks were conducted. Results remained similar when including plot characteristics in the regression (appendix 7.A). This showed that applying more fertilizer, hiring additional labor and cultivating crops in the fertile marshlands are associated with higher yields. Remarkably, yields are nearly 80% higher on intercropped fields. Perhaps, intercropping signals good soil quality.

As a second robustness check, we limited the sample to plots monocropped with beans. Rounding error is easier to define on monocropped yields than on fields with multiple crops. Moreover, several authors attribute the existence of the IR (partially) to differences in crop composition (Barrett, 1996; Assunção and Braido, 2007). This relationship would indeed be found if the most profitable crops, i.e. those with the highest calorific value per hectare, are cultivated on the smallest plots. Restricting the sample to plots monocropped with beans eliminates this explanation for the IR. Beans are the most common staple crop in Burundi, and were cultivated by 1510 households on 2798 plots. Because few households cultivated several plots of beans, we estimated the IR with simple OLS and did not include household fixed effects. This specification included regional dummies, which partially control for differences in soil quality between regions and household characteristics, as explanatory variables. The results were remarkably similar to the base models: rounding was associated with a systematic overestimation of yields, especially on very small plots, and strengthens the inverse productivity-size relationship.

Finally, we extended our definition of rounding and introduced three additional dummies to control for rounding to multiples of 25 kg, 50 kg and 100 kg, which occurred on 7%, 4% and 4% of the fields respectively. Results (table 7.6) showed that rounding to these numbers caused a systematic overestimation of yields and strengthened the IR. Moreover, the systematic overestimation of yields increased with the strength of the rounding error such that rounding to multiples of 5 kg caused much less bias than rounding to multiples of 100 kg. The bias due to rounding to 5 kg decreased (but remained significant) compared to previous estimates (table 7.5) as this dummy also captured the effect of rounding to multiples of 25 kg in the previous estimations.

7.5 Discussion and conclusions

As far as we know, no previous study has explored rounding error in self-reported production as a potential explanation for the inverse productivity *plot-size* relationship. It is, however, intuitive that if yields are systematically more overestimated on smaller plots than larger ones, this bias would partially or fully explain the inverse productivity plot-size relationship. In our view, rounding of

Table 7.6: Extending the definition of rounding to multiples of 25kg, 50kg and 100kg (dependent variable: log of yields)

	OLS	OLS
Land	−0.473***	−0.444***
Production reported as:		
Multiple of 5kg	0.205***	0.306**
Multiple of 10kg	0.530***	1.254***
Multiple of 25kg	1.056***	1.627***
Multiple of 50kg	1.119***	1.836***
Multiple of 100kg	1.816***	2.284***
Interaction between plot size and rounding:		
Multiple of 5kg x plot size		−0.018
Multiple of 10kg x plot size		−0.119***
Multiple of 25kg x plot size		−0.094***
Multiple of 50kg x plot size		−0.115***
Multiple of 100kg x plot size		−0.076*
n	18 754	18 754
R^2	0.22	0.22

Symbols indicate significance levels at: ***≤ 0.01, **≤0.05, *≤0.10

self-reported production to 'round' numbers such as 5 kg or even 25 kg can generate a systematic and significant overestimation of yields on small plots. On small plots even small rounding errors will translate into large errors in estimated yields per hectare, because the small initial rounding error in self-reported production is amplified by dividing through a small plot size. However, testing this phenomenon econometrically is difficult because we have no 'gold' standard to evaluate whether observed production is over or underreported in comparison to the true value of production. Based on simulations, which also included potential measurement error in plot size, we demonstrated that reasonable assumptions about rounding behavior can generate an inverse productivity-size relationship with an elasticity of 3% to 8%. Next, we attempted to test our explanation using a dataset from Burundi.

With simple descriptive statistics, we first showed that the production figures in our data base were very commonly rounded. Although rounding errors in agricultural surveys are not often discussed, we believe that they occur frequently and that this is not unique to our data (Roberts and Brewer, 2001). For instance, Carletto et al. (2013b, 2015a) report that more than 75% of land size was reported as a 'round number' or a commonly accepted fraction of a round number in an agricultural survey in Uganda. We then showed that yields were substantially higher when production was reported as a round number. This result is already interesting because it shows that rounding matters when estimating average yields. However, it is not sufficient to strengthen the inverse productivity-size relationship. This relationship will only be strengthened if overestimation of yields is greater

on small plots than it is on large plots. In a third step, we attempted to estimate the inverse productivity-size relationship including proxies for rounding interacted with plot size. Using several different specifications, the results consistently showed that rounding errors biased yields upwards. In addition to this novel finding, it seemed that the upwards bias of yields was greater on small than on larger plots. This strengthened the inverse productivity plot-size relation, but only to a limited extent.

Our model hinges on the assumption that rounding upwards is more likely than rounding downwards. This assumption was corroborated by the finding that yields are higher when production was rounded. In addition, Carletto et al. (2015a) find some evidence that rounding of self-reported land is asymmetric, with most farmers over estimating the size of their land. Yet, assuming a tendency towards upwards rounding in self-reported production remains questionable and it is difficult to come up with a rational explanation for this rounding behaviour. As far as we know, no other studies have discussed in depth the occurrence of and the effect on yields of 'rounding' or 'heaping' error in self-reported production. Yet, future studies could easily verify if rounding is associated with higher yields. This research could reveal if this assumption of a tendency to round production 'upwards' holds more generally or is an artifact of our data.

Two assumptions, which have not yet been discussed, are essential when considering rounding as a potential explanation for the IR. First, there must be a large number of extremely small plots for small rounding errors to lead to large errors in estimated yields. Although it is difficult to determine a threshold of plot size above which rounding is no longer a major source of bias in plot-specific yield, it is likely that average plot size in many previous studies of the inverse plot size relationship is sufficiently small that rounding errors may (partially) explain their results. Second, yields and plot size need to be estimated jointly. In our study yields were estimated as the ratio of self-reported production over plot size, but in many agricultural surveys it is common practice to estimate yield based on crop cuts within a predefined quadrant (Fermont and Benson, 2011). With this approach yields and plot size are estimated independently and our theoretical argument that yields on small plots are overestimated relative to yields on larger plots no longer holds. Although it is widely recognized that estimating yields of food crops is inherently difficult in developing countries (Jerven, 2013b), most studies about the IR do not discuss how yields were estimated.

This study should be considered as a first tentative attempt to introduce rounding of self-reported production as a potential explanation for the *plot-size* inverse productivity relationship. Perhaps even more importantly, it is one of the first studies that attempts to investigate the impact of 'rounding' or 'heaping error' in self-reported production on statistical inference. We believe that this type of error is relevant for other self-reported variables in household surveys. From this perspective, the effect of rounding on the inverse productivity-size relationship

may be considered as just one example of a much broader research field that deserves more attention.

Appendix

7.A Sensitivity analyses

Table 7.A.1: Estimating the IR including plot characteristics

	Dependent variable: log of yield (kg/ha)	
	OLS	Household fixed effects
Plot size	-0.462***	-0.541***
Production reported as:		
Multiple of 5kg	0.520***	0.466***
Multiple of 10kg	1.028***	0.962***
Interactions between plot size and rounding:		
Multiple of 5kg X plot size	-0.027	-0.016
Multiple of 10kg X plot size	-0.063***	-0.051***
Plot characteristics		
Anti-erosion protection	0.031	0.081***
Plaine	0.124***	-0.012
Marshland	0.431***	0.354***
Inputs		
Fertilizer used (dummy)	-0.005	0.173***
Wage labour used (dummy)	0.217***	0.221***
Hired labour (FBU/ha)	0.000	0.000
Fertilizer (kg/ha)	0.000***	0.000
Intercropped plot (dummy)	0.497***	0.531***
Season B	0.240***	0.235***
Season C	-0.005	0.001
Constant	8.320***	8.757***
n	18 754	18 754
R^2	0.26	0.25

Symbols indicate significance levels at: ***$\leq$ 0.01, **$\leq$0.05, *$\leq$0.10

Table 7.A.2: Restricting the sample to fields monocropped with beans

	Dependent variable: log of yield (kg/ha)	
	OLS	OLS
Plot size	-0.474***	-0.461***
Production reported as:		
Multiple of 5kg	0.546*	0.481*
Multiple of 10kg	1.432***	1.443***
Interactions between plot size and rounding:		
Multiple of 5kg x plot size	-0.014	0.000
Multiple of 10kg x plot size	-0.095**	-0.099**
Plot characteristics		
Anti-erosion protection		-0.104*
Plaine		-0.104
Marshland		0.176*
Inputs		
Fertilizer used (dummy)		0.126**
Wage labour used (dummy)		0.008
Hired labour (FBU/ha)		0.000
Fertilizer (kg/ha)		0.000
Household characteristics		
Age		0.003
Age^2		0.000
Female headed household		-0.011
Season B		0.024
Season C		-0.053
Constant	8.833***	8.945***
Regional dumies included	No	Yes
n	2798	2798
R^2	0.24	0.26

Symbols indicate significance levels at: ***$\leq$ 0.01, **$\leq$0.05, *$\leq$0.10

7.B Determinants of rounding

The tendency of farmers and enumerators to report production numbers as 'round' numbers is an important element in our model. Ideally, rounding should occur randomly and should not correlate with the 'true' production, household characteristics or the crop that is reported. Yet, this condition is certainly violated. For instance, small production numbers such as 3 kg are probably less frequently reported as a round number than larger production numbers. As a result, the probability of observing a rounded production number increases with 'true' production. Similarly, household and crop characteristics may influence whether production is reported as a round number.

In this appendix, we examine the determinants of rounding behavior. We then discuss whether the finding that rounding occurs non-randomly affects our findings. Since the determinants of rounding to 5 kg, 10 kg or higher multiples (25, 50 or 100 kg) are not necessarily the same, three separate probit models were estimated. In the three models, the rounded production numbers were compared to the observations that were not rounded. The independent variables can be grouped in three categories: (1) (log of) production at plot level and its square (2) household characteristics and (3) crop type. The variable crop production has also been used to define whether production was reported as a round number. Hence, by definition, no observations have been rounded to a multiple of, for instance, 25 kg if observed production is smaller than 25 kg. Ideally, 'rounding' should be regressed on the 'true' production, but this quantity is unobserved. As a robustness check, we re-estimated the probit models using 'plot size' rather than 'production at plot level' as explanatory variable. Results, available upon request, are relatively similar. The main difference is that the positive correlation between plot size and rounding was less pronounced than the correlation between production and rounding.

The results of the estimation of the three probit models are shown in table 7.B.1. A first important finding is that the probability of a rounded observation increases with production (figure 7.B.1). The relation between production and rounding is concave: the probability that an observations is rounded tends to stabilize as soon as a certain threshold of production is reached. Production numbers reported as a multiple of 10 kg, for instance, are relatively uncommon if production is smaller than 20 kg (which corresponds to the value 3 on the log-scale), but nearly 40% of the production numbers are a multiple of 10 kg if production is larger than 60 kg. Similar trends are observed for production numbers reported as multiples of 5 kg or multiples of 25, 50 and 100 kg.

Several household characteristics also correlate with the probability of reporting production as a round number. Surprisingly, total landholdings are negatively correlated with rounding in the three specifications. This contradicts expectations

because it seems intuitive that wealthier households (i.e. those cultivating more land) are less likely to know precisely how much they harvested and, therefore, report their production as a round number. Household size and engaging additional labor on the plot correlates positively with rounding in all specifications. Perhaps, the household head recalls total harvest less precisely if he did not cultivate the plot himself.

Table 7.B.1: The determinants of rounding estimated with probit models

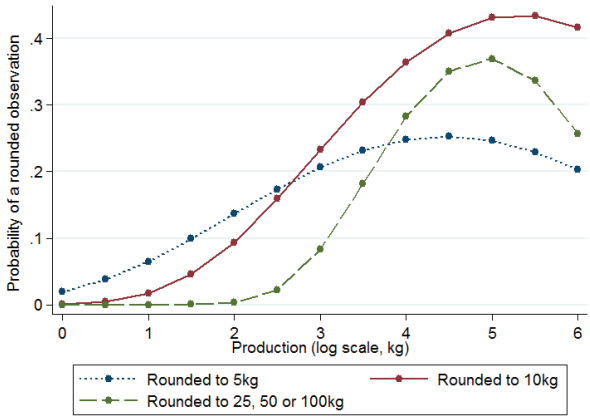
Rounded to a:	Multiple of 5kg	Multiple of 10kg	Multiple of 25kg, 50kg or 100kg
Log of production (kg)	0.640***	1.144***	2.978***
Log of production squared (kg)	-0.0715***	-0.107***	-0.302***
Household characteristics			
Log total landholdings	-0.0544***	-0.0460***	-0.0750***
Literate	0.0314	0.023	0.0283
Age HH head	0.00135	0.00325	-0.00189
Age squared	-0.0000219	-0.0000257	0.000019
Female headed household	-0.0337	-0.0444*	-0.0580*
Hired labour ($1 = yes$)	0.0638***	0.0882***	0.0995***
Household size	0.0139***	0.0140***	0.0199***
Crop type (baseline: other crops)			
Mais	-0.0651*	-0.285***	0.155***
Beans	-0.101***	-0.224***	-0.0643
Peas	-0.111*	-0.409***	0.102
Banana (beer)	-0.342***	-0.392***	-0.873***
Banana (cooking)	-0.398***	-0.354***	-0.922***
Banana (fruit)	-0.455***	-0.465***	-0.816***
Cassava	0.140***	0.212***	0.0957*
Sweet potatoes	0.019	0.187***	-0.223***
Potatoes	0.208***	0.187***	0.259***
Constant	-1.601***	-2.810***	-6.901***
n	43 406	46 014	40 858
Pseudo R^2	0.067	0.187	0.313

Symbols indicate significance levels at: ***$\leq$ 0.01, **$\leq$0.05, *$\leq$0.10

Finally, it turned out that production was more frequently reported as a round number for some of the crops. In all specifications, the crops cassava, sweet potatoes and Irish potatoes were more frequently reported as a round number than most of the other crops. This is an interesting finding that has not been reported previously in the literature. It may be related to the observation that roots and tubers (particularly sweet potatoes and cassava) store better in the soil and are therefore harvested on a daily basis. This may make it more difficult to recall precisely how much was harvested. Bananas, on the other hand, were less likely to be reported as round number. As already mentioned in the data section, bananas were often reported in 'bunches'. Most bunches weighted 8 kg which makes it less likely that total banana harvest is a round number.

Our findings raise several interesting issues about the determinants of rounding,

Figure 7.B.1: The probability of a round production number as function of reported production



which require further research. Given the limitations of our data, we cannot address these issues satisfactory because we only observe reported production and do not observe the 'true' production. Consequently, we do not know if production was rounded or if true production was indeed a round number nor do we know by how much reported production differed from true production. For the same reason, we cannot test whether rounding upwards occurs more frequently that rounding downwards, which is the key assumption in our model offering a new explanation for the plot size inverse productivity relationship.

In the main text, we argued that rounding can cause an overestimation of yields on small plots. Reversed causality is, however, a concern if rounding only occurred if production was large. In this case, it is not rounding that causes high yields, but high yields that cause rounding. The analyses in this appendix confirms that the line of causation runs from rounding to high yields. Although rounding occurs more frequently on larger plots, it also occurs on small plots. For instance, even production numbers between 20 kg and 50 kg (which correspond to 3 and 4 on the log scale) are frequently rounded to multiples of 5 kg (20% of the observations) and multiples of 10 kg (25% of the observations). The finding that the probability of observing rounded production numbers varies with crop type is interesting, but the difference between crops seems too small to drive our results. Moreover, our findings remained similar when restricting the sample to fields monocropped with beans (see table 7.A.2).

CHAPTER 8

Conclusion

8.1 Recapitulation of the research objectives

The standard approach in applied socio-economic work assumes that concepts such as poverty and food security are correctly measured, and uses advanced econometric techniques in an attempt to establish causal relations between concepts. This dissertation approaches applied socio-economic work from a different angle. It does not assume that concepts are correctly measured, but instead examines the arduous task of transforming raw data into robust evidence. It looks at the pitfalls involved in collecting data and quantifying concepts and investigates how to deal with imperfectly measurement concepts in order to obtain reliable quantitative knowledge about the world we live in. This dissertation studies three related research questions: (i) how is data collected? (ii) how can we quantify concepts? and (iii) given that concepts are always imperfectly measured, how do these imperfections affect causal links between concepts? The dissertation is divided into three parts, corresponding to the three research questions.

The first part illustrates that data quality is at least as important as data availability. Socio-economic data is often collected by international institutions or national administrations which have their own objectives and face technical constraints during the data collection process. This being the case, this thesis demonstrates that numbers are not simply an objective representation of the world. They have to be interpreted within the context in which they were generated. This is illustrated in **chapter 2**, which uses data related to agricultural reforms in Rwanda as a case study. It is shown that the success of large-scale agricultural reforms in Rwanda depends on the data used to evaluate it. The statistics of the FAO showed a much greater increase in agricultural yields since the implementation of the reforms than statistics derived from household and agricultural surveys. The discrepancy may be explained by a combination of the inherently difficult task of collecting reliable agricultural data and of political incentives to overestimate the numbers to show that the reforms have worked. This study contributes to the small, but growing literature that is concerned with data quality in Sub-Saharan Africa. This literature, of which Morten Jerven is one of the leading figures, stresses that most data is not collected by disinterested researchers, but is collected within a given context, often by government institutions that have their own objectives. Analyzing and interpreting data without taking this context into account may therefore lead to dubious results. While most research focuses on inaccurate measurement of economic growth, I focused on the agricultural sector for which data availability and quality is arguably an even more serious concern.

The second part studies how concepts can be quantified. I focus on measuring one tangible concept, crop area, and two less tangible concepts, food security and poverty. When researchers want to measure something – be it land area or food security – they develop measurement instruments. Good measurement instruments have to satisfy three criteria: the measurement instrument needs to

valid, accurate and precise under varying circumstances. Evaluating if a particular measurement instrument meets these criteria requires comparing the outcome of the measurement with a gold standard. This is often challenging in socio-economic work since a gold standard may not exist or may only be, at best, considered a good proxy for the measured concept under certain circumstances. Additionally, to test if a measurement instrument works independently from the context in which it is used, instruments need to be tested in many different circumstances. This is also challenging for social researchers as they cannot control the environment in which they work and only passively observe the world (see section 1.3.1 for more details).

In **chapter 3** the accuracy and precision of GPS in measuring crop area is evaluated. In this case a gold standard, the tape and compass method, exists which makes the evaluation straightforward. It is shown that GPS measures crop area accurately, but that the precision of the measurement increases with plot size. GPS measurement is preferable to tape and compass methods for areas of land larger than 1000 m^2. This is a very relevant finding for statistical offices in many developing countries, which are currently debating whether they can use GPS to measure plot size in agricultural surveys (Carletto et al., 2015b). The study also contributes to the growing attention being paid to designing high-quality household surveys.

The other three chapters in the second part deal with measurement instruments of two less tangible concepts, poverty and food security, for which gold standards do not exist. These chapters contribute to the literature on poverty and food security indicators. These indicators are already frequently used to monitor and evaluate development programmes, although their validity has not yet been tested extensively (Coates, 2015). The case studies illustrate three different aspect of validity: cross-sectional validity, inter-temporal validity and internal validity. **Chapter 4** evaluated the Progress out of Poverty Index (PPI). This simple indicator has been developed to determine if a household lives below the poverty line. It does not require detailed expenditure data, which is the standard approach for estimating poverty rates. Instead, it is an asset-based indicator and only requires data on household assets, such as ownership of a radio. The PPI can therefore be implemented with limited resources and is easy to interpret. Using expenditure data as the gold standard, we show that the PPI distinguishes poor from non-poor households. Thus, this indicator is cross-sectionally valid. Moreover, the indicator is, to some extent, robust in relation to changing circumstances and is a valuable tool in development programmes. **Chapter 5** evaluates an indicator of food security, the Household Food Insecurity Access Scale (HFIAS) in which the calorific content of total annual food production is used as a proxy for food security. Although this indicator is useful for distinguishing food secure from food insecure households over the same period, it cannot be used to monitor food insecurity over time. It turns out that food production in the research area decreased over a five year

period, while the HFIAS pointed towards an improvement in the food security situation. Consequently, the HFIAS is cross-sectionally, but not inter-temporally valid. This is a new finding, not previously reported in the literature. **Chapter 6** looks at yet another indicator of food security, the Household Dietary Diversity Score (HDDS). What is innovative in this case study is that we borrowed a methodology from psychometric research, Rasch analysis, to assess the internal validity of the indicator. In contrast to the other methodologies used in this part of the PhD, this approach does not require a gold standard to evaluate the measurement instrument. Rather, the methodology assesses the internal validity of the measurement tool. In other words, it investigates whether the different questions about food consumption that define the measurement instrument measure the same underlying construct of food insecurity. The results suggests that the HDDS is not internally valid and should not be used to measure food insecurity.

The third part of this dissertation goes a step further. It starts by recognizing that concepts are always imperfectly measured. Yet, imperfectly measured concepts can still inform us about the world. However, it is important to investigate how causal relations between concepts are affected by imperfect measurement and how we can deal with imperfect measurement to obtain reliable knowledge. Here it is important to make the distinction between random and systematic measurement errors. The former cause imprecise descriptive statistics and estimates, while the latter can cause imprecise and inaccurate measurements and can cause spurious patterns in the data. As such, systematic measurement error is a threat to obtaining reliable knowledge (see section 1.3.4 in the introduction for more details). **Chapter 7** illustrates that systematic measurement error can generate spurious correlations in the data, showing that accurate and precise measurement is a sine qua non for reliable quantitative knowledge. The chapter shows that the tendency of farmers to report production numbers in multiples of 5 kg or 10 kg can generate a spurious negative correlation between farm size and yields. Hence the stylized fact of the inverse productivity-size relationship can partially be attributed to systematic measurement errors. This case study is noteworthy since it offers a new explanation for the inverse productivity-size relationship. Furthermore, it contributes to the small literature about systematic measurement error.

8.2 Some additional thoughts

Besides the specific findings of the different case studies, three overarching issues emerge from the case studies: data quality, the use and abuse of socio-economic indicators and the relevance (or lack thereof) of evidence-based policies. Although none of the case studies focuses specifically on these topics, taken together they still provide interesting insights into these questions. These overarching themes are discussed in this section, together with some policy recommendations and suggestions for further research.

8.2.1 Improving data quality

Good quality data is an essential step in scientific work. Even the most advanced econometric models still require the input of good quality data. One aspect of data quality is simply 'wrong' numbers. Using 'wrong' numbers has to be avoided to guarantee that wrong research findings do not become accepted truths (see **chapter 2** for an example). Another aspect of data quality is measurement error. Measurement error is always present in applied research. Reducing random measurement error improves descriptive statistics and the precision of estimates. Moreover, as illustrated in this dissertation, systematic measurement error can bias statistical analysis. In my view there are two strategies that can improve data quality: that academic journal articles pay more attention to data quality and a compulsory requirement that datasets be published alongside academic publications.

Many researchers in the social sciences are concerned with the quality of the data they use in their analyses. Yet, data quality is only rarely fully discussed in the data section of academic articles (Schrodt, 2014). It is often just assumed that data are of sufficient quality (Woods, 2014). The 'representativeness of the data' is, on the other hand, often discussed at length. It is indeed important that a sample is 'representative' of a region, a socio-economic group or a sector to enable the research findings to be generalized. Yet one can only consider making generalizations if one believes that the findings are accurate within the sample. Data quality should thus be given as much, if not more, priority as representativeness. This will occur when academic journals encourage discussion of both. At the very least, researchers publishing their work should discuss how the data was collected, by whom, for what purpose, and any potential threats to the precision and accuracy of the data. It is equally important to include a discussion of how potential problems with data quality, such as systematic measurement error, may affect the principal conclusions. This is even more critical if the researchers were not involved in the data collection process, but downloaded or purchased the data from another source. Researchers involved in data collection are more likely to be aware of the strengths and weaknesses of their data (Smith, 2008). If academic journals demanded an explicit discussion of data quality this would drive researchers to reflect upon the quality of their data which would, in turn, improve research quality. This would require that academic journals change their policies to recognize that problems with data quality should be acknowledged as a strength of the study, and not necessarily as a weakness. Currently it is often the case that researchers have an incentive to hide weaknesses in the data they use.

A second way to improve data quality is compulsory publication of the raw data alongside the published academic article. This strategy is currently receiving much attention from scholars and academic institutions (Borgman, 2012; Ghent University, 2015; Hanson et al., 2011; King, 2011; The Economist, 2013). It is argued that open access to datasets would improve research quality (Wicherts et al., 2011)

since it would facilitate the replication of research findings, encourage researchers to pay more attention to data quality and avoid valuable datasets getting lost for future use. Within the framework of my research, three arguments in favor of sharing raw data are particularly relevant: storing and managing valuable datasets, the need for different datasets to check data quality and the need for panel data to validate food security and poverty indicators and to detect measurement error.

First of all, valuable datasets get lost too often. Several surveys with a small sample size have certainly been conducted in Burundi in the last decade by governments, NGOs and research institutions but this data is not publicly available. Moreover, information hidden within the data is often not teased out because of insufficient human or financial capital. As a result, there is no reliable time series information about key indicators such as population growth, agricultural production or land use. Voluntary or even mandatory publication of these datasets would avoid them getting lost, encourage research about Burundi and would increase the availability of reliable statistics. This could, in turn, lead to better policies. Data sharing requires a legal framework and infrastructure to store the data. Both are currently not readily available. Unfortunately, most datasets used in this dissertation – with the exception perhaps of the household surveys in Rwanda – will probably be lost within a decade. This is unfortunate because the data – if carefully analyzed – could help in designing better policies.

Secondly, open access to data would help to improve data quality. The easiest way to check data quality is to compare findings between datasets. For instance, if several different surveys (even with a small sample size) conducted in similar periods and regions, but set up by different organizations using different methodologies, all find similar results, one can be fairly confident about the robustness of the findings. If there are important discrepancies between surveys, one could then examine why they occurred. This could improve the design of surveys and questionnaires. Similarly, to test the robustness of measurement instruments, it is essential that an instrument is tested under many different circumstances. The indicators studied in this dissertations (PPI, HFIAS, HDDS) are frequently used by development programmes. But, given that the data gathered is not publicly available, only a rather limited number of studies have been done to evaluate the precision and accuracy of these indicators.

Thirdly, many research questions can only be answered adequately with panel data. One of the reason that so few papers study the 'inter-temporal validity' of indicators (**chapter 5**) is partly due to the lack of panel data. Studying 'inter-temporal validity' of constructs requires, by definition, panel data, but identifying causal relations is also easier with panels than with cross-sectional data. Moreover, panel data help us to detect outliers or poorly measured variables more than cross-sectional data. If datasets are publicly available, researchers may be encouraged to conduct follow-up surveys, leading to panel datasets over longer periods of time.

Open data is a necessary, but not sufficient, condition for improving data quality in academic research. An example is the agricultural statistics collected and disseminated by FAO. Although widely acknowledged to be of meager quality (**chapter 5**), researchers continue to download and analyze the data. The reason for this is convenience, since it is the only dataset that offers time series of agricultural production since 1961 for more than one hundred countries. It is difficult to develop strategies to avoid poor quality data becoming embedded within statistical systems. Should one consider drastic actions, such as removing all observations that are deemed unreliable from publicly available datasets? Or is it sufficient to signal that the observation is of poor quality and leave it to the wisdom of the individual researcher and reviewers whether or not to use the data? What should be done when better estimates become available? Should the previous data just be replaced with the new and better estimates, taking the risk of losing previous datasets and, hence, making it impossible to replicate research based on the 'wrong datasets'? Such questions need to be addressed at an institutional level and require an open debate within the academic community.

Finally, the convenience of open data may crowd out efforts of individual researchers to collect their own data. One strength of the research field of development economics, compared to many other fields in economics, is that researchers still engage in collecting data themselves, rather than downloading or purchasing it. This is valuable because those researchers often have a good feel for the strengths and weaknesses of their own data and understand the context in which the data were collected. With access to high-quality open data, every incentive to engage in the time-consuming process of data collection may disappear and this may reduce the quality of research. In addition, researchers may be tempted to analyze the same easily accessible datasets over and over again. Schrodt (2014), for instance, roughly estimates that the Oneal-Russett datasets (about democratic peace) have been analyzed over 3000 times[1]. It is highly unlikely that the new analyses lead to new insights. Hence there seems to be a trade-off between the advantages of open data of high-quality with a large sample size, which can only be collected by larger networks of researchers or international institutions, and those of private high-quality data with a smaller sample size collected by individual researchers. The interaction between open data, efforts to collect new data and research quality deserves further research.

8.2.2 Socio-economic indicators

Transforming raw data into measured concepts is a demanding task. It requires a set of rules that defines the process of transformation. Indicators, such as poverty or food security indicators, are examples of such rules **(chapter 4 – 6)**. Socio-economic indicators summarize data into one number that is easy to interpret and

[1] One can think about many other datasets in other research fields that are very popular. For instance, the GTAP data for international trade, the Penn World Tables for purchasing power comparisons, the FADN datasets to study the impact of the Common Agricultural Policy, ...

can be compared between countries and over time. Because indicators facilitate communication with a broader audience, the number of indicators to measure all kinds of phenomena – from food insecurity to corruption to modern slavery – has exploded in recent years (Kelley and Simmons, 2015). This dissertation includes three case studies about indicators: one country-specific poverty indicator and two food security indicators, and assesses their precision, accuracy and robustness with regard to changing circumstances. But the three case studies also hold some more general lessons about socio-economic indicators in general, and indicators of food security and poverty in particular.

"Never take an indicator at its face value" is undoubtedly the main message derived from the case studies. One needs country-specific information to interpret an indicator as well as some basic knowledge about the aims and scope of the indicator. This allows us to put the information summarized by the indicator within a broader context and to realize that every indicator only captures, at best, one aspect of reality. Reducing complexity is indeed the main strength and weakness of all indicators.

It is good to regularly question whether an indicator measures what it pretends to do. By doing so, researchers and policymakers remain aware of the strengths and shortcomings of that indicator. This prevents an indicator taking on a life of its own and becoming more 'real' than reality (Desrosières, 2002/1993), meaning that the complexity of a reality is hidden behind the indicator. The many different available indicators for food security can be beneficial since it encourages researchers to think constantly about which indicator is best suited for their particular project. In addition, regularly evaluating the quality of indicators continuously helps to improve them.

This brings us to an important dilemma: should the development of, and the competition between, indicators of similar concepts be encouraged or should we aim to standardize indicators? Competition has several advantages: it reveals the strengths and weaknesses of different indicators; leads to a continuous improvement of indicators; increases the choice of indicators, allowing researchers and policy makers to choose the one that is the most appropriate under certain conditions. It also prevents an indicator favoured by a powerful elite trumping other potential candidates. On the other hand, standardization is required in order to compare the outcome of measurements between development programmes, between countries and over time. If everyone were to measure living standards with their own favorite index, that may even change from time to time, it just would not be possible to compare countries or to study changes over time. From a research perspective, the standardization of indicators simplifies impact evaluations and systematic reviews of the academic literature.

The tension between competition and standardization coincides with a tension between micro-level and macro-level research. When working with micro-level

data, an indicator should be context-specific and sensitive to small changes, rather than performing well in a more general context. At the micro-level, indicators are useful if they are simple to use and can be quickly implemented, as discussed at length in chapter 4. However, these characteristics may reduce the precision and accuracy of the instrument. For macro-level studies, it is essential to use an index that works in all settings, and less important to have an index that is highly sensitive to small changes since one wants to study global trends. In other words, referring back to equation 1.2 in the introduction, indicators for micro-level studies require great sensitivity $\frac{\partial f}{\partial x} \gg 0$, while indicators for macro-level studies require robustness to changing circumstances $\frac{\partial f}{\partial OC} = 0$. In practice there is often a trade-off between the two. An example of a sensitive, but not very robust measurement instrument for poverty is the Progress out of Poverty Index (PPI) (chapter 4), which can only be used in a single country and needs to be regularly updated to remain accurate. The Multi Dimensional Poverty Index (MPI), currently a very popular poverty indicator, is likely to be less sensitive than the PPI, but is more robust and can be used to compare poverty rates between countries and over time (Alkire and Foster, 2011). The question of too much competition versus too much standardization of socio-economic indicators depends on the research area. At present, ongoing interest in the development of indicators – be it at the global or the micro-level – is driving us towards more competition rather than more standardization[2] .

Finally, it is worth asking if the poverty and food security indicators discussed in this thesis are not really two sides of the same coin. From a theoretical point of view, there is only a subtle difference between poverty and food security. One can be poor and food secure. But, according to most definitions, being extremely poor always implies being food insecure[3]. From a practical point of view, the difference between poverty and food security is even more blurred as most poverty and food security indicators are likely to select the same households. From a policy perspective, few will argue that eradicating poverty is very different from eradicating hunger. This is also implicitly acknowledged by the Millennium Development Goals, which do not differentiate between extreme poverty and food security, but aim to eradicate both extreme poverty and hunger (MDG 1). But if there are so few differences between food security and poverty, is it then necessary to spend so much time and effort on the development of food security indicators and on surveys

[2] In many universities, including Ghent University, many people complain about using articles published in (top) academic journals listed on Web of Science as the sole indicator of research quality, arguing (among others) that this indicator favours publishing 'low-hanging fruit', rather than investing time and effort in truly innovative research. Perhaps, there is too much standardization and too little competition between indicators of research quality at universities!

[3] The term 'nutritional security' is gaining prominence in the field and is likely to replace the term 'food security' in the next decade. Nutritional insecurity is broader than food security as it also focuses on phenomena such as obesity and lack of micro-nutrients. In this paragraph, I use the traditional definition of food insecurity, that is, insufficient energy intake. None of the indicators studies in this PhD aim to measure 'nutritional security'.

primarily concerned with food security? I consider food security indicators to be just another type of poverty indicator. They compete with more frequently-used indicators of poverty and may sometimes be more suited for a particular research or development project, but they are not fundamentally different.

8.2.3 Evidence-based policy

The growing demand for numbers and quantitative studies is partially the result of the popularity of 'evidence-based policies'. Evidence-based policies are defined as *"helping people make well informed decisions about policies, programmes and projects by putting the best available evidence at the heart of policy development and implementation"* (Davies et al., 2000). In other words, the aim is to critically evaluate and design new policies using scientifically rigorous methods. The gold standard of evidence-based policies is the Randomized Controlled Trial (RCT) (Reiss, 2013). RCTs are, however, expensive and cumbersome and are therefore only conducted for large-scale interventions. In consequence, evidence-based policies in practice often involve strengthening monitoring and evaluation and using more 'quantitative indicators' (Hoey, 2015; Pérouse de Montclos, 2012). While it is undeniable that well-executed RCTs can improve policies, it is more doubtful whether 'more numbers' can do so (Whitfield, 2012; Hoey, 2015)[4]. The main reason is that RCTs are set up with the explicit objective of evaluating a policy, while numbers are often just collected to provide situational background. One needs to ask whether and how these numbers lead to better policies.

Nearly all datasets used in this dissertation have been collected with the objective of enhancing the evidence base and formulating sensible policies. I will highlight one example: agricultural surveys. I use data from an agricultural survey in Burundi in several chapters. In the first chapter I also use information from an agricultural survey in Rwanda. Agricultural surveys in developing countries are used to collect detailed information on agricultural production and land use. The aim of the surveys is to develop agricultural policies, but also to compile national accounts. Several multilateral organizations such as the World Bank have recently set up new initiatives to improve the quality of agricultural surveys (World Bank and United Nations and Food and Agricultural Organization, 2010).

Typically, the outcomes of agricultural surveys are official reports loaded with statistics about food production, livestock and land use by region. Yet, these reports do not discuss how these statistics can inform policies. They tend not to present their findings in a way that is useful for developing rural policies. For instance, reporting that the average landholding per household is 0.76 ha in Rwanda does not teach us anything about how to develop the agricultural sector. Policy-

[4]Some scholars have argued that even RCTs do not help in designing better policies (Cartwright and Hardie, 2012; Deaton, 2010). The main argument is a lack of external validity: RCTs can tell if a policy worked in a certain context, but this does not guarantee that new policies, implemented in a different context, will also work.

makers want to know if 0.76 ha of land is sufficient to survive or to understand the reasons for the small average size of landholdings. It is thus doubtful whether such numbers translate directly into better policies. The reason is that numbers per se, or even socio-economic indicators, are not (yet) the form of information upon which policy makers can act. Policy-relevant information requires dialogue, informed not only by numbers and measurement, but also by qualitative studies, between researchers and policy makers.

The findings of chapter 2 neatly illustrate why numbers do not necessarily influence policies. In this chapter agricultural yields in Rwanda are estimated using several datasets. Yields differ substantially between datasets ranging from 1200 kg/ha to over 2000 kg/ha. But whatever the 'correct' number is, it is unlikely that the lowest estimate would result in a different policy than the highest estimate. The finding that numbers do not (always) influence policies is not new. Deaton (2011), for instance, wonders why the upwards revision by 500 million people estimated to live in poverty by the World Bank did not cause any major reaction among the international community. Similarly, Upton et al. (2015) noted that the worldwide prevalence of undernourished people was estimated at 762 million in 1990, but would have been close to one billion using today's method. Deaton (1997) has suggested that global measures have only a limited effect on international policymaking. He argues that global statistics are only popular among scholars interested in long-term trends, but are of little practical use. Perhaps, the same is true for national statistics, such as yields in Rwanda or Burundi. Or, perhaps, estimates of 'yields' do not contain any information that helps policymakers to design policies? Knowing that yields are low is one thing, designing policies to increase them is something completely different.

If so much information has no real implications for policy, why then is so much data collected? Many international organizations engaged in the collection and dissemination of statistics implicitly or explicitly assume that numbers feed into the policy debate (Howlett and Morgan, 2010). However, a careful study of evidence-based policies in developed countries has concluded that *"policy decisions are not deduced primarily from facts and empirical models, but rather from politics, judgement and debate"* (Head, 2010). This is likely to be also true in the context of a developing country. Hence, it is essential to gain a better insight into how numbers and empirical information can be effectively used to shape policies. Arguably, this is the first question that should be addressed before even considering setting up an expensive survey. While there is already an extensive literature on evidence-based policy in developed countries (Head, 2010) and on the uptake of research by policymakers (Amara et al., 2004; Boaz et al., 2008), there is surprisingly little research on the interplay between statistics and policies in developing countries. This would be a highly relevant research question for policymakers as well as for international donors who currently fund most data collection efforts in developing countries.

The previous critical notes about the excessive collection of data with limited relevance should not lead us to conclude that quantitative research in the applied sciences should be abandoned altogether. Numbers can be informative and can lead to better policies, but only in a conducive environment[5]. As already mentioned, what defines such a conducive environment deserves further research and is beyond the scope of this dissertation. Moreover, given the need to monitor the Sustainable Development Goals (SDGs) the demand for numbers in developing countries is likely to grow even more (Jerven, 2014a). The simple fact that policymakers are so keen to use numbers is already sufficient reason for researchers to engage in the quest for accurate numbers. Policymakers need researchers to operationalize the process of collecting, analyzing and interpreting data. Moreover, if there is explicit demand for numbers from policymakers, it is more likely that these numbers will also have an impact on policies compared to a situation in which numbers are merely collected for research purposes or because of international obligations.

It is important, however, that we do not assume that numbers and quantitative evidence are the only set of evidence bases relevant for designing and evaluating policies (Boumans, 2015; Cartwright, 2009). Many researchers, policymakers and international institutes consider quantitative studies as the gold standard of knowledge and of evidence-based policies. They prefer quantitative studies (in particular, RCTs and meta-analyses in systematic reviews) to careful considering all the available evidence. Here researchers have an important role to play. They could attempt to convince policymakers – and parts of their own community – that quantitative knowledge or 'facts' are not necessarily more informative for designing policies than other sets of evidence such as theoretical models, simulations, qualitative arguments, historical case studies and insights from practitioners[6]. What is important is that the quality of each of the different sets of evidence bases is carefully weighed against the others by experts in the field (Pawson, 2002). Researchers – when operating without a hidden agenda and with full transparency – are well-placed to make expert judgements about the quality and the relevance of the different evidence bases and to synthesize the – often contradictory – evidence.

8.3 Concluding remarks

Many economist believe that 'objective' knowledge does exist (Boumans, 2015; Reiss, 2014). They assume that 'facts' can be filtered out from the noisy environ-

[5]The inverse is also true: reliable numbers can improve policies, but poor numbers often signal poor policies.

[6]The television program 'Fact check' of VRT (The Flemish Radio and Television Broadcasting Organization) regularly discusses the accuracy of numbers cited by policymakers. By doing so, it emphasizes that numbers are also man-made and subject to assumptions. This is an important step in the right direction.

ment in which we live in and that relations, that is, mathematical laws, between these 'facts' can be established. This idea was magnificently stated by Nobel-prize winner Milton Friedman is his seminal paper 'The methodology of positive economics'.

> *I venture the judgement, however, that currently in the Western world, and especially in the United States, differences about economic policy among disinterested citizens derive predominant from different predictions about the economic consequences of taking action – differences that in principle can be eliminated by the progress of positive economics - rather than from fundamental differences in basic values, differences about which men can ultimately only fight* (Friedman, 1953).

In this dissertation, I have demonstrated that men not only fight about values, but also – as Friedman calls it – about 'factual evidence'. As I show, observing 'facts' is not straightforward. They are constructed from raw data through a perilous process that requires many questionable assumptions. It starts with the development of a concept which is subsequently translated into a set of rules which determine how raw data are transformed into measurement. Even when the rules are clearly established, the more mundane task of data collection is not without problems. Imperfectly measured concepts may lead to erroneous relations between concepts. The final information is thus man-made and therefore neither objective nor value-free. As such, it is perfectly rational to fight over 'facts' and not only over values.

The conclusion to this dissertation can be concluded with the following words, attributed to Einstein, *'not everything that counts can be counted, and not everything that can be counted counts'*. Perhaps, I could further add *'Even if it can be counted, it may not be counted correctly'*.

Bibliography

Adato, M., Carter, M. R., May, J., 2006. Exploring poverty traps and social exclusion in South Africa using qualitative and quantitative data. The Journal of Development Studies 42(2), 226–247.

Addinson, C., Boto, I., Mofolo, L., 2015. Briefing nr. 40 data: the next revolution for agriculture in ACP countries.

Ali, D. A., Deininger, K., 2014. Is there a farm-size productivity relationship in African agriculture? Evidence from Rwanda. World Bank Policy Research Working Paper (6770).

Ali, D. A., Deininger, K., Goldstein, M., 2014. Environmental and gender impacts of land tenure regularization in Africa: pilot evidence from Rwanda. Journal of Development Economics 110, 262–275.

Alkire, S., Foster, J., 2011. Counting and multidimensional poverty measurement. Journal of Public Economics 95(7), 476–487.

Alkire, S., Santos, M. E., 2010. Acute multidimensional poverty: A new index for developing countries. United Nations development programme human development report office background paper (2010/11).

Alkire, S., Santos, M. E., 2014. Measuring acute poverty in the developing world: robustness and scope of the Multidimensional Poverty Index. World Development 59(0), 251–274.

Altazin, Y., 2014. Les paysans exclus du succès économique rwandais (idées). Le Monde: March 4th 2014.

Amara, N., Ouimet, M., Landry, R., 2004. New evidence on instrumental, conceptual, and symbolic utilization of university research in government agencies. Science Communication 26(1), 75–106.

Amitsa, 2011. Rwanda: rapport mensuel des prix. Nairobi: Regional Agricultural Input Market Information System.

André, C., Platteau, J. P., 1998. Land relations under unbearable stress: Rwanda caught in the Malthusian trap. Journal of Economic Behavior & Organization 34(1), 1–47.

Angrist, J. D., Pischke, J.-S., 2010. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. Journal of Economic Perspectives 24(2), 3–30.

Ansoms, A., 2008a. A Green Revolution for Rwanda? The political economy of poverty and agrarian change. University of Antwerp, Institute of development policy and management (IOB).

Ansoms, A., 2008b. Striving for growth, bypassing the poor? A critical review of Rwanda's rural sector policies. Journal of Modern African Studies 46(1), 1–32.

Ansoms, A., Hilhorst, T., 2014. Losing your land: dispossession in the Great Lakes. African Issues, James Currey, an imprint of Boydell & Brewer.

Ansoms, A., McKay, A., 2010. A quantitative analysis of poverty and livelihood profiles: the case of rural Rwanda. Food Policy 35(6), 584–598.

Ansoms, A., Rostagno, D., 2012. Rwanda's vision 2020 halfway through: what the eye does not see. Review of African Political Economy 39(133), 427–450.

Ansoms, A., Verdoodt, A., Van Ranst, E., 2008. The inverse relationship between farm size and productivity in rural Rwanda. University of Antwerp, Institute of Development Policy and Management-Discussion Paper 2008.

Arimond, M., Ruel, M. T., 2004. Dietary diversity is associated with child nutritional status: evidence from 11 demographic and health surveys. The Journal of Nutrition 134, 2579–2585.

Arimond, M., Wiesmann, D., Becquey, E., Carriquiry, A., Daniels, M. C., Deitchler, M., Fanou-Fogny, N., Joseph, M. L., Kennedy, G., Martin-Prével, Y., Torheim, L. E., 2010. Simple food group diversity indicators predict micronutrient adequacy of women's diets in 5 diverse, resource-poor settings. The Journal of Nutrition 140, 2059–2069.

Assunção, J. J., Braido, L. H., 2007. Testing household-specific explanations for the inverse productivity relationship. American Journal of Agricultural Economics 89(4), 980–990.

Atkin, D., 2013. Trade, tastes, and nutrition in India. American Economic Review 103(5), 1629–63.

Atkinson, A. B., 2001. The strange disappearance of welfare economics. Kyklos 54(2-3), 193–206.

Attaran, A., 2005. An immeasurable crisis? A criticism of the millennium development goals and why they cannot be measured. PLoS medicine 2(10), 955.

Ballard, T., Kepple, A., Cafiero, C., 2013. The food insecurity experience scale: development of a global standard for monitoring hunger worldwide. Rome: FAO.

Baltagi, B., 2008. Econometric analysis of panel data, volume 1. John Wiley & Sons.

Banerjee, A., Duflo, E., 2011. Poor economics: a radical rethinking of the way to fight global poverty. Public Affairs.

Bar, H. Y., Lillard, D. R., 2012. Accounting for heaping in retrospectively reported event data–a mixture-model approach. Statistics in medicine 31(27), 3347–3365.

Barrett, C. B., 1996. On price risk and the inverse farm size-productivity relationship. Journal of Development Economics 51(2), 193–215.

Barrett, C. B., 2010. Measuring food insecurity. Science 327(5967), 825–828.

Barrett, C. B., Marenya, P. P., McPeak, J., Minten, B., Murithi, F., Oluoch-Kosura, W., Place, F., Randrianarisoa, J. C., Rasambainarivo, J., Wangila, J., 2006. Welfare dynamics in rural Kenya and Madagascar. The Journal of Development Studies 42(2), 248–277.

Bateman, B. W., 2001. Make a righteous number: social surveys, the men and religion forward movement, and quantification in American economics. History of political economy 33(5), 57–85.

Baulch, B., Hoddinott, J., 2000. Economic mobility and poverty dynamics in developing countries. The Journal of Development Studies 36(6), 1–24.

Becquey, E., Martin-Prevel, Y., Traissac, P., Dembélé, B., Bambara, A., Delpeuch, F., 2010. The household food insecurity access scale and an index-member dietary diversity score contribute valid and complementary information on household food insecurity in an urban West-African setting. The Journal of Nutrition 140(12), 2233–2240.

Beegle, K., De Weerdt, J., Friedman, J., Gibson, J., 2012. Methods of household consumption measurement through surveys: experimental results from Tanzania. Journal of Development Economics 98(1), 3–18.

Benjamin, D., 1995. Can unobserved land quality explain the inverse productivity relationship? Journal of Development Economics 46(1), 51–84.

Bennett, M., 1941. Wheat in national diets. Wheat Studies 18(2), 37–76.

Bhalla, S. S., Roy, P., 1988. Mis-specification in farm productivity analysis: the role of land quality. Oxford Economic Papers 40(1), 55–73.

Blackwell, M., Honaker, J., King, G., In press. Unified approach to measurement error and missing data: overview. Sociological Methods and Research .

Bland, J. M., Altman, D., 1986. Statistical methods for assessing agreement between two methods of clinical measurement. The lancet 327(8476), 307–310.

Bland, J. M., Altman, D. G., 1995. Comparing methods of measurement: why plotting difference against standard method is misleading. The Lancet 346(8982), 1085–1087.

Blaug, M., 2002. Ugly currents in modern economics. Fact and Fiction in Economics, Cambridge UP pp. 35–56.

Blauw, S. L., Franses, P. H., 2011. The impact of mobile telephone use on economic development of households in Uganda.

Boaz, A., Fitzpatrick, S., Shaw, B., 2008. Assessing the impact of research on policy: a review of the literature for a project on bridging research and policy through outcome evaluation. Policy Studies Institute. London: Kings College .

Bogaert, P., Delinc, J., Kay, S., 2005. Assessing the error of polygonal area measurements: a general formulation with applications to agriculture. Measurement Science and Technology 16(5), 1170.

Bond, T., Fox, C., 2001. Applying the Rasch model. Fundamental measurement in the human sciences. Lawrence Erlbaum Associates, Inc., Mahwah, NJ.

Borgman, C. L., 2012. The conundrum of sharing research data. Journal of the American Society for Information Science and Technology 63(6), 1059–1078.

Boumans, M., 2005a. Measurement in economic systems. Measurement 38(4), 275–284.

Boumans, M., 2005b. Measurement outside the laboratory. Philosophy of science 72(5), 850–863.

Boumans, M., 2013. The role of models in measurement outside the laboratory. Measurement 46(8), 2908–2912.

Boumans, M., 2015. Science outside the laboratory: measurement in field science and economics. Oxford University Press.

Boumans, M. J., 2009. Truth versus precision. In: Logic, methodology and philosophy of science: proceeding of the twelfth international congress, eds., P. Hájek, L. Valdés-Villanueva, D. Westerstahl, pp. 257–269.

Bound, J., Krueger, A. B., 1989. The extent of measurement error in longitudinal earnings data: Do two wrongs make a right? National Bureau of Economic Research .

Bromley, D. W., 2008. Volitional pragmatism. Ecological Economics 68(12), 1–13.

Brown, K., Peerson, J., Kimmons, J., Hotz, C., 2002. Options for achieving adequate intake from home-prepared complementary foods in low-income countries.

Bundervoet, T., 2010. Assets, activity choices, and civil war: evidence from Burundi. World Development 38(7), 955–965.

Cafiero, C., Melgar-Quiñonez, H. R., Ballard, T. J., Kepple, A. W., 2014. Validity and reliability of food security measures. Annals of the New York Academy of Sciences 1331(1), 230–248.

Carletto, C., Gourlay, S., Murray, S., Zezza, A., 2014. Welcome to fantasyland: comparing approaches to land area measurement in household surveys. Washington, DC: World Bank.

Carletto, C., Gourlay, S., Winters, P., 2015a. From guesstimates to gpstimates: land area measurement and implications for agricultural analysis. Journal of African Economies 24(5), 593–628.

Carletto, C., Jolliffe, D., Banerjee, R., 2013a. The emperor has no data! Agricultural statistics in Sub-Saharan Africa. Washington, DC: World Bank.

Carletto, C., Jolliffe, D., Banerjee, R., 2015b. From tragedy to renaissance: improving agricultural data for better policies. The Journal of Development Studies 51(2), 133–148.

Carletto, C., Savastano, S., Zezza, A., 2013b. Fact or artifact: the impact of measurement errors on the farm size-productivity relationship. Journal of Development Economics 103(0), 254–261.

Carroll, R. J., Ruppert, D., Stefanski, L. A., Crainiceanu, C. M., 2012. Measurement error in nonlinear models: a modern perspective. CRC press.

Carter, M. R., Barrett, C. B., 2006. The economics of poverty traps and persistent poverty: an asset-based approach. The Journal of Development Studies 42(2), 178–199.

Carter, M. R., Wiebe, K. D., 1990. Access to capital and its impact on agrarian structure and productivity in Kenya. American Journal of Agricultural Economics 72(5), 1146–1150.

Cartwright, N., 2009. Evidence-based policy: what's to be done about relevance? Philosophical Studies 143(1), 127–136.

Cartwright, N., Hardie, J., 2012. Evidence-based policy: a practical guide to doing it better. Oxford University Press, USA.

Casillas, A., Schulz, E. M., Robbins, S. B., Santos, P. J., Lee, R. M., 2006. Exploring the meaning of motivation across cultures: IRT analyses of the goal instability scale. Journal of Career Assessment 14(4), 472–489.

Chang, H., 2004. Inventing temperature: Measurement and scientific progress. Oxford University Press.

Chayanov, A. V., 1926/1986. AV Chayanov on the theory of peasant economy. Manchester University Press.

Chen, S., Fonteneau, F., Jütting, J., Klasen, S., 2013. Towards a post 2015 framework that counts: aligning global monitoring demand with national statistical capacity development. Paris21: Discussion Paper Series.

Chesher, A., Schluter, C., 2002. Welfare measurement and measurement error. The Review of Economic Studies 69(2), 357–378.

Clay, D. C., Byiringiro, F. U., Kangasniemi, J., Reardon, T., Sibomana, B., Uwamariya, L., Tardif-Douglin, D., 1995. Promoting food security in Rwanda through sustainable agricultural productivity: meeting the challenges of population pressure, land degradation, and poverty.

Coates, J., 2015. Food Insecurity Measurement, chapter 3, p. 51. CRC press.

Coates, J., Swindale, A., Bilinsky, P., 2007. Household Food Insecurity Access Scale (HFIAS) for measurement of household food access: indicator guide (v. 3). Washington, DC: Food and Nutrition Technical Assistance Project.

Cochet, H., 1998. Burundi: questions on the origin and differentiation of an agrarian system. African Economic History (26), 15–62.

Cochet, H., 2004. Agrarian dynamics, population growth and resource management: the case of Burundi. GeoJournal 60(2), 111–122.

Collier, P., Dercon, S., 2014. African agriculture in 50 years: smallholders in a rapidly changing world? World Development 63(0), 92–101.

CountryStat-Burundi, 2013. Retrieved from http://countrystat.org/home.aspx?c=BDI (June 2013).

D'Ambrosio, C., Frick, J. R., 2012. Individual wellbeing in a dynamic perspective. Economica 79(314), 284–302.

DANE, 2011. Anexo pobreza según departamentos. Colombia: Departamento Administrativo Nacional de Estadìstica.

Datt, G., Ravallion, M., 1998. Farm productivity and rural poverty in India. The Journal of Development Studies 34(4), 62–85.

Davies, H., Nutley, S., Smith, P., 2000. Introducing evidence-based policy and practice in public services. What works pp. 1–12.

Dawson, J. W., DeJuan, J. P., Seater, J. J., Stephenson, E. F., 2001. Economic information versus quality variation in cross-country data. Canadian Journal of Economics pp. 988–1009.

Dawson, N., Martin, A., Sikor, T., 2016. Green revolution in Sub-Saharan Africa: Implications of imposed innovation for the wellbeing of rural smallholders. World Development 78, 204–218.

De Ayala, R. J., 2013. Theory and practice of item response theory. Guilford Publications.

De Groote, H., Traoré, O., 2005. The cost of accuracy in crop area estimation. Agricultural systems 84(1), 21–38.

Deaton, A., 1997. The analysis of household surveys: a microeconometric approach to development policy. World Bank Publications.

Deaton, A., 2010. Instruments, randomization, and learning about development. Journal of economic literature 48(2), 424–455.

Deaton, A., 2011. Measuring development: different data, different conclusions. In: 8th AFD-EUDN Conference, Paris.

Deininger, K., Carletto, C., Savastano, S., Muwonge, J., 2012. Can diaries help in improving agricultural production statistics? Evidence from Uganda. Journal of Development Economics 98(1), 42–50.

Deitchler, M., Ballard, T., Swindale, A., Coates, J., 2010. Validation of a measure of household hunger for cross-cultural use. Washington, DC: Food and Nutrition Technical Assistance II Project (FANTA-2).

Derrickson, J. P., Fisher, A. G., Anderson, J. E. L., 2000. The core food security module scale measure is valid and reliable when used with Asians and Pacific islanders. The Journal of Nutrition 130(11), 2666–2674.

Desiere, S., 2015. The carbon footprint of academic conferences: Evidence from the 14th EAAE congress in Slovenia. EuroChoices. .

Desiere, S., D'Haese, M., Niragira, S., 2015a. Assessing the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale (HFIAS) in Burundi. Public Health Nutrition pp. 1–11.

Desiere, S., Niragira, S., D'Haese, M., 2015b. Cow or goat? Population pressure and livestock keeping in Burundi. Agrekon. .

Desiere, S., Staelens, L., D'Haese, M., 2016. When the data sources writes the conclusion: evaluating agricultural policies. Journal of Development Studies. .

Desiere, S., Vellema, W., D'Haese, M., 2015c. A validity assessment of the progress out of poverty index (PPI^{TM}). Evaluation and Program Planning 49(0), 10–18.

Desrosières, A., 2002/1993. The politics of large numbers: A history of statistical reasoning. Harvard University Press.

Detry, J.-F., 2008. Programme d'appui aux populations vulnérables de la province de Ruyigi: analyse de la situation de sécurité alimentaire dans 3 communes de la province de Ruyigi. Brussels: Coopération Technique Belge.

Devarajan, S., 2013. Africa's statistical tragedy. Review of Income and Wealth 59(S1), S9–S15.

Diamond, P. A., Hausman, J. A., 1994. Contingent valuation: is some number better than no number? The Journal of Economic Perspectives pp. 45–64.

Dinh, T., Zeller, M., 2010. Development of operational poverty indicators in Northern Vietnam. Sustainable land use and rural development in mountainous regions of Southeast Asia.

Diskin, P., 1999. Agricultural productivity indicators measurement guide.

Drewnowski, A., Henderson, S., Driscoll, A., Rolls, B., 1997. The dietary variety score: assessing diet quality in healthy young and older adults. Journal of the American Dietetic Association 97, 266–271.

Druilhe, Z., Barreiro-Hurlé, J., 2012. Fertilizer subsidies in Sub-Saharan Africa. ESA Working paper No. 12-04. Rome: FAO.

Dufour, D. L., Staten, L. K., Reina, J. C., Spurr, G., 1997. Living on the edge: dietary strategies of economically impoverished women in Cali, Colombia. American Journal of Physical Anthropology 102, 5–15.

Enste, D. H., Schneider, F. G., 2000. Shadow economies: size, causes, and consequences. Journal of Economic Literature 38(1), 77–114.

European Evaluation Network for Rural Development, 2014. Defining proxy indicators for rural development programmes (working paper).

Evenson, R. E., Gollin, D., 2003. Assessing the impact of the green revolution, 1960 to 2000. Science 300(5620), 758–762.

Faber, M., Schwabe, C., Drimie, S., 2009. Dietary diversity in relation to other household food security indicators. International Journal of Food Safety, Nutrition and Public Health 2(1), 1–15.

Fafchamps, M., Shilpi, F., 2008. Subjective welfare, isolation, and relative consumption. Journal of Development Economics 86(1), 43–60.

FAO, 1982. Estimations des superficies cultivées et des rendements dans les statistiques agricoles. Etude FAO: développement économiqe et social 22. Rome: FAO.

FAO, 1996. Rome declaration on world food security. Rome: FAO.

FAO, 2012a. Action plan of the global strategy to improve agricultural and rural statistics. Rome: FAO.

FAO, 2012b. Guidelines for measuring household and individual dietary diversity. Rome: Food and Agriculture Organization of the United Nations (FAO), the Food and Nutrition Technical Assistance (FANTA) Project.

FAO, 2014. Reference manual: An insight into countrystat- food and agriculture data network. Rome: FAO.

Feder, G., 1985. The relation between farm size and farm productivity: the role of family labor, supervision and credit constraints. Journal of Development Economics 18(2), 297–313.

Fermont, A., Benson, T., 2011. Estimating yield of food crops grown by smallholder farmers. Washington, DC: International Food Policy Research Institute.

Finkelstein, L., 2005. Problems of measurement in soft systems. Measurement 38(4), 267–274.

Foster, A. D., Rosenzweig, M. R., 1996. Technical change and human-capital returns and investments: evidence from the green revolution. The American Economic Review pp. 931–953.

Friedman, M., 1953. The methodology of positive economics. Essays in positive economics 3(3).

Fuller, W. A., 2009. Measurement error models, volume 305. John Wiley & Sons.

Ghent University, 2015. Data management. Retrieved from https://www.ugent.be/en/research/research-staff/organisation/datamanagement.

Gibson, J., Beegle, K., De Weerdt, J., Friedman, J., 2013. What does variation in survey design reveal about the nature of measurement errors in household consumption? The World Bank: Policy Research Working Paper (6372) .

Gibson, J., Kim, B., 2007. Measurement error in recall surveys and the relationship between household size and food demand. American Journal of Agricultural Economics 89(2), 473–489.

Goodman, L. A., Kruskal, W. H., 1979. Measures of association for cross classification. Springer-Verlag, New York, NY.

GoR, 2000. Vision 2020. Kigali: Ministry of Finance and Economic Planning.

GoR, 2005. Organic law determining the use and management of land in Rwanda (n. 08/2005 of 14/07/2005). Kigali: Official Gazette of the Republic of Rwanda.

GoR, 2006. Preliminary poverty update report: integrated living conditions survey 2005/06. Kigali: National Institute of Statistics of Rwanda.

GoR, 2009. Strategic plan for the transformation of agriculture in Rwanda - Phase II (PSTA II) final report. Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2010. National agricultural survey 2008 (NAS 2008). Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2012a. EICV 3: Thematic report agriculture. Kigali: National Institute of Statistics of Rwanda.

GoR, 2012b. The evolution of poverty in Rwanda from 2000 to 2011: results from the household surveys (EICV). Kigali: National Institute of Statistics of Rwanda.

GoR, 2012c. Farm land use consolidation in Rwanda: assessment from the perspective of agriculture sector. Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2012. Rwanda demographic and health survey 2010: Final report. Kigali: National Institute of Statistics of Rwanda.

GoR, 2012. The third integrated household living conditions survey (EICV3). Kigali: National Institute of Statistics of Rwanda.

GoR, 2013. Seasonal agricultural survey report 2013. Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2014. National fertilizer policy. Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2015a. Crop Intensification Program (CIP). Retrieved from www.minagri.gov.rw/index.php?id=618 (February 2015).

GoR, 2015b. Privatized fertilizer importation and distribution system in Rwanda: Contribution to boosting productivity. Kigali: Ministry of Agriculture and Animal Resources.

GoR, 2015c. Rwanda poverty profile report. Kigali: National Institute of Statistics of Rwanda.

Grameen Foundation, 2014a. Global report on poverty measurement with the Progress out of Poverty Index.

Grameen Foundation, 2014b. Progress out of poverty. Retrieved from http://www.progressoutofpoverty.org/ (October 2014).

Groves, R. M., Couper, M. P., 2012. Nonresponse in household interview surveys. John Wiley & Sons.

Günther, I., Maier, J. K., 2014. Poverty, vulnerability, and reference-dependent utility. Review of Income and Wealth 60(1), 155–181.

Haavelmo, T., 1944. The probability approach in econometrics. Econometrica 12, iii–115.

Hanson, B., Sugden, A., Alberts, B., 2011. Making data maximally available. Science 331(6018), 649.

Harris, L., 1991. Stock price clustering and discreteness. Review of Financial Studies 4(3), 389–415.

Hatloy, A., Hallund, J., Diarra, M. M., Oshaug, A., 1999. Food variety, socioeconomic status and nutritional status in urban and rural areas in Koutiala (Mali). Public Health Nutrition 3(1), 57–65.

Head, B. W., 2008. Three lenses of evidence-based policy. Australian Journal of Public Administration 67(1), 1–11.

Head, B. W., 2010. Reconsidering evidence-based policy: key issues and challenges. Policy and Society 29(2), 77–94.

Headey, D., Ecker, O., 2012. Improving the measurement of food security.

Headey, D., Ecker, O., 2013. Rethinking the measurement of food security: from first principles to best practice. Food Security 5, 327–343.

Headey, D. D., 2013. The impact of the global food crisis on self-assessed food security. The World Bank Economic Review 27(1), 1–27.

Hoddinott, J., Yohannes, Y., 2002. Dietary diversity as a food security indicator. Food consumption and nutrition division discussion paper 136.

Hoey, L., 2015. Show me the numbers? Examining the dynamics between evaluation and government performance in developing countries. World Development 70(0), 1–12.

Howlett, P., Morgan, M. S., 2010. How well do facts travel? The dissemination of reliable knowledge. Cambridge University Press.

Hsiao, C., 2003. Analysis of panel data, volume 34. Cambridge university press.

Hyslop, D. R., Imbens, G. W., 2001. Bias from classical and other forms of measurement error. Journal of Business & Economic Statistics 19(4), 475–481.

Imai, K., Yamamoto, T., 2010. Causal inference with differential measurement error: nonparametric identification and sensitivity analysis. American Journal of Political Science 54(2), 543–560.

IMF, 2013a. Regional economic outlook: Sub-Saharan Africa keeping the pace. Washington, DC: IMF.

IMF, 2013b. World economic outlook: Rwanda economic overview. Washington, DC: IMF.

INEC, 2006. Las condiciones de vida de los Ecuatorianos.

Ingelaere, B., 2010. Peasants, power and ethnicity: a bottom-up perspective on Rwanda's political transition. African Affairs 109(435), 273–292.

International Coffee Organization, 2013. Historical data of total production. Retrieved from http://www.ico.org/new_historical.asp (January 2013).

International Initiative for Impact Evaluation (3ie), 2015. From influence to impact: 3ie strategy 2014-2016.

Irz, X., Lin, L., Thirtle, C., Wiggins, S., 2001. Agricultural productivity growth and poverty alleviation. Development Policy Review 19(4), 449–466.

Jayne, T. S., Yamano, T., Weber, M. T., Tschirley, D., Benfica, R., Chapoto, A., Zulu, B., 2003. Smallholder income and land distribution in Africa: implications for poverty reduction strategies. Food Policy 28(3), 253–275.

JCGM, 2008. International vocabulary of metrology – basic and general concepts and associated terms (VIM). Joint Commitee for Guides in Metrology: ISO/IEC Guide 99.

Jensen, R., Miller, N., 2010. A revaled preference approach to measuring undernutrition and poverty using calorie shares.

Jensen, R. T., Miller, N. H., 2008. Giffen behavior and subsistence consumption. The American economic review 98(4), 1553.

Jerven, M., 2013a. Briefing: For richer, for poorer: GDP revisions and Africa's statistical tragedy. African Affairs 112(446), 138–147.

Jerven, M., 2013b. Poor numbers: how we are misled by African development statistics and what to do about it. Cornell University Press.

Jerven, M., 2014a. Benefits and costs of the data for development targets for the post-2015 development agenda. Copenhagen consensus center.

Jerven, M., 2014b. The political economy of agricultural statistics and input subsidies: evidence from India, Nigeria and Malawi. Journal of Agrarian Change 14(1), 129–145.

Jerven, M., Johnston, D., 2015. Statistical tragedy in Africa? Evaluating the data base for African economic development. The Journal of Development Studies 51(2), 111–115.

Jick, T. D., 1979. Mixing qualitative and quantitative methods: triangulation in action. Administrative Science Quarterly .

Jones, A. D., Ngure, F. M., Pelto, G., Young, S. L., 2013. What are we assessing when we measure food security? A compendium and review of current metrics. Advances in Nutrition 4, 481–506.

Kahler, C. W., Strong, D. R., 2006. A Rasch model analysis of DSM-IV alcohol abuse and dependence items in the national epidemiological survey on alcohol and related conditions. Alcoholism: Clinical and Experimental Research 30(7), 1165–1175.

Kalibata, A., Roy, A., 2015. The fertile roots of Rwanda's green revolution (opinion). The Guardian online: February 19th.

Keita, N., Carfagna, E., 2009. Use of modern geo-positioning devices in agricultural censuses and surveys: use of GPS for crop area measurement. Paper presented at the Bulletin of the International Statistical Institute, the 57th Session, 2009, Durban.

Kelley, J. G., Simmons, B. A., 2015. Politics by number: indicators as social pressure in international relations. American Journal of Political Science 59(1), 55–70.

Kennedy, G., Berardo, A., Papavero, C., Horjus, P., Ballard, T., Dop, M., Delbaere, J., Brouwer, I. D., 2010. Proxy measures of household food consumption for food security assessment and surveillance: comparison of the household dietary diversity and food consumption scores. Public Health Nutrition 13(12), 2010–2018.

Kennedy, G. L., Pedro, M. R., Seghieri, C., Nantel, G., Brouwer, I. D., 2007. Dietary diversity score is a useful indicator of micronutrient intake in non-breastfeeding Filipino children. The Journal of Nutrition 137, 472–477.

Kimhi, A., 2006. Plot size and maize productivity in Zambia: is there an inverse relationship? Agricultural Economics 35(1), 1–9.

King, G., 2011. Ensuring the data-rich future of the social sciences. Science 331(6018), 719–721.

Knueppel, D., Demment, M., Kaiser, L., 2010. Validation of the Household Food Insecurity Access Scale in rural Tanzania. Public Health Nutrition 13(3), 360–367.

Koopmans, T. C., 1947. Measurement without theory. The Review of Economic Statistics pp. 161–172.

Kuhn, T. S., 1961. The function of measurement in modern physical science. Isis pp. 161–193.

Lamb, R. L., 2003. Inverse productivity: land quality, labor markets, and measurement error. Journal of Development Economics 71(1), 71–95.

Larsen, A., Lilleor, H., 2013. Beyond the field: impact of farmer field schools on food security and poverty alleviation. Copenhagen, Denmark.

Larson, D. F., Otsuka, K., Matsumoto, T., Kilic, T., 2014. Should African rural development strategies depend on smallholder farms? An exploration of the inverse-productivity hypothesis. Agricultural Economics 45(3), 355–367.

Leroy, J. L., Ruel, M., Frongillo, E. A., Harris, J., Ballard, T. J., 2015. Measuring the food access dimension of food security: a critical review and mapping of indicators. Food and Nutrition Bulletin 36(2), 167–195.

Linacre, J., 2002. Understanding Rasch measurement: optimizing rating scale category effectiveness. Journal of Applied Measurement 3(1), 85–106.

Lobao, L., Brown, L., 1998. Development context, regional differences among young women, and fertility: the Ecuadorean Amazon. Social Forces 76, 819–848.

Luttmer, E. F. P., 2005. Neighbors as negatives: relative earnings and well-being. The Quarterly Journal of Economics 120(3), 963–1002.

Maes, K. C., Hadley, C., Tesfaye, F., Shifferaw, S., Tesfaye, Y. A., 2009. Food insecurity among volunteer AIDS caregivers in Addis Ababa, Ethiopia was highly prevalent but buffered from the 2008 food crisis. The Journal of Nutrition 139(9), 1758–1764.

Maier, M. H., Imazeki, J., 2012. The data game: controversies in social science statistics. ME Sharpe.

Mair, P., Hatzinger, R., 2007. CML based estimation of extended Rasch models with the eRm package in R. Psychology Science 49(1), 26.

Marques, E. S., Reichenheim, M. E., de Moraes, C. L., Antunes, M. M., Salles-Costa, R., 2014. Household food insecurity: a systematic review of the measuring instruments used in epidemiological studies. Public Health Nutrition FirstView, 1–16.

Martin-Prevel, Y., Becquey, E., Tapsoba, S., Castan, F., Coulibaly, D., Fortin, S., Zoungrana, M., Lange, M., Delpeuch, F., Savy, M., 2012. The 2008 food price crisis negatively affected household food security and dietary diversity in urban Burkina Faso. The Journal of Nutrition 142(9), 1748–1755.

Marysse, S., Ansoms, A., Cassimon, D., 2007. The aid 'darlings' and 'orphans' of the Great Lakes region in Africa. The European Journal of Development Research 19(3), 433–458.

Maslow, A., 1943. A theory of human motivation. Psychological Review 50(4), 370–396.

Matul, M., Kline, S., 2003. Scoring change: prizma's approach to assessing poverty.

Maxwell, D., Coates, J., Vaitla, B., 2013. How do different indicators of household food security compare? Empirical evidence from Tigray. Somerville, MA: Feinstein International Center, Tufts University .

Mayer-Schönberger, V., Cukier, K., 2013. Big data: a revolution that will transform how we live, work, and think. Houghton Mifflin Harcourt.

McCloskey, D. N., 1983. The rhetoric of economics. Journal of Economic Literature pp. 481–517.

McKay, A., Greenwell, G., 2007. Methods used for poverty analysis in Rwanda update note.

Mensink, J., 2012. Poverty measures: from production to use. Thesis, London School of Economics and Political Science.

Miller, S. J., 2015. Benford's Law: Theory and Applications. Princeton University Press.

Monitor Group, 2013. The business case for investing in the import and distribution of fertilizer in Rwanda. Kigali: Ministry of Agriculture, USAID.

Morgan, M. S., 1991. The history of econometric ideas. Cambridge University Press.

Morgan, M. S., 2001. Making measuring instruments. History of political economy 33(5), 235–251.

Moursi, M. M., Arimond, M., Dewey, K. G., Trèche, S., Ruel, M. T., Delpeuch, F., 2008. Dietary diversity is a good predictor of the micronutrient density of the diet of 6-to 23-month-old children in Madagascar. The Journal of Nutrition 138(12), 2448–2453.

Murphy, S. P., Allen, L. H., 2003. Nutritional importance of animal source foods. The Journal of Nutrition 133, 3932–3935.

Night, G., Asiimwe, P., Gashaka, G., Nkezabahizi, D., Legg, J., Okao-Okuja, G., Obonyo, R., Nyirahorana, C., Mukakanyana, C., Mukase, F., et al., 2011. Occurrence and distribution of cassava pests and diseases in Rwanda. Agriculture, ecosystems & environment 140(3), 492–497.

Niragira, S., D'Haese, M., D'Haese, L., Ndimubandi, J., Desiere, S., Buysse, J., 2015. Food for survival diagnosing crop patterns to secure lower threshold food security levels in farm households of Burundi. Food and Nutrition Bulletin 36(2), 196–210.

OECD, 2005/2008. Paris declaration and Accra agenda for action. Paris: OECD.

Opsomer, J. D., Jensen, H. H., Pan, S., 2003. An evaluation of the U.S. department of agriculture food security measure with generalized linear mixed models. The Journal of Nutrition 133(2), 421–427.

Pallant, J. F., Tennant, A., 2007. An introduction to the Rasch measurement model: an example using the Hospital Anxiety and Depression Scale (HADS). British Journal of Clinical Psychology 46, 1–18.

Partchev, I., Partchev, M. I., Suggests, M., 2009. Package irtoys .

Pawson, R., 2002. Evidence-based policy: the promise of 'realist synthesis'. Evaluation 8(3), 340–358.

Pawson, R., Wong, G., Owen, L., 2011. Known knowns, known unknowns, unknown unknowns: the predicament of evidence-based policy. American Journal of Evaluation .

Pérez-Escamilla, R., 2012. Can experience-based household food security scales help improve food security governance? Global Food Security 1(2), 120–125.

Pérouse de Montclos, M.-A., 2012. Humanitarian action in developing countries: who evaluates who? Evaluation and program planning 35(1), 154–160.

Piketty, T., 2014. Capital in the 21st century. Cambridge: Harvard University .

Pinstrup-Andersen, P., 2009. Food security: definition and measurement. Food Security 1(1), 5–7.

Pischke, S., 2007. Lecture notes on measurement error. Lecture Notes on Measurement Error .

Poister, T. H., 2008. Measuring performance in public and nonprofit organizations. John Wiley & Sons.

Ponocny, I., 2001. Nonparametric goodness-of-fit tests for the Rasch model. Psychometrika 66(3), 437–459.

Porter, T. M., 1996. Trust in numbers: the pursuit of objectivity in science and public life. Princeton University Press.

Porter, T. M., 2001. Economics and the history of measurement. History of political economy 33(5), 4–22.

Pottier, J., 2006. Land reform for peace? Rwanda's 2005 land law in context. Journal of Agrarian Change 6(4), 509–537.

Pritchard, M. F., 2013. Land, power and peace: tenure formalization, agricultural reform, and livelihood insecurity in rural Rwanda. Land Use Policy 30(1), 186–196.

Przeworski, A., Teune, H., 1966. Equivalence in cross-national research. Public Opinion Quarterly 30(4), 551–568.

Randall, S., Coast, E., 2014. Poverty in African households: the limits of survey and census representations. The Journal of Development Studies pp. 1–16.

Rasch, G., 1960. Probabilistic models for some intelligence and attainment tests. University of Chicago Press, Chicago.

Rask, K., Rask, N., 2014. Measuring food consumption and production according to resource intensity: the methodology behind the cereal equivalent approach.

Rauch, B., Göttsche, M., Brähler, G., Engel, S., 2011. Fact and fiction in EU-governmental economic data. German Economic Review 12(3), 243–255.

Ravallion, M., 2012. Poverty lines across the world.

Ravallion, M., 2014. Can we trust shoestring evaluations? The World Bank Economic Review 28(3), 413–431.

Ravallion, M., Chen, S., Sangraula, P., 2009. Dollar a day revisited. The World Bank Economic Review 23(2), 163–184.

Ravallion, M., Datt, G., Walle, D., 1991. Quantifying absolute poverty in the developing world. Review of Income and Wealth 37(4), 345–361.

Regassa, N., Stoecker, B. J., 2012. Household food insecurity and hunger among households in Sidama district, southern Ethiopia. Public Health Nutrition 15(07), 1276–1283.

Reiss, J., 2013. Philosophy of economics: a contemporary introduction. Routledge.

Reiss, J., 2014. Struggling over the soul of economics: objectivity versus expertise, pp. 131–152. Springer.

République du Burundi, 2006. Monographie de la commune Ngozi. Bujumbura.

République du Burundi, 2013a. Enquête nationale agricole du Burundi: 2011-2012 (enab). Bujumbura.

République du Burundi, 2013b. Enquête nationale agricole du Burundi 2011-2012 (ENAB). Bujumbura.

Reyntjens, F., 2004. Rwanda, ten years on: from genocide to dictatorship. African Affairs 103(411), 177–210.

Rizzo, M., 2011. Rural wage employment in Rwanda and Ethiopia: a review of the current policy neglect and a framework to begin addressing it. Working Paper no. 103. Geneva: International Labour Office.

Roberts, J. M., Brewer, D. D., 2001. Measures and tests of heaping in discrete quantitative distributions. Journal of Applied Statistics 28(7), 887–896.

Rodrik, D., 2008. The new development economics: we shall experiment, but how shall we learn? John F. Kennedy School of Government, Harvard University.

Rose, D., Charlton, K. E., 2002. Quantitative indicators from a food expenditure survey can be used to target the food insecure in South Africa. The Journal of Nutrition 132(11), 3235–3242.

Ruel, M. T., 2003. Operationalizing dietary diversity: a review of measurement issues and research priorities. The Journal of Nutrition 133, 3911–3926.

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

Sahyoun, N. R., Nord, M., Sassine, A. J., Seyfert, K., Hwalla, N., Ghattas, H., 2014. Development and validation of an Arab family food security scale. The Journal of nutrition 144(5), 751–757.

Salzberger, T., Sinkovics, R. R., Schlegelmilch, B. B., 1999. Data equivalence in cross-cultural research: a comparison of classical test theory and latent trait theory based approaches. Australasian Marketing Journal (AMJ) 7(2), 23–38.

Sandefur, J., Glassman, A., 2015. The political economy of bad data: evidence from African survey and administrative statistics. The Journal of Development Studies 51(2), 116–132.

Schedler, A., 2012. Judgment and measurement in political science. Perspectives on Politics 10(01), 21–36.

Schneeweiss, H., Komlos, J., Ahmad, A., 2010. Symmetric and asymmetric rounding: a review and some new results. AStA Advances in Statistical Analysis 94(3), 247–271.

Schøning, P., 2005. Handheld GPS equipment for agricultural statistics surveys: Experiments on area-measurements done during fieldwork for the Uganda pilot census of agriculture, 2003. Rapporter Statistisk sentralbyrå .

Schoors, K., 2000. A note on building a database on Russian banks: fieldwork against the odds .

Schreiner, M., 2010. A simple poverty scorecard for Rwanda .

Schreiner, M., 2014. How do the poverty scorecard and the PAT differ? .

Schrodt, P. A., 2014. Seven deadly sins of contemporary quantitative political analysis. Journal of Peace Research 51(2), 287–300.

Sen, A. K., 1962. An aspect of Indian agriculture. Economic Weekly 14(4-6), 243–246.

Skoufias, E., Davis, B., De La Vega, S., 2001. Targeting the poor in Mexico: an evaluation of the selection of households into Progresa. World Development 29(10), 1769–1784.

Smith, E., 2008. Pitfalls and promises: the use of secondary data analysis in educational research. British Journal of Educational Studies 56(3), 323–339.

Smith, L. C., 1998. Can FAO's measure of chronic undernourishment be strengthened? Food Policy 23(5), 425–445.

Sprangers, M. A. G., Schwartz, C. E., 1999. Integrating response shift into health-related quality of life research: a theoretical model. Social Science & Medicine 48(11), 1507–1515.

Stefanski, L., 2000. Measurement error models. Journal of the American Statistical Association 95(452), 1353–1358.

Steyn, N., Nel, J., Nantel, G., Kennedy, G., Labadorios, D., 2006. Food variety and dietary diversity scores in children: are they good indicators of dietary adequacy? Public Health Nutrition 9(5), 644–650.

Stiglitz, J. E., 2009. GDP fetishism. The Economists' Voice 6(8).

Swindale, A., Bilinsky, P., 2006. Household dietary diversity score (HDDS) for measurement of household food access: indicator guide. Washington, DC: Food and Nutrition Technical Assistance Project (FANTA), Academy for Educational Development.

Swindale, A., Ohri-Vachaspati, P., 2005. Measuring household food consumption: a technical guide. Washington, DC: Academy for Educational Development, Food and Nutrition Technical Assistance Project (FANTA).

Tennant, A., Conaghan, P. G., 2007. The Rasch measurement model in rheumatology: What is it and why use it? When should it be applied, and what should one look for in a Rasch paper? Arthritis Care & Research 57(8), 1358–1362.

The Economist, 2013. Unreliable research: trouble at the lab. October 19th 2013.

The Economist, 2014. Ranking the rankings. November 8th 2014.

The Economist, 2015a. India's health data. June 27th 2015.

The Economist, 2015b. It's not what you spend. May 23rd 2015.

Thirtle, C., Lin, L., Piesse, J., 2003. The impact of research-led agricultural productivity growth on poverty reduction in Africa, Asia and Latin America. World Development 31(12), 1959–1975.

Thorne-Lyman, A. L., Valpiani, N., Sun, K., Semba, R. D., Klotz, C. L., Kraemer, K., Akhter, N., de Pee, S., Moench-Pfanner, R., Sari, M., Bloem, M. W., 2010. Household dietary diversity and food expenditures are closely linked in rural Bangladesh, increasing the risk of malnutrition due to the financial crisis. The Journal of Nutrition 140(1), 182S–188S.

Toledo Vianna, R., Hromi-Fiedler, A., Segall-Correa, A., Pérez-Escamilla, R., 2012. Household food insecurity in small municipalities in Northeastern Brazil: a validation study. Food Security 4(2), 295–303.

Tripathi, L., Mwangi, M., Abele, S., Aritua, V., Tushemereirwe, W. K., Bandyopadhyay, R., 2009. Xanthomonas Wilt: a threat to banana production in East and Central Africa. Plant Disease 93(5), 440–451.

UNDP, 2010. The Real Wealth of Nations: Pathways to Human Development. Human Development Report Office (HDRO), United Nations Development Programme (UNDP).

Upton, J. B., Cissé, J. D., Barrett, C. B., 2015. Food security as resilience: Reconciling definition and measurement .

USAID, 2009. Livelihoods zoning "plus" activity in Burundi: a special report by the famine early warning system network (FEWS NET). Bujumbura: USAID.

USAID, 2014. Poverty assessment tools.

Van de Walle, D., 1998. Targeting revisited. The World Bank Research Observer 13(2), 231–248.

Van Leeuwen, M., 2010. Crisis or continuity? Framing land disputes and local conflict resolution in Burundi. Land Use Policy 27(3), 753–762.

Vandercasteelen, J., Dereje, M., Minten, B., Taffesse, A. S., 2013. Scaling-up adoption of improved technologies: the impact of the promotion of row planting on farmers' teff yields in ethiopia. Licos Discussion paper.

Varian, H. R., 2014. Big data: new tricks for econometrics. The Journal of Economic Perspectives 28(2), 3–27.

Vellema, W., Casanova, A. B., Gonzalez, C., DHaese, M., 2015. The effect of specialty coffee certification on household livelihood strategies and specialisation. Food Policy 57, 13–25.

Vellema, W., Desiere, S., D'Haese, M., 2016. Verifying validity of the household dietary diversity score: an application of rasch modelling. Food and Nutrition Bulletin. .

Verpoorten, M., Arora, A., Stoop, N., Swinnen, J., 2013. Self-reported food insecurity in Africa during the food price crisis. Food Policy 39, 51–63.

Versailles, B., 2012. Country learning notes: Rwanda: Performance contract (imihigo). London: Overseas Development Institute (ODI).

Verwimp, P., 2013. Peasants in power: the political economy of development and genocide in Rwanda. Springer Netherlands.

Wang, H., Heitjan, D. F., 2008. Modeling heaping in self-reported cigarette counts. Statistics in medicine 27(19), 3789.

Webb, J., Mainville, N., Mergler, D., Lucotte, M., Betancourt, O., Davidson, R., Cueva, E., Quizhpe, E., 2004. Mercury in fish-eating communities of the Andean Amazon, Napo River Valley, Ecuador. EcoHealth 1(2), 59–71.

Webb, P., Coates, J., Frongillo, E. A., Rogers, B. L., Swindale, A., Bilinsky, P., 2006. Measuring household food insecurity: why it's so important and yet so difficult to do. The Journal of Nutrition 136(5), 1404S–1408S.

WFP, 2008. Food consumption analysis: calculation and use of the food consumption score in food security analysis. Rome: World Food Programme.

Whitfield, S., 2012. Evidence-based agricultural policy in Africa: critical reflection on an emergent discourse. Outlook on Agriculture 41(4), 249–256.

WHO, 2014. Trade, foreign policy, diplomacy, and health: glossary of globalization, trade and health terms. Retrieved from http://www.who.int/trade/glossary/story028/en/ (March 2014).

Wicherts, J. M., Bakker, M., Molenaar, D., 2011. Willingness to share research data is related to the strength of the evidence and the quality of reporting of statistical results. PloS one 6(11).

Witt, J., Kakabadse, Y., Ortiz, R., Maldonado, L., 1999. Zonas intangibles de la Amazonia Ecuatoriana.

Woods, D., 2014. The use, abuse and omertà on the "noise" in the data: African democratisation, development and growth. Canadian Journal of Development Studies / Revue canadienne d'études du développement 35(1), 120–135.

World Bank, 2013. Rwanda economic update: maintaining momentum with a special focus on Rwanda's pathway out of poverty. Washinton, DC: World Bank.

World Bank and United Nations and Food and Agricultural Organization, 2010. Global strategy to improve agricultural and rural statistics. World Bank Other Operational Studies 12402. Washington,DC: World Bank.

Young, K., Ashby, D., Boaz, A., Grayson, L., 2002. Social science and the evidence-based policy movement. Social Policy and Society 1(03), 215–224.

Zeller, M., Sharma, M., Henry, C., Lapenu, C., 2006. An operational method for assessing the poverty outreach performance of development policies and projects: results of case studies in Africa, Asia, and Latin America. World Development 34(3), 446–464.

Ziliak, S. T., McCloskey, D. N., 2008. The cult of statistical significance: how the standard error costs us jobs, justice, and lives. University of Michigan Press.

List of Figures

List of Tables

APPENDIX A

Summary

Data are key to empirical research. But data by themselves are not yet information. Raw numbers need to be transformed into measurements and, finally, into robust evidence, which can be used to help designing evidence-based policies. In this thesis, three different steps in this transformation are examined: (i) collecting good-quality data; (ii) quantifying concepts and (iii) accounting for the imperfections in quantified concepts to obtain robust evidence. Different challenges are encountered at every step.

This thesis focuses on household survey data from developing countries collected by universities, NGOs or (inter)national institutions with the explicit objective of 'enhancing the evidence base'. Household surveys are still the most important source of information in developing countries where administrative data are often incomplete and where 'big data', such as data from mobile phones, are still in their infancy. This is unlikely to change in the near future. Monitoring the implementation of the Sustainable Development Goals is likely to increase the demand for household surveys even further. More awareness about the process of transforming raw numbers from household survey into robust evidence is therefore indispensable.

The first critical step towards robust evidence is collecting high quality data since using 'wrong numbers' will lead to the 'wrong results'. It is often argued that the lack of data in developing countries impedes the design of sensible policies.

Perhaps even more critical, however, are data of poor quality that are used to design policies or to support far-reaching reforms. The first case study in this thesis illustrates that this is indeed a real threat. Different datasets that purport to measure the impact of large-scale and controversial agricultural reforms on yields in Rwanda provide very different results. However, only the most positive estimates have been incorporated into the international data management system of the FAO, amplifying the risk that these numbers will be accepted as the 'truth' and possibly used for policy design elsewhere.

The second step in the transformation of raw numbers into robust evidence requires quantifying theoretical concepts. The difficulty here is that these concepts are often not directly observable. Household surveys, for instance, are frequently designed to measure the concepts of poverty or food security. Yet, these concepts are not directly observable and require the development of measurement instruments. These measurement instruments are based on a set of rules that define how observable household characteristics should be translated into the unobservable concept. The development of such measurement instruments is challenging and involves making many different assumptions. Moreover, one can always question whether the final measurement instrument measures the concept it is intended to measure and under what circumstances it measures the concept precisely and accurately. Addressing these questions in the social sciences is notoriously difficult because of the lack of gold standards or the absence of benchmarks against which a newly developed measurement instrument can be assessed. Moreover, the validity of measurement instruments should ideally be tested in many different contexts. However, in practice, social scientists work outside of a laboratory and cannot manipulate the context in which they operate.

In this thesis, the challenge of quantifying concepts is illustrated by evaluating the validity of four measurement instruments: GPS to measure the directly observable concept of land area and three poverty and food insecurity indicators, which quantify unobservable concepts. The evaluation of GPS measurement of land area is straightforward as it can be assessed against the gold standard of compass and rope measurement. The evaluation of food security and poverty indicators requires more creativity since gold standards are unavailable. The three case studies of poverty and food security indicators are used to illustrate three different aspect of validity: cross-sectional validity, inter-temporal validity and internal validity. The first indicator, the Progress out of Poverty Index (PPI) in Rwanda, is benchmarked against expenditure data. It turns out that this indicator is cross-sectionally valid, that is, it consistently distinguishes poor from non-poor households. The second indicator, the Household Food Insecurity Access Scale (HFIAS), is benchmarked against total agricultural production. This indicator is cross-sectionally valid, but its inter-temporal validity is questionable. While total food production decreased over a period of five years, the HFIAS pointed towards an improved food security situation over the same period. This implies that the

indicator cannot be used to monitor the evolution of food security over time. The third food security indicator, the Household Dietary Diversity Score (HDDS), is not assessed against an external benchmark. Instead, its internal validity is evaluated using Rasch models. In other words, it is analyzed if the different food groups included in the HDDS measure a single underlying concept. This is not the case, raising the question of what the HDDS actually measures.

Even with good-quality data and excellent measurement instruments, concepts may still be imprecisely or inaccurately measured. Hence, the third and final step of the transformation of raw numbers into robust evidence consists of accounting for these imperfections when establishing (causal) relations between two (or more) imperfectly measured concepts. To illustrate the relevance of accounting for measurement error, it is shown that imprecise measurement of the harvest at plot level can generate a spurious, negative correlation between productivity and plot size. This has implications for the stylized fact of the inverse productivity-size relationship.

The transformation of raw numbers into robust evidence is a long journey with several steps along the way, all of which are decisive for the final outcome. At every step, new challenges need to be tackled. This requires skilful interventions by researchers and an open discussion about the minimum set of assumptions needed to overcome the challenges. These steps also hold some implications for the interpretation of the final outcome of the journey: robust evidence. A first policy implication is that the academic community pays more attention to the issue of data quality. The compulsory publication of the data alongside journal articles would be an important first step in this process. In addition, studying systematic measurement error can help to limit bias in empirical work and to improve survey design. A second implication has to do with the development of measurement instruments, and in particular, poverty and food security indicators. There is definitely a demand for indicators that can quickly estimate the prevalence of poverty and food insecurity at a regional level in order to monitor development programmes, target the most vulnerable household and design policies. Yet, with so many indicators in existence, choosing the one that is most useful for the purpose at hand is complicated since every indicator has its own strengths and weaknesses. More validation exercises of existing indicators could help to clarify the circumstances under which a particular indicator works and/or is useful. An important advantage of these 'validity exercises' is that researchers will remain keenly aware of the shortcomings of a particular indicator, which are likely to be context-specific. Given the existence of so many indicators one can argue that the validation of existing indicators should be prioritized over the development of yet more indicators. Finally, we should remain aware that the principal driver for funding the collection and interpretation of raw numbers is the call for more 'evidence-based policy'. The main - and perhaps unexpected - lesson of this thesis is that 'quantitative evidence' should not be considered the gold standard for the

design of evidence-based policies. Quantitative evidence is man-made and needs to be complemented by other sets of evidence when designing policies. Researchers should be at the forefront of weighing the quality of different evidence bases and of attempting to synthesize them.

APPENDIX B

Samenvatting

Data staan centraal in empirisch onderzoek. Ruwe data op zich is echter nog geen informatie. De data moet eerst worden gemanipuleerd om theoretische concepten te kwantificeren, die op zijn beurt kunnen worden gebruikt om relaties tussen empirische fenomenen te onderzoeken. Pas na deze laatste stap in de transformatie van ruwe data tot informatie wordt de data ook daadwerkelijk bruikbaar om het beleid te ondersteunen. Dit doctoraat analyseert drie stappen van deze transformatie: (1) de data collectie; (2) het kwantificeren van theoretische concepten en (3) de zoektocht naar (causale) relaties tussen imperfect gemeten concepten. Elke stap wordt gekenmerkt door andere uitdagingen.

Dit doctoraat focust op data uit enquêtes afgenomen bij gezinnen in ontwikkelingslanden door universiteiten, NGO's en (inter)nationale organisaties met als doel om 'evidence-based' beleid te ontwikkelen. Deze enquêtes blijven de belangrijkste bron van informatie in ontwikkelingslanden waar administratieve gegevens dikwijls beperkt zijn en 'big data' zoals data van mobiele telefonie nog in hun kinderschoenen staan. Het is niet erg waarschijnlijk dat dit snel zal veranderen in de nabije toekomst. Zo zal het opvolgen van de Sustainable Development Goals de vraag naar enquêtes, bijvoorbeeld rond armoede en honger, enkel maar doen toenemen. Daarom blijft het van groot belang om het proces van de transformatie van ruwe data in informatie goed te begrijpen.

De eerste belangrijke stap in dit proces is het verzamelen van ruwe data van hoge

kwaliteit. Fouten bij het verzamelen van de data leiden immers per definitie tot verkeerde conclusies. Er wordt vaak gesteld dat het gebrek aan data in ontwikkelingslanden het uitwerken van een aangepast beleid bemoeilijkt. Het is wellicht nog gevaarlijker wanneer data van onbetrouwbare kwaliteit worden gebruikt om ingrijpende hervormingen door te voeren. Dat dit een reëel risico is, wordt gellustreerd met een eerste case study uit Rwanda. Het blijkt dat data uit verschillende bronnen leiden tot erg verschillende conclusies met betrekking tot de toename van de voedselproductie sinds het invoeren van ingrijpende en controversiële landbouwhervormingen in 2007. De internationale dataset van de FAO rapporteert echter enkel de meest positieve schattingen. Hierdoor bestaat er een reëel risico dat deze statistieken als de waarheid zullen worden beschouwd en mogelijk zullen worden gebruikt om gelijkaardige hervormingen in andere landen door te voeren.

De tweede stap in de transformatie van ruwe data in informatie vereist het kwantificeren van theoretische concepten. De moeilijkheid is hier dat deze concepten meestal niet eenvoudig te oberveren of definiëren zijn. Veel enquêtes hebben bijvoorbeeld als doel om armoede en voedselzekerheid te meten. Deze concepten zijn niet onmiddellijk waarneembaar waardoor er meetinstrumenten dienen te worden ontwikkeld om ze te kunnen kwantificeren. Meetinstrument in de sociale wetenschappen zijn gebaseerd op een set van regels die éénduidig vastleggen hoe het concept moet worden gemeten. Deze regels zijn tot op zekere hoogte arbitrair doordat er assumpties m.b.t. het concept nodig zijn om kwantificatie mogelijk te maken. Daardoor kan men zich op het einde van de rit steeds afvragen of het meetinstrument nu wel degelijk het concept meet en in welke omstandigheden het concept accuraat en precies wordt gemeten. Deze vragen beantwoorden is erg complex in de sociale wetenschappen omdat er maar zelden goudstandaarden bestaan waartegen de validiteit van het nieuw meetinstrument kan worden afgetoetst. Bovendien is het belangrijk dat nieuwe meetinstrumenten worden getest onder verschillende omstandigheden. In praktijk kunnen sociale wetenschappers de omstandigheden waarin ze het meetinstrument testen niet controleren, waardoor er steeds twijfels blijven bestaan of het meetinstrument ook in een andere context tot goede resultaten zal leiden.

Dit doctoraat illustreert de uitdagingen bij het kwantificeren van concepten met vier case studies. Een eerste case study onderzoekt of GPS een geschikt meetinstrument is om landoppervlaktes te meten. Deze analyse is eenvoudig omdat er een goudstandaard bestaat voor het meten van oppervlaktes waarmee de resultaten van metingen met GPS kunnen worden vergelijken. Omdat er geen goudstandaarden bestaan voor de concepten voedselzekerheid en armoede, is de evaluatie van meetinstrumenten die deze concepten meten complexer. De drie case studies rond indicatoren voor armoede en voedselzekerheid kijken naar drie verschillende aspecten van validiteit waaraan een goede indicator moet voldoen: cross-sectionele validiteit, inter-temporele en interne validiteit. De eerste indicator, de PPI (Progress Out of Poverty Index), werd ontwikkeld om armoede in Rwanda

te meten. De kwaliteit van deze indicator werd bepaald door te vergelijken met consumptie data. Het blijkt dat deze indicator geschikt is om arme gezinnen te onderscheiden van rijkere gezinnen. Dit toont de cross-sectionele validiteit van deze indicator aan. Om de validiteit van de tweede indicator rond voedselzekerheid, de HFIAS (Household Food Insecurity Access Scale), te testen werd deze gecorreleerd met de totale, jaarlijkse landbouwproductie van het gezin. Dit bevestigde evenzeer de cross-sectionele validiteit van deze indicator, maar zijn inter-temporele validiteit bleek beperkt. Gezinnen waarvan de totale voedselproductie daalde over een periode van vijf jaar waren, volgens de indicator, toch minder voedsel onzeker geworden over diezelfde periode. Deze paradox betekent dat deze indicator niet kan worden gebruikt om de evolutie van voedselzekerheid te bestuderen. In de laatste case study werd gekeken naar de HDDS (Household Dietary Diversity Score), een indicator voor voedselzekerheid. In tegenstelling tot de vorige case studies, werd deze indicator niet vergeleken met een andere maatstaf. In plaats daarvan werd de interne validiteit van de indicator onderzocht. Er werd met andere woorden onderzocht of de voedselgroepen, waaruit de indicator is opgebouwd, allemaal gecorreleerd zijn met éénzelfde latente variabele. Dit is niet het geval, waardoor de vraag kan worden gesteld wat de HDDS nu precies meet.

Zelfs met data van hoge kwaliteit en uitstekende meetinstrumenten van essentieel, blijft het mogelijk dat het concept niet voldoende accuraat en precies is gemeten. In de derde en laatste deel van dit doctoraat worden met deze imperfecties rekening gehouden bij de zoektocht naar (causale) relaties tussen imperfect gemeten concepten. Het belang hiervan wordt gellustreerd door aan te tonen dat het inaccuraat meten van landbouwproductie kan leiden tot een bedrieglijke negatieve correlatie tussen productiviteit en land. Dit is mogelijks een nieuwe verklaring voor de 'stylized fact' van de inverse relatie tussen land en productiviteit.

De transformatie van ruwe data in betrouwbare informatie bestaat uit verschillende stappen, die elk van essentieel belang zijn in dit proces. Bij elke stap dienen nieuwe moeilijkheden te worden overwonnen. Onderzoekers spelen daarbij een belangrijke rol. Het is tevens van groot belang dat er een open discussie wordt gevoerd rond keuze voor specifieke assumpties die nodig zijn om de moeilijkheden te overwinnen. De verschillende stappen hebben evenzeer implicaties voor de interpretatie van het eindproduct, de informatie, waarvan we er drie belichten.

Een eerste beleidsimplicatie is dat de academische wereld, en in het bijzonder de wetenschappelijke 'journals', meer aandacht zouden moeten besteden aan data kwaliteit. De verplichte publicatie van de data samen met het wetenschappelijk artikel zou daarbij reeds een eerste belangrijke stap zijn. Daarenboven kan het bestuderen van systematische meetfouten helpen om fouten in empirisch werk te vermijden en om de kwaliteit van enquêtes verder te verbeteren.

Een tweede implicatie heeft betrekking tot het ontwikkelen van meetinstrumenten, in het bijzonder voor het kwantificeren van armoede en voedselzekerheid. Er is

in elk geval vraag naar indicatoren die op een eenvoudige en betrouwbare manier armoede en voedselzekerheid kunnen meten op lokaal niveau zodat ontwikkelingsprogramma's kunnen focussen op de meest kwetsbare groepen en tegelijkertijd de impact van hun programmas kunnen evalueren. De overvloed aan indicatoren, elk met hun eigen sterktes en zwaktes, maakt het echter moeilijk om te bepalen welke indicator het meest geschikt is in een specifieke situatie. Dit vraagt om meer validiteitsstudies die kijken onder welke omstandigheden een specifieke indicator goed werkt. Een bijkomend voordeel hiervan is dat onderzoekers zich bewust zullen blijven van de beperkingen van een specifieke indicator, die wellicht erg context specifiek zijn. Omdat er reeds zoveel indicatoren bestaan lijkt het nuttiger om te focussen op het valideren van bestaande indicatoren dan op het ontwikkelen van nieuwe instrumenten.

Tenslotte blijft het belangrijk om te beseffen dat de belangrijkste drijfveer – en de financiële middelen – voor het verzamelen en analyseren van ruwe data het versterken van 'evidence-based' beleid is. De belangrijkste – en ongetwijfeld onverwachte – conclusie van dit doctoraat is dat 'kwantitatieve informatie' niet moet worden beschouwd als de goudstandaard bij het ontwikkelen van 'evidence-based' beleid. Kwantitatieve informatie is, net zoals kwalitatieve informatie, het resultaat van een stapsgewijs proces waarbij verschillende impliciete en expliciete assumpties dienen te worden gemaakt bij elke stap. Kwantitatieve informatie moet daarom aangevuld worden met andere vormen van informatie bij het ontwikkelen van een beleid. Onderzoekers kunnen een belangrijke rol spelen bij het beoordelen van de kwaliteit van de verschillende vormen van informatie en bij het samenvatten van deze informatie.

APPENDIX C

Curriculum vitae

Academic degrees

2007 – 2010: BSc. in Civil Engineering: applied physics, Ghent University
2010 – 2011: MSc. in Economics, Ghent University

Additional trainings

2011 Information cycle BTC, Belgian Technical Cooperation.
2012 Advanced academic Writing, University Language Center, Ghent University.
2013 Practical French 5, University Language Center.
2013 Categorical data analysis, Ghent University.
2013 Item Response Theory, Flames, KU Leuven.
2013 English Proficiency for Presentations,University Language Center.
2014 Survey analysis, ICES, Ghent University.

Peer reviewed publications

1. Desiere, S., Staelens, L., D'Haese, M., 2016. When the data sources writes the conclusion: evaluating agricultural policies. Journal of Development Studies. (Accepted)

2. Vellema, W., Desiere, S., D'Haese, M., 2016. Verifying validity of the household dietary diversity score: an application of rasch modelling. Food and Nutrition Bulletin. (Accepted)

3. Desiere, S., 2015. The carbon footprint of academic conferences: Evidence from the 14th EAAE congress in Slovenia. EuroChoices. (Accepted, A2-journal)

4. Desiere, S., Niragira, S., D'Haese, M., 2015b. Cow or goat? Population pressure and livestock keeping in Burundi. Agrekon. (Accepted)

5. Desiere, S., D'Haese, M., Niragira, S., 2015a. Assessing the cross-sectional and inter-temporal validity of the Household Food Insecurity Access Scale (HFIAS) in Burundi. Public Health Nutrition pp. 1–11

6. Desiere, S., Vellema, W., D'Haese, M., 2015c. A validity assessment of the progress out of poverty index (PPI^{TM}). Evaluation and Program Planning 49(0), 10–18

7. Niragira, S., D'Haese, M., D'Haese, L., Ndimubandi, J., Desiere, S., Buysse, J., 2015. Food for survival diagnosing crop patterns to secure lower threshold food security levels in farm households of Burundi. Food and Nutrition Bulletin 36(2), 196–210

Participation to conferences

1. Boserup versus Malthus: does population pressure drive agricultural intensification? Evidence from Burundi. August 2015, ICAE, Milan

2. The inverse productivity-size relationship: can it be explained by the rounding of self-reported production? June 2015, 6th EAAE PhD Workshop, Rome.

3. Boserup versus Malthus: does population pressure drive agricultural intensification? Evidence from Burundi. April 2015, 89th AES Conference, Warwick.

4. A validity assessment of the Progress out of Poverty Index $(PPI)^{TM}$ for Rwanda. August 2014, EAAE, Ljubljana.

5. Testing the Household Food Insecurity Access Scale (HFIAS) in Burundi: inconsistent over time? May 2013, 5th EAAE PhD Workshop, Leuven.

6. A green revolution or business as usual? The impact of agricultural reforms in Rwanda. December 2013, GAPSYM7, Ghent.

ISBN 978-9-0598984-8-6