On the Construction of a Consumption Aggregate for Inequality and Poverty Analysis

Giulia Mancini
Giovanni Vecchi

WORLD BANK GROUP

March 2022

1818 H Street NW
Washington DC 20433
Telephone: 202-473-1000
Internet: www.worldbank.org

Report design
Spaeth Hill

Table of contents

Acknowledgments

We are grateful to Martin Ravallion and Salman Zaidi for reviewing these guidelines. We thank Benu Bidani, Carlos Rodríguez-Castelán, Kristen Himelein, and Carolina Sánchez-Páramo for coordinating this project, and for providing helpful comments. We thank Raul Andres Castaneda Aguilar, Nicola Amendola, João Pedro Azevedo, Federico Belotti, Andrea Brandolini, Cesar Cancho, Luigi Cannari, Gero Carletto, Giovanni D'Alessio, Andrew Dabalen, Carolina Diaz-Bonilla, Olivier Dupriez, Maria Gabriela Farfan Bertran, Stefano Fenoaltea, Elizabeth Foster, Margaret Grosh, Dean Jolliffe, Juan Muñoz, Sergio Olivieri, Berk Özler, Marco Ranzani, Silvia Redaelli, Ernesto Savaglio, Dhiraj Sharma, Erwin Tiongson, Roy van de Weide, Ruslan Yemtsov, and Alberto Zezza for useful comments. The authors remain responsible for any remaining errors. We thank Piero Conforti for providing access to the Rural Livelihoods Information System (RuLIS) database. We thank Sédi-Anne Boukaka for excellent research assistance.

This research was sponsored by the Global Unit of the Poverty and Equity Global Practice of the World Bank.

1. Introduction

It has been 20 years since Angus Deaton and Salman Zaidi's *Guidelines for Constructing Consumption Aggregates for Welfare Analysis* first appeared.[1] The paper was conceived as a how-to guide for practitioners in the field of poverty measurement, at the time a small and highly specialized crowd. In fact, its impact in the field of applied poverty research has been far-reaching and enduring beyond expectations. The *Guidelines* are now a key reference for welfare analysts worldwide. In the last five years alone, they have been downloaded 3,154 times (World Bank Documents and Reports Database), while just 2 percent of the World Bank's "knowledge products" surpass 1,000 downloads over a five-year period (Doemeland and Trevino 2014).[2]

Why has Deaton and Zaidi (2002)—henceforth DZ—become so influential? Three answers come to mind. First, the paper deliberately targeted an unmet demand for guidance on the construction of a welfare indicator. The other fundamental building blocks of poverty measurement—poverty lines, poverty measures, and survey data—had all been the subjects of influential publications during the 1990s (Ravallion 1994, 1998; Grosh and Munoz 1996; Deaton 1997). Another factor of DZ's impact is that it successfully straddles the line between theory and practice, without selling either short. The authors lay out solid microeconomic theoretical foundations, and consistently refer back to that framework to resolve the myriad large and small dilemmas facing welfare analysts. At the same time, they remain unabashedly pragmatic: they offer straightforward, concrete recommendations, without ever shying away from reaching a conclusion on ambiguous cases. Finally, the *Guidelines* were developed as part of a massive harmonization effort in the field of poverty measurement, the Living Standards Measurement Study (LSMS), whose reach has only expanded since. As a consequence, DZ's relevance has not just endured, but arguably increased over time.

Two decades after publication, scholars and practitioners alike wonder whether DZ's recommendations still apply, and if not, what is the current "best practice." Among the former, scholarship on the theory and practice of welfare measurement has continued to advance. Among the latter, it is common to encounter a propensity for what is current. Throughout the years spent disseminating DZ within technical assistance and capacity building programs in national statistical offices around the world, we were often gently nudged: "Does a more updated reference exist?". "Newest" does not equal "best," of course—there is no need for an improved version of Dante's Divine Comedy—but the concerns of DZ's readers deserve to be addressed.

---

[1] Following the publication of Deaton's *The Analysis of Household Surveys* (1997), the *Guidelines* were commissioned by Margaret Grosh, then head of the World Bank Living Standards Measurement Study (LSMS) team. Writing began in the summer of 1998, and the document was first circulated as a Princeton working paper in 1999 (Deaton and Zaidi, 1999). The final version, which we refer to throughout this document, was published with minimal alterations in the LSMS Working Paper Series in 2002 (Deaton and Zaidi, 2002).

[2] The Documents and Reports Database's download statistics refer to 2014 onward, while Doemeland and Trevino's refer to the years 2008 to 2012. According to Google Scholar, total citations of the *Guidelines* tally up to 1,234, but traditional academic metrics may be a poor impact measure in this instance, given the *Guidelines*' relevance for applied work in nonacademic circles (such as national statistical offices) where citations are not always used.

This is the main goal of the present document. The ways a "'new" DZ could be envisioned are, we believe, on a spectrum. At one end is *replacement:* an altogether revised set of guidelines, which would render DZ obsolete. We believe this is not attainable, other than by the original authors themselves. Even so, a complete rewriting would not be useful. In many aspects, the *Guidelines* are as sound as ever. At the other end of the spectrum, then, is *repetition*: a transliteration, a digest of DZ. This is clearly pointless. Our wish is to place this work somewhere in the middle: a *reevaluation* of sorts, acknowledging the developments in the literature since the late 1990s, as well as the changing needs of the *Guidelines*' users. To the extent possible in this difficult middle ground where some repetition is inevitable, our choice has been to make this paper self-contained, a feature we believe crucial in making it truly useful. However, the ties to DZ remain close and explicit throughout, and while the reader's familiarity with the original paper is not obligatory, it is certainly helpful.

The contributions of this document can be summarized with three questions. First, do DZ's recommendations stand the test of time, in view of the literature that appeared during the past two decades? Second, when that is not the case, which new guidelines can be put in place? And third, to what extent are DZ's recommendations actually followed in the construction of official poverty measures worldwide? While the two former questions have a *normative* nature, the latter has to do with the *positive* evaluation of the *Guidelines*' de-facto impact—this helps to pinpoint areas where a stronger harmonization effort is needed. Our empirical assessment of the international practice of constructing consumption aggregates is based on the methodological documentation accompanying recent official poverty estimates in 137 countries (Appendix A).

Our target audience largely overlaps with DZ's current heterogeneous group of readers: analysts tasked with constructing consumption aggregates ("those actually doing the calculations") are an obvious focus, but we also hope to reach, more generally, students, economists, statisticians, and other professionals interested in the use and dissemination of poverty measures, as well as statistical officers involved in the production of survey data for poverty analyses. We made every effort to iron out a few asperities that, based on our experience, still make some parts of DZ—namely its theoretical introduction—a tough read for less-tech nical audiences. Relative to DZ, some topics are given more space, and others less, in consideration of the fact that, during 20 years of applied poverty work, practices have solidified and the emphasis on specific issues has changed. Among the topics that are treated in more detail than in the *Guidelines* are those having to do with the quality of survey data (including questionnaire design, nonresponse, and outliers), and those related to the sensitivity and reproducibility of results.

The document is organized as follows. Sections 2 and 3 cover the theoretical underpinnings of welfare measurement, and the choice between income and consumption as a welfare measure. Section 4 covers the practical aspects of constructing a consumption aggregate from survey data in its fundamental components (food, nonfood nondurables, durables, and housing). Sections 5 and 6 discuss adjustments to the consumption aggregate (for cost-of-living differences as well as household size and composition). Section 7 covers data issues (missing and extreme values). Section 8 discusses sensitivity analysis, and section 9 tackles the reproducibility of results. Section 10 provides an updated set of recommendations, setting DZ's original ones beside our assessment two decades later. Appendix C, on income aggregation, and Appendix E, on questionnaire design, are noteworthy.

2. Theory of Welfare Measurement

In this section, we cover the theoretical underpinnings for the measure of welfare whose construction from survey data is discussed in the rest of the document. Welfare (or well-being, or the living standard) is comprised of many facets, not all of them monetary (think of health), or even directly measurable (think of "freedom"). In this report, we adopt a definition of welfare based exclusively on *material* well-being. Thus, we acknowledge at the outset that "welfare" is a big word, used here in a much narrower sense than its common meaning suggests. Our focus will be on one (only one) of many possible dimensions of the standard of living, namely consumption. This is the traditional approach of welfare economics (Slesnick 2001, 8—9), which DZ followed.[3]

Topics overlap with Section 2 of DZ (*Theory of the Measurement of Welfare*), although we make an effort to further spell out the theoretical foundations of welfare measurement for the less-tech nical reader. No matter how strong the temptation to skip the theoretical material, we recommend not to. There is no such thing as good measurement in the absence of a theoretical framework: this part is about theory, but data and their use in empirical analysis remain the ultimate goal.

Readers with a strong background in economics will find this section to be a review of familiar concepts, though they should benefit from seeing them explicitly linked to the end goal of measuring poverty. Readers with only some background in economics will enjoy the highest return from engaging with section 2: a lack of familiarity with the *theoretical* material discussed in the rest of this section is, in our experience, the most significant barrier to a full understanding of *practical* choices made when constructing a welfare measure.

The material of this section has been organized in the following way. Section 2.1 provides a nontechnical overview of the conceptual framework underlying welfare measurement. The upshot of the section is that standard consumer theory indicates *consumption expenditure* as the ideal measure for individual welfare. Section 2.2 provides a rigorous, but still accessible, theoretical framework for the ideas in section 2.1. Finally, in section 2.3 we discuss the recommendations stemming from the theory, in light of developments in the literature and international practice.

[3] Alternative approaches to thinking about welfare exist, most notably Amartya Sen's capability approach (Sen 1985, 1987, 1993), which broadens the theoretical framework of the traditional welfarist setup, and addresses some of its shortcomings (Ravallion 2016, ch 3; Ravallion 2020). However, the capability approach will not be considered here: despite its extraordinary influence on the conceptualization of welfare measurement, its empirical implementation remains a major challenge (Brandolini and D'Alessio 2001; Comin, Qizilbash, and Alkire 2008; Vecchi 2017). Some other important alternatives to monetary welfare measurement are mentioned in section 2.3.

2.1 A bird's-eye view of the theory of welfare measurement

The key assumption of the traditional approach to welfare measurement is that an individual's welfare depends on the consumption of a bundle of goods and services, which we denote by $q = (q_1, q_2, \ldots, q_n)$. This is certainly a narrow depiction of human experience, perhaps unacceptably so, as it may appear to some. In fact, at least in theory, it is not as restrictive as it seems—we could think of q as containing *all* goods that matter for well-being, including items that are not normally thought of as "consumer goods" (health, education, leisure, and so on).[4]

Thus, the complex problem of measuring welfare is recast as a simpler task, that of comparing consumption bundles (that is, baskets of goods and services). Still, determining which of two bundles yields the most welfare to a consumer remains problematic: naturally the answer depends on the consumer's own *tastes*. Economic theory provides a solution. Standard consumer theory thinks of individuals as being able to rank every possible consumption bundle consistently in order of preference, from least to most preferred. This ranking may be translated into numbers—in fact, into a mathematical function that, given any bundle, returns its rank in the preference ordering of the consumer (better liked bundles correspond to higher numbers). Economists call such a function the *utility function*.

The appeal of the economic concept of utility—a number expressing how "happy" a consumer is of one bundle, relative to all other alternatives—to a welfare analyst should be apparent. It suggests the following equivalence:

$$\text{welfare} = u = v(q)$$

where u is the level taken on by the *utility function* $v(.)$ and q is any consumption bundle.

We are moving closer to the ultimate goal of measuring welfare: we can associate it to a number, which is directly linked to the subjective tastes of the consumer, and not to any normative idea of what is "best" for her. What remains unresolved is the problem of measuring utility—a number, yes, but one that describes an abstract concept—using real-life data. To move forward, the theory posits four assumptions:[5]

1. A rational individual chooses the most preferred consumption bundle, given her tastes and her budget constraint. Equivalently, we say that the consumer maximizes her welfare, that is, she maximizes utility.
2. All individuals are alike, that is, they have the same tastes (preferences) and needs.
3. A price exists for each of the goods that contribute to the consumer's well-being.
4. All individuals face the same set of prices.

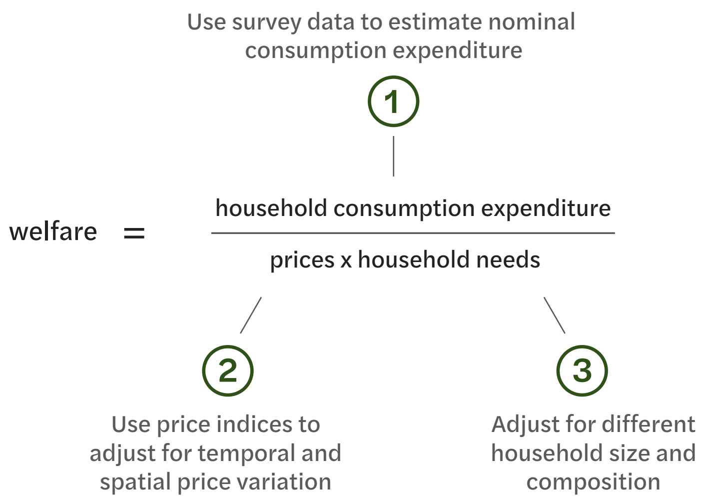
[4] Part of this section draws from a couple of excellent background papers prepared by Erzo Luttmer for the World Bank's 2001 *Croatia: Economic Vulnerability and Welfare Study* (Croatia 2001).

[5] As we shall see, these strong assumptions can and will be "relaxed."

This set of four assumptions allows to define a *metric* for utility—essentially, a unit of measurement—and to do so without having to specify the shape or nature of the function $u(q)$ any further. In particular, economic theory shows that the level of welfare (utility) derived from a consumption bundle q can be represented by the *monetary cost* of the bundle. This is a key result, and this is why economists say that the cost of the consumption bundle is a *money-metric utility* function. Why does current consumption expenditure capture an individual's welfare? The reason is that the individual could have bought a cheaper bundle of goods, but she did not; hence, under the assumption that the consumer maximizes her welfare, she must get a higher level of welfare from the current bundle of goods than from any cheaper bundle of goods.

While the intuition is simple, a formal proof of this result requires some work (which we do in section 3). Far from being an embellishment, such formalization really is essential in this context. A solid theoretical framework is needed to reign in the arbitrariness of intuition.

FIGURE 2.1. Construction of a welfare indicator consistent with consumer theory



SOURCE: Authors' elaboration.

The final step to obtain a welfare indicator consists in a few adjustments to bridge the gap between overly simplifying assumptions, as are the four listed above, and real life. Figure 2.1 summarizes these adjustments and provides a roadmap for the discussion to be developed in the rest of this document. First, theory rests on the hypothesis that all goods and services that concur to welfare be considered, but also that a price exists for each of them. In practice, this is rarely the case: markets may simply not exist for some goods, and survey data most often provide limited and imperfect information. In section 4, we discuss the compromise between a comprehensive theoretical definition and a measure that may be computed using the data available in practice. We have also assumed that individuals face the same set of prices, while in practice we often observe variation in the cost of living: prices vary depending on location and time period (spatial and temporal price variation). In section 5 we discuss adjustments based on price indices to ensure that differences in expenditure indeed reflect differences in consumption (and thus in welfare), not merely in prices. Finally, theory assumes that all individuals have the same tastes and needs—in section 6 we discuss how to relax this hypothesis.

2.2 Economic foundations of welfare measurement

The use of *consumption expenditure* as a measure of individual welfare is the pillar of an entire approach to the measurement of welfare, poverty, and inequality. DZ make it their first recommendation; the aim of this section is to review the theoretical basis for it. Ultimately, we will only need six equations to accomplish our task.

The starting point is consumer theory, the field of economic theory that welfare economists use, a topic covered in any microeconomics course.[6] We imagine a simple economy with only two goods, 1 and 2 (this setting can be readily extended to any number of goods). Quantities of each good are indicated by q_1 and q_2. A combination of goods, $q = (q_1, q_2)$, is a *bundle*. Note that while q_1 and q_2 are scalars, q is a vector. Consumer theory describes how a consumer chooses how much of each good to purchase, given her tastes, and given that she can only afford bundles whose cost does not exceed her budget, which we denote by x. If we denote the prices of the goods with $p = (p_1, p_2)$ then p_1q_1 is the amount of money the consumer spends on good 1, p_2q_2 is the amount of money that goes to good 2, and the *budget constraint* can be written as $p_1q_1 + p_2q_2 \leq x$, which says that the value of consumed goods (left-hand side) cannot exceed the consumer's income (right-hand side). If we replace the "less than or equal to" sign ($\leq$) with an equal sign we obtain the *budget line*, $p_1q_1 + p_2q_2 = x$, which identifies the bundles that just exhaust the consumer's income.

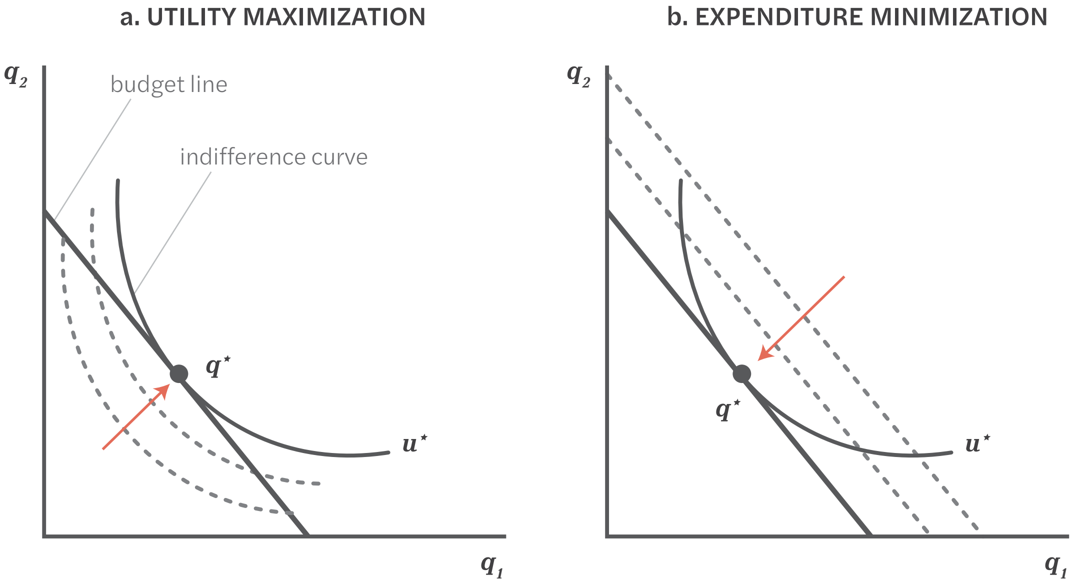
In general, there will be many different bundles that are affordable—which one will be chosen depends on the consumer's preferences. Preferences are described by means of the *utility function*. The utility function, $u = v(q)$, is a mathematical device to assign a number to a bundle: given any two bundles, q^1 and q^2, utility will be higher for whichever one the consumer likes the best. If the consumer prefers q^1 to q^2, then $v(q^1) > v(q^2)$, or, equivalently, $u^1 > u^2$; if she is indifferent between the two bundles, then $v(q^1) = v(q^2)$, or $u^1 = u^2$. The utility function facilitates the task of describing the consumer's preferences—it transforms a complex task, comparing and ranking combinations of goods and services, into a simple one, comparing numbers. Thanks to the utility function, ranking bundles boils down to comparing utility levels.[7]

[6] The reader is referred to introductory-level textbooks (e.g., Varian 2010, ch. 2–5), or to advanced ones (e.g., Deaton and Muellbauer 1980, ch. 2; Varian 1992, ch. 7; Mas-Colell, Whinston, and Green 1995, ch. 3; Jehle and Reny 2011, ch. 1).

[7] Varian (2010, 54): "In Victorian days, philosophers and economists talked blithely of utility as an indicator of a person's overall well-being. Utility was thought of as a numeric measure of a person's happiness. Given this idea, it was natural to think of consumers making choices so as to maximize their utility, that is, to make themselves as happy as possible. The trouble is that these classical economists never really described how we were to measure utility. How are we supposed to quantify the "amount" of utility associated with different choices? Is one person's utility the same as another's? (...) Because of these conceptual problems, economists have abandoned the old-fashioned view of utility as being a measure of happiness. Instead, the theory of consumer behavior has been reformulated (...), and utility is seen only as a way to describe preferences. Economists gradually came to recognize that all that mattered about utility as far as choice behavior was concerned was whether one bundle had a higher utility than another—how much higher didn't really matter."

Given this setup, economic theory builds a model of consumer behavior: in essence, rational consumers are assumed to maximize utility. Individual choice (which is the "best" bundle?) is seen as an optimization problem, constrained by tastes, budgets, and market prices. This problem is visualized in figure 2.2. Panel a of the figure shows the budget constraint (the line with negative slope), and the utility function. Utility is represented by *indifference curves*: each curve represents a set of bundles that leave the consumer indifferent, that is, all bundles that yield the same level of utility. Pick any point (any bundle q) and calculate the corresponding level of utility $u = v(q)$: the indifference curve through q contains all bundles that are equally well liked by the consumer (she is indifferent between choosing either one among them). While movements along an indifference curve leave utility constant, jumping from one curve to another does change the level of the utility. The arrow in the figure shows the direction of the preferred bundles: the further away an indifference curve is from the origin, the higher the consumer's utility. Thus, the consumer maximizes her utility by choosing a bundle that lies on the most outward curve possible. The choice is constrained by her budget: tangency between the budget line and the indifference curve is as far as the consumer can go, so that q^* is the bundle that maximizes her utility, given her budget. When q^* is chosen, then the consumer achieves a utility level equal to $u^* = v(q^*)$, which we use to label the specific indifference curve that contains q^*.

FIGURE 2.2. The consumer maximizes utility or minimizes expenditure



SOURCE: Authors' elaboration.

Panel b in figure 2.2 shows how the consumer makes the same decision as in panel a, by solving a mirror-image problem. This time, the choice to be made is the following: which bundle can be purchased at the minimum expense, while still achieving utility level u^*, precisely the same as in panel a? Graphically, utility is set at u^*, and the consumer jumps from one budget line to the other, toward the origin, until she achieves the line tangent to the indifference curve. The same solution as in panel a, q^*, is found, but the mechanism that leads to it does not imply maximization of utility given a budget, but rather, minimization of the expenditure required to achieve a certain level of utility.

The mechanism illustrated in figure 2.2 can be described with more precision by introducing some mathematical notation. The theory discussed so far can be rephrased as follows:

$$\max u = v(q) \quad \text{subject to} \quad p \cdot q = x \tag{2.0}$$

$$\min x = p \cdot q \quad \text{subject to} \quad v(q) = u \tag{2.1}$$

Equation (2.0), called the consumer's *original problem*, says that the consumer maximizes her utility subject to her budget constraint, and its graphical illustration is panel a of figure 2.2. Equation (2.1), the consumer's *dual problem*, says that the consumer minimizes the expenditure required to attain a certain level of utility, and is illustrated by panel b of figure 2.2. It turns out that for the purpose of measuring poverty, equation (2.1) is more useful than equation (2.0), so in the rest of this section we focus on equation (2.1).[8]

The solution to the minimization problem is the minimum cost of attaining the level of utility u at prices p. Clearly, the minimum cost will vary with u—all other things being equal, the higher the utility level that the consumer wants to achieve, the higher the minimum expenditure required. This idea is captured by the *cost (or expenditure) function*, which we denote as follows:

$$c(u, p) = x \tag{2.2}$$

To interpret the expenditure function in equation (2.2), consider the following mental experiment. Let us fix the prices p faced by the consumer, and let us pick any target level of utility u: what is the minimum amount that the consumer needs to spend, in order to achieve utility u at prices p? The expenditure function answers this question, and the answer is x.

Now that the definition of the cost function is clarified, we introduce more realism to the model and allow for multiple households (we use the superscript h to denote household h) whose utility we want to compare. Because different consumers may face different prices (differences in the cost of living arise over time, or across areas of a country, for instance), comparisons are only valid if we control for differences in purchasing power and keep prices fixed: we denote a set of reference prices by p^0 (more details shortly). This notation allows us to introduce the *money metric utility* (MMU) function:

$$u_m^h = c(u^h, p^0) \tag{2.3}$$

Despite its name being somewhat intimidating, equation (2.3) has a simple economic interpretation: MMU, indicated by u_m^h (the subscript m evokes the concept of money) is the minimum cost for household h of reaching utility level u^h, at prices p^0. Why is MMU so important? Three more equations will deliver an answer.

Using calculus, DZ rewrite equation (2.3) as follows:

$$u_m^h = c(u^h, p^0) \approx p^0 \cdot q^h \tag{2.4}$$

[8] The "dot-product" $p \cdot q$ in equations (2.0) and (2.1) denotes $\Sigma_k p_k q_k$. Notation is consistent with Deaton and Muellbauer (1980), p. 37, which is still considered the main reference for the theory of consumption-based welfare measurement.

Equation (2.4) makes it explicit that MMU u_m^h is simply the cost of a bundle (that is, q^h evaluated at prices p^0. The approximately equal symbol (≈) is a consequence of the math needed to obtain equation (2.4) from equation (2.3), but can be safely ignored in our discussion:[9] we say that, according to equation (2.4), MMU can be approximated by the minimum cost of the bundle q^h chosen by household h, valued at reference prices p^0.

With each step, we are moving from abstract to concrete: equation (2.4) can be written in a more convenient form by introducing the following price index:

$$P^h = \frac{p^h \cdot q^h}{p^0 \cdot q^h} \tag{2.5}$$

P^h in equation (2.5) is known as *Paasche index* (table 5.1 in section 5.1.1 provides more details, not needed at this stage). Similarly to other price indices, Paasche is a device for comparing two price vectors, such as p^h and p^0, by means of a scalar. The Paasche index in equation (2.5) compares prices actually faced by household h, p^h, to the reference set of prices p^0, using q^h as weights. Unlike the Laspeyres index, where weights would be fixed, the Paasche index uses individual weights for each household. Rewriting equation (2.4) after multiplying and dividing its right-hand side by $p^h q^h$, and noting that $p^h q^h = x^h$, produces the following key result:

$$u_m^h \approx \frac{x^h}{P^h} \tag{2.6}$$

Equation (2.6) says that MMU, u_m^h, can be approximated by total household expenditure x^h, deflated with a Paasche price index P^h.

We have arrived at the finish line. Equation (2.6), corresponding to equation 2.6 in DZ's paper, is possibly the single most important equation for welfare measurement within the framework under consideration. Additional adjustments are required, as we shall see, to account for a number of other issues—for instance the fact that we care about individuals rather than households—but the upshot is that equation (2.6) establishes a link between

[9] This is a long and technical footnote, that can be skipped without compromising one's understanding of the general point of this section. To obtain equation (2.4) from equation (2.3), we need to expand the function $c(u^h, p^0)$ around p^h. In math, to expand a function means to transform the function into a polynomial form (e.g. $a_0 + a_1x + a_2x^2 + ...$) – see Chiang (1984, 256–57). In particular, we need to apply the so-called Taylor expansion. Given a function $y = f(x)$, the Taylor expansion consists in transforming the function around a point x_0 into the following polynomial:

$$f(x) = \frac{f(x_0)}{0!} + \frac{f'(x_0)}{1!}(x - x_0) + \frac{f''(x_0)}{1!}(x - x_0)^2 + \cdots$$

where f', f'' denote the first and second derivatives with respect to x of the function. If we use a first-order Taylor approximation, then the general formula simplifies to $f(x) \approx f(x_0) + f'(x_0)(x - x_0)$. DZ apply a first-order Taylor expansion to the cost function $c(u^h, p^0)$ around p^h. This gives:

$$\begin{aligned} c(u^h, p^0) &\approx c(u^h, p^h) + \frac{\partial c(u^h, p^h)}{\partial p^h}(p^0 - p^h) \\ &\approx c(u^h, p^h) + q^h(p^0 - p^h) \end{aligned}$$

where we applied the Shephard's lemma described in Deaton and Muellbauer (1980, 37–40), according to which the partial derivatives of the cost function $c(u^h, p)$ with respect to prices are the (Hicksian or *compensated*) demand functions ($\partial c(u^h, p^h)/\partial p^h \equiv q^h$). Finally, we note that $c(u^h, p^h) = p^h q^h$, so that the equation above further simplifies to $c(u^h, p^0) \approx p^h q^h + q^h(p^0 - p^h)$, that is, $c(u^h, p^0) \approx p^0 q^h$, which corresponds to (2.4). The approximation hinges on the fact that p^0 are not too different from p^h, which is the point where the function is being approximated.

standard economic *theory* (the *left*-hand side, the utility that the consumer derives from a certain consumption bundle) and *practice* (the *right*-hand side, household spending as recorded by survey data, deflated by a price index that adjusts for differences in purchasing power). Equation (2.6) represents the answer that economic theory gives to the question at the very beginning of this section: "How to proxy individual welfare?" DZ's recommendation, based on the material reviewed so far, is clear-cut: "an attempt should be made to use Money Metric Utility (MMU) and to calculate the Paasche price indices with individual household weights."

DZ discuss the possibility of using a Laspeyres index instead of a Paasche: after all, the former is much more popular than the latter, simpler, and easier to explain to policy makers; while the Paasche index is virtually never produced by national statistical offices (NSOs), the calculation of a Laspeyres index is routine for most of them.[10] Would the substitution of the Paasche index in equation (2.6) with a Laspeyres index lead to equivalent results? DZ show that the answer is negative: their argument is a subtle but essential one to understand the preference accorded to MMU.

Let us put equation (2.6) aside for a moment, and consider an alternative indicator of individual welfare, the so-called *welfare ratio* (WR), defined by Blackorby and Donaldson (1987) as the ratio of household expenditure to the expenditure required to match the poverty line: $w^h_{ratio} = x^h/z$, where z denotes the poverty line. The welfare ratio is a pure number that expresses, for each household h, how many times the household can purchase the poverty line basket. If w^h_{ratio} equals, say, 1.5, this means that household consumption expenditure is 1.5 times the value of the poverty line. DZ reformulate w^h_{ratio} so that it is expressed as total household expenditure divided by a price index, which allows for a direct comparison with equation (2.6). The result is that the welfare ratio can be rewritten precisely as total household expenditure adjusted by a Laspeyres index, L^h_z:

$$u^h_{wratio} = \frac{x^h}{L^h_z} \qquad (2.7)$$

Equation (2.7) is simply a transformation (a money-metric representation) of w^h_{ratio}, and will also be referred to as the WR—consistently with DZ's framework and vocabulary.[11] Comparing equations (2.7) to (2.6) leads to the following conclusion: *using a Laspeyres price index to adjust nominal household expenditure is effectively equivalent to using a WR, and not MMU, as a measure of living standards*. Because the two are, in general, different measures of individual welfare, the Paasche index cannot be replaced by a Laspeyres index without altering the nature of how individual welfare is being measured.

What is the difference between the two measures, exactly, and why should MMU in equation (2.6) be preferred to WR in equation (2.7)? The answer is technical. Dividing household expenditure x^h by z to obtain WR, far from being an innocent normalization, is responsible for severing the link between household consumption expenditure (which is what we observe

[10] Differences between price index formulas are discussed at length in section 5.

[11] The suffix z is a reminder that the price index L^h_z uses the goods and services contained in the bundle underlying the poverty line z as reference weights. The math that leads to expressing the welfare ratio as in equation (2.7) is not complicated, but it is omitted here to avoid cluttering the text (see Deaton and Zaidi 2002, 11).

in *practice*) and utility (the concept that we use in *theory* to establish a welfare measure). If the welfare ratio in equation (2.7) is no longer approximately equal to the cost function defined in equation (2.2), then the consequence is that it may fail to measure welfare correctly: it is possible for someone to become better off, that is, to increase her utility, and yet have her welfare ratio decrease. This cannot happen with MMU (Blackorby and Donaldson 1987). This explains why DZ end up recommending the use of MMU in equation (2.6) over WR in equation (2.7).

In fact, the *Guidelines* make it clear that "use MMU" should be taken as a recommendation, not a prescription. DZ were well aware of an irreducible trade-off associated with the choice between equations (2.6) and (2.7). This subtle but important point is worth clarifying. DZ note that, in general, MMU is better than WR as an *individual* welfare measure: MMU is an "exact" measure, as it ranks households consistently with the utility-based theory reviewed in this section, while WR does not necessarily do so. On the other hand, Blackorby and Donaldson (BD) (1987) note that, when individual welfare measures are aggregated—because the interest is in estimating inequality or poverty, for instance, and the computation of inequality and poverty indices requires the aggregation of individual welfare indicators—WR is better than MMU.[12] While DZ attach more value to the first property, and hence suggest to stick to MMU, BD attach more importance to the second property, and hence opt for WR. All in all, the lesson here is that when the analyst sets to the task of measuring individual welfare, she should be prepared to pay a price: if MMU is chosen, then the price is (potential) inaccuracy in distributionally sensitive types of analysis (e.g., cost-benefit analysis). If WR is chosen, the price is the analysis hinging on an inexact individual welfare measure.[13] DZ conclude in favor of what they see as the lesser evil, MMU, and argue that the analyst should *at least attempt* to compute a Paasche price index, and obtain the welfare indicator as in equation (2.6).

To recap, in DZ's view, the use of WR is a second-best strategy, but one well worth implementing when factors other than theoretical considerations play a role. When the estimation of a reliable Paasche index (necessary for calculating MMU) is not a viable strategy—for instance because of a lack of suitable high-quality information—the risk is to compromise the "transparency and simplicity" of the analysis, and the recommendation is to use a Laspeyres index (under the assumption that official estimates exist and are reliable). When both indices are available, the sensitivity analysis discussed in section 8 provides a simple way to figure out the impact of this choice on the statistics of interest.

[12] Unlike MMU, WR offers protection against situations where "rich-to-poor" transfers, for instance, translate into a welfare improvement (Fleurbaey and Maniquet 2011, ch. 1; Decanq, Fleurbaey, and Shokkaert 2015, 95).

[13] If one is willing to assume that preferences are *homothetic* (that is, if consumers can be assumed to have the same consumption pattern irrespective of their income level), then MMU and WR fare equally well both as individual welfare measures and as building blocks of other aggregate welfare measures (in fact, in that specific case WR *is* a form of MMU)—see Ravallion (1998, p. 4). This is of little practical help, however, as homothetic preferences are rarely supported by the evidence.

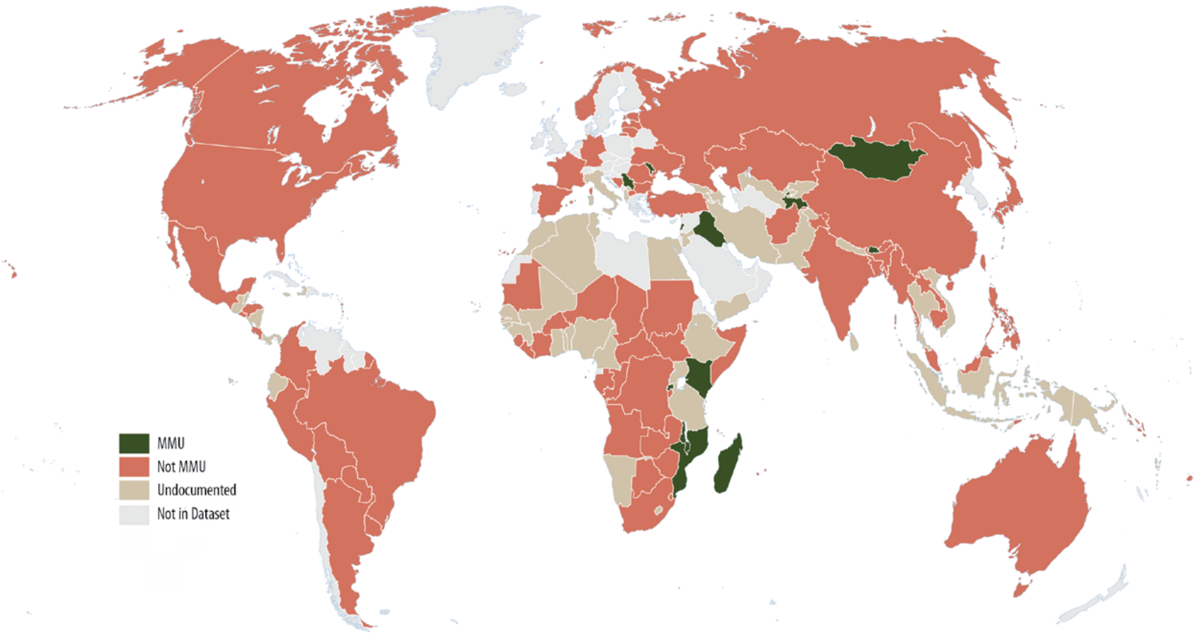
2.3 Discussion

At the roots of the argument for representing welfare by MMU is the theoretical apparatus reviewed in the present section, standard consumer theory. This framework has not even marginally changed in the last 50 years. Nor have recent theoretical developments solved the trade-off underlying the choice between MMU and WR. Therefore, we see no reason to question DZ's fundamental recommendation to grant a preference to measuring welfare using total household expenditure divided by a Paasche price index, and replace the latter with a Laspeyres index when empirical difficulties get in the way of producing accurate estimates.

To what extent is this general recommendation followed in practice? Map 2.1 is constructed on the basis of the technical and methodological documentation underlying official (monetary) poverty estimates in 137 countries (Appendix A elaborates on the construction of this methodological database and its sources). One conclusion can be drawn immediately: only for a scant minority of countries—11, to be exact—can we definitively conclude that the welfare indicator is household expenditure divided by a Paasche index. Granted, a lack of documentation may be hiding more cases of compliance with the *Guidelines*. For 46 countries, marked as "undocumented" in Map 2.1, we could not come to a firm conclusion (mostly because we lack details on the deflation of the nominal aggregate). Even so, at least on first impression, Map 2.1 shows that DZ's first recommendation has not found a wide application in practice.

What drives this result? Of the 68 countries documenting non-MMU approaches, about one-half use a different *numerator* (income, rather than expenditure). A discussion of the merits of income as a welfare indicator deserves no less attention than it did when DZ included it in the table of contents of the *Guidelines*, and it is tackled in section 3 of this document. The remaining countries depart from DZ's recommendation by using a different *denominator*, that is, by deflating household expenditure with indices other than Paasche, or, in a few cases, by not deflating at all (nominal expenditure is the welfare indicator). The adjustment for cost of living differences requires ample discussion, too, in light of a growing literature arguing that price indices available in practice often perch on too narrow a coverage and fail to provide analysts with an accurate proxy of the index that theory demands. These issues are examined in detail in section 5. For now, suffice it to say that below the surface of the blunt categories depicted in Map 2.1 is a lively debate, involving both academics and practitioners, that saw many important contributions since the early 2000s, and whose frontier is moving as we write.

MAP 2.1. Countries that use MMU (expenditure divided by a Paasche index) as a welfare measure



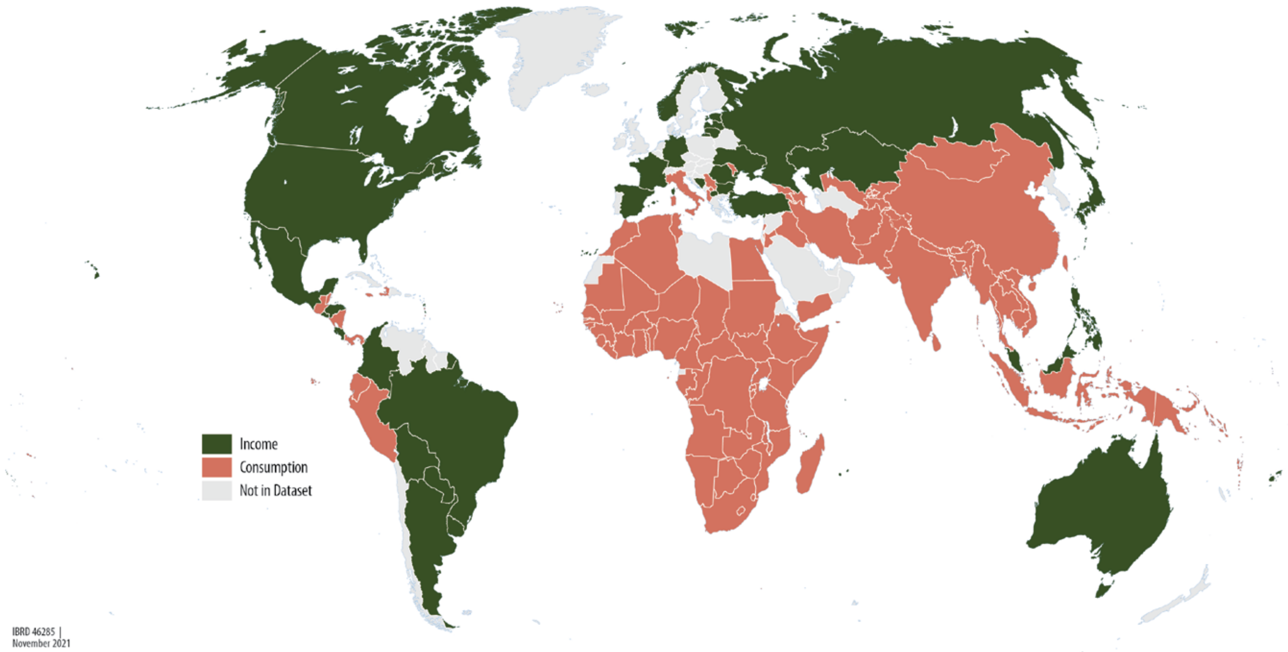
NOTE: For the purposes of this map, "expenditure" can be any aggregate of household expenditures.
SOURCE: Authors' elaboration of the dataset described in Appendix A.

By way of conclusion, we offer a general consideration. Taking a step back to reconsider the MMU versus WR debate two decades later makes the discussion seem rather narrow, when compared with the amount of attention garnered by alternative approaches to the measurement of poverty. The European Union, for instance, has focused on *social exclusion* (Atkinson and Da Voudi 2000; Atkinson et al. 2002; etc.), a concept that implies deprivation in a wide range of economic and social indicators or functionings of living standards. More recently, World Bank (2017) has called attention to the *multidimensional poverty* approach (Alkire et al. 2015; World Bank 2017, 2018), and to *subjective* welfare assessments (Ravallion 2016). Summarizing the discussion on the merits of these evolving methodologies is not within our reach—not in the space of a few pages—and it ultimately is not necessary for the purposes of these guidelines. While these developments are likely to fuel the debate in the near future, they do not call for a complete overhaul of the monetary approach to welfare measurement. Rather, they are a useful complement for what currently remains the "workhorse" of poverty measurement and poverty comparisons worldwide.

3. Consumption or Income?

"Among economic measures of living standards, the main competitor of a consumption-based measure is a measure based on income" (DZ, 11). This is as true now as it was at the time of publication of the *Guidelines*, as shown in Map 3.1. The map shows a neat subdivision of the world into "camps"—Africa and South-East Asia opting for consumption, and the Americas, Europe, and Central Asia adopting income.

MAP 3.1. Consumption versus income-based welfare measurement



SOURCE: Authors' elaboration of the dataset presented in appendix A.

DZ devote a whole section to discussing the relative merits of consumption and income before concluding that "consumption is a theoretically more satisfactory measure of well-being." (p. 21). That is because standard consumer theory points at total consumption expenditure as a utility-consistent measure of welfare (section 2 of this document). However, within the simple, single period model of consumer behavior that motivates this preference, consumption and income are really one and the same: all income is consumed, all consumption is financed by income. Within this basic framework, the choice between the two is inconsequential. But the theory can be readily extended to accommodate a more realistic depiction of consumer choice, one in which decisions are made and funds are allocated over multiple time periods. In this case, of course, consumption and income cease to be identical, the difference being *saving* (or borrowing, that is negative saving). For instance, individuals may want to save at one point in their lives, consuming less than their income, and later dissave, or borrow, to be able to consume more than their income. Ultimately, DZ argue, that in this more complex setting the choice between income and consumption becomes tied with another question: *over which period of time* do we want to measure welfare? This question leads us to reviewing the so-called "smoothness argument" in favor of consumption.

3.1 The "smoothness argument"

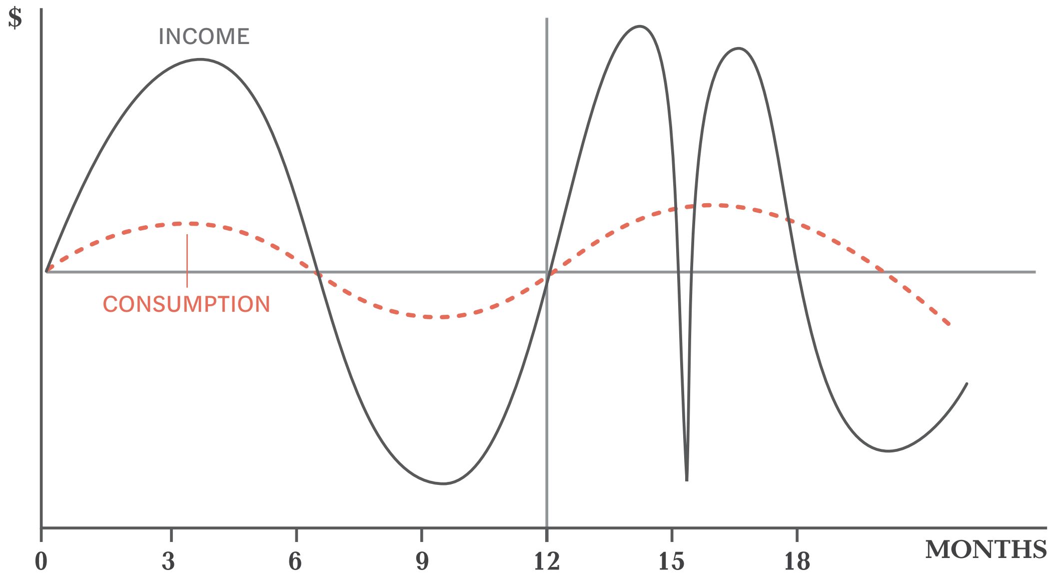
Are there reasons to be interested in a very short-term measure of welfare, capturing living standards over a period as short as one or two days? The answer is clearly negative—knowledge of an individual's poverty status during such a short reference period would be both conceptually uninteresting and of little to no use in practice. At the other extreme is lifetime welfare: a measure of individual well-being from cradle to grave. This concept is not as easy to dismiss, as there might be a conceptual interest for such a measure, but the practical difficulties of estimating lifetime living standards are likely to be unsurmountable. Between the two extremes is a continuum of potential choices. True, an "instantaneous" measure of living standards over one or two days may be uninteresting, but a short-term measure over one or two months may not. In some circumstances, such as the COVID-19 pandemic, people suddenly losing their (meager) labor incomes could quickly fall into desperation, even in rich countries, if they could not rely on sufficient savings. In such a context, access to information on poverty status even for a period as short as one month may be crucial for making the right policy choices. Scenarios requiring such granular measurement tend to be rare in practice—as DZ suggest, "on balance, and for most purposes, there is a widespread agreement that *a year* is a sensible practical compromise for the measurement of welfare" (p. 14). If that is the case, then the next question is: which among consumption and income provides the best measure of living standards *over a year*? In DZ's view, what ends up tipping the scales in favor of consumption is an essentially empirical argument, one that hinges on the notion of uncertainty.

In low-income economies, and particularly in rural areas, households are exposed to shocks that can cause a sudden decrease or increase of income or consumption: a bad harvest or job loss, an unplanned large expense due to illness, an inheritance. . . . Figure 3.1 provides a stylized description of how income (grey line) and consumption (red line) vary over time. Both variables have short-term ups and downs, but income fluctuations are found to be more frequent and severe: in other words, consumption is *smoother* over time than income is.[14]

[14] There is also a theoretical explanation of this phenomenon, known as the *permanent income hypothesis* (PIH), suggested by Friedman (1957). The idea is that people make their consumption decisions based not much on their current income, but instead on what they expect to earn in the long run. If so, then their spending will not change whenever their income changes; spending will be affected by unexpected income changes that are perceived as *permanent*, but only marginally by those that seem *temporary*. Under the assumption that most unexpected income changes are temporary, then consumption should be less volatile than income (Christiano 1987).

There is abundant empirical evidence supporting this idea.[15] The general finding of these studies is that consumption smoothing is real and significant, and that it is larger than income smoothing, although not complete (Morduch 1995, 107).[16]

FIGURE 3.1. Consumption and income fluctuations over time



SOURCE: Adapted from Zaidi (2007).

Nevertheless, "even limited smoothing gives consumption a practical advantage over income in the measurement of living standards because observing consumption over a relatively short period, even a week or two, will tell us a great deal more about annual—or even longer period—living standards than will a similar observation on income." (DZ, 14).[17]

This sums up the "smoothness argument" in favor of consumption. Although this point is regularly cited when income and consumption are compared as candidate measures of long-term living standards, it has been criticized, too. For instance, Ravallion (1994, 14) points out that, when thinking about which measure, income or consumption, is more volatile, one should think of whether fluctuations tend to be common across households or not.

---

[15] Paxson (1993) shows that this is the case for Thailand, Chaudhuri and Paxson (2002) for India, Khandker (2009) for Bangladesh, Jalan and Ravallion (1999) for rural China, Genoni (2012) for Indonesia, Deaton (1992) and Grimard (1997) for Côte d'Ivoire, Asfaw and von Braun (2004) for rural Ethiopia, Kaminski, Christiaensen, and Gilbert (2016) for Tanzania, Skoufias and Quisumbing (2005) for a pool of five countries (Bangladesh, Ethiopia, Mali, Mexico, and Russia, investigated in more detail by as many International Food Policy Research Institute [IFPRI] projects).

[16] Ravallion (1994) notes that consumption can be a "noisy" welfare indicator. In a certain context, consumption smoothing can be very limited—see, for instance, Deaton (1992) and Wagstaff (2007).

[17] Going back to the short-term (one or two months) measurement of living standards in situations of fast-paced economic change, this is a case where income may be the better indicator. Thanks to some limited savings, consumption could fall much less than income, but its level would fail to capture the true worrying condition of all those people who have insufficient assets to maintain their living standards for long. Income may capture the severity of their situation and their vulnerability better than consumption.

To understand why this is relevant, consider that the shocks that affect income and consumption can arise from either covariate risks or idiosyncratic risks. Covariate risks impact many households at the same time: uncertainties associated with harvest failure (due to droughts, floods, and other climatic events), social unrest, and policy shocks (e.g., changes in taxation, land reforms, and bans on migration) are typical examples. On the other hand, idiosyncratic risks, such as illness, shortage of agricultural inputs, death and illness of the livestock, crime and banditry, and more (Dercon 2005a), are those that affect individual households, in isolation.

Both types of risks cause income and consumption fluctuations over time, which is bad news for the analyst; but covariate shocks have one redeeming quality, that is, even though they generate intertemporal variability, they hit all households in similar ways, so that the position of households relative to each other at any given point in time is (more or less) preserved. In this scenario, the *level* of estimated long-term welfare for any one family might be off, but its *ranking* relative to others would be correct. Idiosyncratic shocks, on the other hand, generate variability over time and *also* alter the relative position of households at any one time, precisely because they are not felt uniformly across the population. In this case, both the *level* and the *ranking* of the estimated welfare of a family hit by a calamity would be off with respect to their long-term values.

There is reason to believe that, while income is more volatile than consumption over time, its variation has high covariance across households (think of a "bad season" and its impact on incomes in a rural village: certainly large, but similar across households). On the other hand, consumption may not vary as much over time, but this variation may be more idiosyncratic, related to personal circumstances (think of the need to finance a wedding, a funeral, or face a health emergency: the impact of these events on the level of consumption of a family may not be as large, but it would be isolated, and cause a re-ranking of that family in the overall distribution of living standards). If this is the case, consumption, while "smoother," would not necessarily be the better measure of long-term living standards, if welfare rankings across households are what we truly care about. Empirical evidence of this is supplied by Chauduri and Ravallion (1994).

Admittedly, the discussion over variability is just the tip of the iceberg. The question on whether to use income or consumption as a measure of individual welfare is a long-standing one that is far from settled today, and it has resulted in many difficult-to-balance lists of advantages and disadvantages of the two alternative measures. The next section takes stock of this broader debate regarding the merits of the two measures, and the likely addition of other candidates.

3.2 Further considerations on the consumption versus income debate

As the literature on the choice between income and consumption continues to grow, so do the attempts to summarize its findings, motivated in large part by a need to communicate the stakes of the choice between income and consumption effectively to practitioners, statistical officers, and policy makers around the world. As a result, different versions of a 2×2 "pros and cons" matrix comparing the two candidates have been widely used in teaching materials and technical reports, one notable example being the *World Bank Handbook of Poverty and Inequality* (Haughton and Khandker 2009, 30).

This "matrix approach" is undoubtedly effective in giving a bird's eye view of the opposing arguments the literature has brought up since the publication of the *Guidelines;* in fact, an expanded version with updated references can be found in appendix B. However, it is unlikely to help determine which measure, income or consumption, ultimately comes out on top. Not only is the relative importance of each advantage and drawback essentially arbitrary—how much is consumption's "smoothness" (third bullet in column 2, table B.1) worth, in comparison to income's "cost effectiveness" (third bullet in column 1, table B.1), for instance?[18]—but the very nature of the two measures is different. The comparison of income and consumption does not fit neatly into a "positives versus negatives" grid: a change of perspective may recast some of the listed "pros" as "cons," and vice versa. Said plainly, *the choice between consumption and income depends on the purpose of the analysis* (Atkinson 2015; 35).

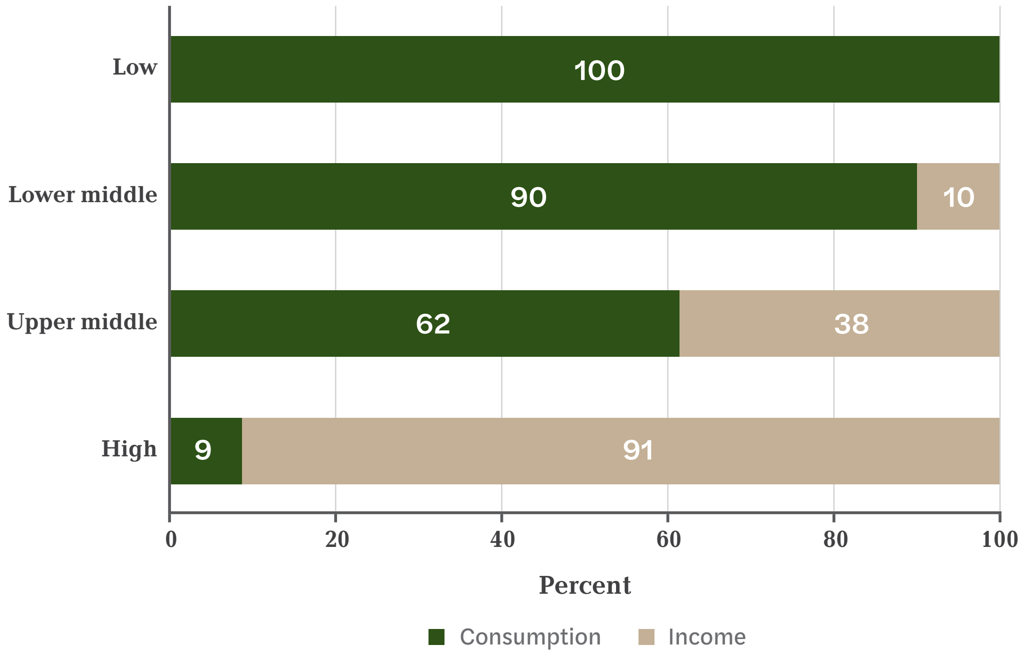
The choice of the welfare indicator can be framed as a decision on how poverty itself ought to be defined. If one believes that, in order to avoid being poor, individuals must experience a minimum *living standard* (having enough food to eat, adequate shelter, and whatever else may be considered "basic" in a given context), then consumption, which reflects the actual use of goods and services, is the natural metric. On the other hand, one may think that in order to escape poverty, individuals must have access to a minimum amount of *resources*, regardless of how these may ultimately be used. In this case, income would be a more fitting measure of welfare. This argument has been summarized as a contrast between a measure of *actual* welfare (consumption) and a measure of *potential* welfare (income) (Atkinson 2015, 35; World Bank 2015, 32; Ravallion 2016, 157). A similar reasoning applies to inequality; here the use of income may be justified by a concept of inequality that goes beyond mere achievement and incorporates other aspects of "being rich," such as the power that wealth can convey: "This power may be exercised over one's family, as with the passing on of wealth to heirs, or more generally in such ways as control of the media or influence with political parties. (...) Income is indeed a means to an end, but its reach goes much wider than consumption." (Atkinson 2015, 37).

These alternative interpretations of poverty and inequality correlate with the stage of development of a country, and the structure of its economy. A concept of well-being based on actual welfare (consumption) is well suited to contexts where material deprivation is a serious concern, and on top of that, measuring income may be completely unfeasible due to the prevalence of self-employment and informal work (Beegle et al. 2016); an idea of welfare that centers on opportunity and potential (income) is more in tune with contexts where policy may target "minimum rights" to resources (Atkinson 1989, 2019) and where inequality is a major concern. Certainly, there is also a path dependence which solidifies the preference for one or the other measure in practice: once an approach becomes established, breaking away from it (designing new surveys, collecting different data, losing comparability with

[18] The problem with assuming these implicit trade-offs between alternatives is similar to the one arising with the use of composite indices; see, for instance, Ravallion (2010); Amendola, Gabbuti, and Vecchi (2018).

past estimates) is costly. Figure 3.2 confirms that there is a divide in welfare measurement practices which correlates with a country's position in the global distribution of income.

FIGURE 3.2. Consumption versus income, by World Bank country income group



SOURCE: Authors' elaboration of the dataset presented in appendix A.

Furthermore, there are important *practical* reasons for turning to income as a complementary, or even preferred, measure of welfare. Information on total household income is useful even when the welfare indicator of choice is consumption. The analysis of household income and its sources is a key chapter of any poverty profile: a low capacity of generating income is among the drivers of poverty (e.g., Botswana 2015), and inequality (e.g., Mauritius 2019), and one can hardly imagine any investigation on the causes of poverty that does not consider income generating capacity.[19] Nor is it rare that the quality of available data thwarts the effort to construct a reliable measure of consumption. Survey estimates may be plagued by measurement error that no imputation or adjustment can entirely "fix"—for instance, the collection of food data using nonstandard measurement units may have been problematic, generating unreliable estimates for this major component of consumption. In such cases, alternative welfare indicators need to be considered, either as an instrument for cross-validation (do consumption and income tell consistent stories?) or as a candidate for replacing a poorly measured consumption aggregate.

For all these reasons, the scenarios where welfare analysts find themselves working with income data are increasingly common. Reviewing the theoretical underpinnings of income as a concept, the difficulties of matching the concept with real-life survey data, the adjustments

[19] Recommendation 4 of the Atkinson commission highlights the importance of examining the relation between consumption and income (World Bank 2017, 37).

likely needed to perfect the match, are challenging topics that lay outside the perimeter of these guidelines. The interested reader is directed to appendix C, where we provide a pragmatic strategy for constructing an income aggregate. We lean on an external source, the *Canberra Group Handbook on Household Income Statistics*, which has been the international standard for both welfare analysts and national accountants since the mid 1990s when the community of experts gathered under the aegis of the Canberra Group began to address conceptual, definitional, and practical problems that national and international statistical agencies faced in the area of household income distribution statistics (UNECE 2011).

In conclusion, the preference for consumption-based measures originally expressed by DZ can be attenuated, to the extent that the analyst is interested in measuring poverty by using context-specific standards (for instance, when a country grows rich and the scope for embracing an income-based conceptual framework increases), or because of purely pragmatic reasons (for instance, a consumption aggregate cannot be produced with the desired accuracy). On the other hand, there is a need for international organizations to analyze poverty as a *global* phenomenon, epitomized by the Sustainable Development Goals. This brings about the need for at least some degree of cross-country comparability in poverty and inequality estimates. Given the preference for consumption-based measurement within the poorest regions of the world, there is little doubt that DZ's recommendation is not to be ignored, either, if we want to keep these countries "on board" in a global perspective.

Arguably, we are heading toward a world where income-based measures, as well as other alternatives to consumption-based welfare indicators (including multidimensional and subjective poverty measures) also have a place, in what is becoming an increasingly integrated approach, where different measures complement each other, as demonstrated most notably by the recommendations of the Atkinson commission (World Bank 2017). In this forward-looking perspective, one of the priorities that come to mind is that of accounting for household assets and liabilities. In 2008, the President of the French Republic, Nicolas Sarkozy, unsatisfied with the present state of statistical information about the economy, created the "Commission on the Measurement of Economic Performance and Social Progress." As many as six Nobel laureates contributed to the production of the final report, which is often referred to as the *Stiglitz-Sen-Fitoussi Report*, and the following excerpt ("Recommendation 3"), which we take out of the 300 pages, makes the point:

> *"Income and consumption are crucial for assessing living standards, but in the end they can only be gauged in conjunction with information on wealth. A household that spends its wealth on consumption goods increases its current well-being but at the expense of its future well-being (...) we need comprehensive accounts of assets and liabilities (...)"*
>
> Stiglitz, Sen, and Fitoussi (2009, 13)

The report did not remain a dead letter. In 2013, OECD (2013) presented the results of an internationally agreed project (the *Framework for Statistics on the Distribution of Household Income, Consumption and Wealth*, or ICW Framework) to support the joint analysis of micro-level statistics on household income, consumption, and wealth. The task is demanding from the practical side and still unsettled from a conceptual viewpoint (e.g., Brandolini, Magri, and Smeeding 2010). Yet, the integration of income and household wealth with consumption information is at the top of the list of challenges facing welfare analysts in the near future (Dang, Jolliffe and Carletto, 2019). In fact, as the COVID-19 pandemic washes over the global economy, it is hard to understate the importance of collecting data needed to assess and monitor household financial fragility (Clark, Lusard, and Mitchell 2021; Demertzis, Dominguez-Jiménez, and Lusardi 2020) or resilience (Gambacorta, Rosolia, and Zanichelli 2022; McKnight and Rucci 2020).

4. Constructing the Nominal Consumption Aggregate

Economic theory lays the foundations for the welfare measure and justifies the choice of a consumption-based welfare measure (section 2)—after that, comes construction. Indeed, the first task in the computation of a consumption-based welfare measure consists in assembling the so-called nominal consumption aggregate (NCA) from survey data. The construction of the NCA is the subject of the present section (which largely overlaps with section 3 of DZ's *Guidelines*).

In section 4.1, we lay out the "first principles" that guide the process of aggregation of elementary consumption expenditures recorded by surveys. We identify four criteria that descend directly from the conceptualization of the welfare indicator as a measure of consumption. The criteria serve as a guide for the analyst's decisions on which items should be aggregated, and how. Sections 4.2—4.5 review specific recommendations for the construction of the NCA, summarizing any relevant developments in the literature and the international practice since the *Guidelines* were circulated. The discussion is organized by main components of the aggregate: food items (section 4.2), nonfood nondurable items (section 4.3), durables (section 4.4), and housing (section 4.5).

4.1 Four fundamental criteria

The NCA may be defined as *the value of all goods and services consumed by members of the household during the reference period*. Each word in this definition is important and has subtle implications for the conceptual and empirical dilemmas that often arise in practice. For instance, should we count the value of leisure time as consumption? Should funeral expenses be seen as positively contributing to well-being, and hence included in the NCA? Two households use the same vehicle, but while one car is owned, the other is provided by an employer—should they be considered equally well off? Should a household that spends nothing on housing, because they live in their own house, be considered "poorer" than a household that pays rent every month?

A few general criteria may be derived from the definition of NCA and the underlying theory to guide the analyst's decisions in situations such as those mentioned above. Ravallion (1994) and Lanjouw (2009) both compile some general principles for the construction of the consumption aggregate, and we take advantage of those contributions. Sometimes, fulfilling the criteria may not be entirely feasible, given empirical constraints, but approaching them should always be a priority when selecting next-best solutions.

We label the four criteria as follows: (1) comprehensiveness, (2) relevance, (3) typical consumption, and (4) valuation.

Criterion (1) is the *comprehensiveness* of the aggregate, given that the NCA is "the value of *all* goods and services consumed." Ravallion (1994, 13) introduces this concept under the name of "goods coverage," and writes that "consumption should cover all monetary expenditures on goods and services consumed plus the monetary value of all consumption from income in kind, such as food produced on the family farm, and the value of owner-occupied housing." Naturally, this prescription clashes with the empirical difficulty of assigning a monetary value to some items that are regularly consumed, but not exchanged on a market; publicly provided goods and services, such as policing or public infrastructure, are a notable example (other cases are discussed in section 4.3). In principle, however, the consumption aggregate should leave behind as little as possible.

Criterion (2), *relevance*, refers to the difference between consumption and expenditure, and emphasizes that what matters is the former. If a household buys a loaf of bread, the amount spent should only enter the consumption aggregate—and thus become *relevant* for welfare measurement—once the bread is actually eaten. This is because it is the *use* of goods, not their mere acquisition, that contributes to well-being. The distinction is all but philosophical: for most items, household surveys do not collect information on consumption, but on expenditure (purchase value), and the analyst faces the problem of "estimating consumption from expenditure" (World Bank 2017, 40; Atkinson 2019, 60). This is unproblematic when it comes to items for which the difference between expenditure and consumption is null or negligible in practice: for instance, most food items are perishable, and if they are purchased during the reference period, we can safely assume they are also consumed in the same period. On the other hand, there are many items for which expenditure does not approximate the value of consumption well, or at all. A typical example is that of durable goods: the expenditure incurred to acquire, say, a washing machine, does not just reflect the use (consumption) of the washing machine during the reference period, but rather its enjoyment during a much longer, multiyear time frame. Another case of divergence between expenditure and consumption is that of goods that are consumed, but never purchased, for instance, homegrown food. As we shall see in the rest of section 4, the solution to these problems involves estimating the value of consumption. The idea of relevance also helps to exclude those transactions that are, indeed, money flowing out of the household's budget, but that do *not* represent current consumption; this is the case of purchases of financial assets (which qualify as savings or investment, contributing to future rather than current consumption), or loan repayments (which may be interpreted as the financing of past consumption, at least partly).[20]

Criterion (3) prescribes that the NCA represents *typical consumption*. When we say the NCA represents welfare *during the reference period*, the underlying assumption is that what is observed in that time interval will be a good representation of the welfare *typically* enjoyed by households during a generic year. However, any empirical evidence on a household's consumption will reflect *contingent* behaviors, those that took place in that particular year or month. If a household spends a fortune on a special celebration during the survey period, such as a marriage, the resulting spike in measured consumption is genuine enough, but unrepresentative of typical living standards for that household. This argument leads to the exclusion of infrequent (or "lumpy," or "bulky") expenditures from the consumption aggregate (and the consequent dilemma of which expenditures are to be considered infrequent). The choice of excluding lumpy expenditures is not, however, entirely uncontroversial. In all likelihood, exceptional expenditures will displace other spending (that is, a household will probably cut back on some of its other expenses in order to afford the big payment). The displacement will be greater for households that are unable to draw on savings or borrow, that is, poorer families and families having to shoulder large expenses that they have not had the chance to prepare for (as in the case of a catastrophic shock). The question, then, is whether spending net of the lumpy components is *more* typical than total spending. Arguably, if there is displacement, neither of the two measures—net or total—is representative of long-run consumption; in fact, both are noisy proxies of it. Ultimately, because we do not observe long-run consumption, and we have no way to ascertain the size of the displacement of current expenditure, we cannot know for sure which of the two proxies is, in fact, the noisiest. A pragmatic strategy is to continue to exclude the shortlist of expenditures that are usually considered lumpy (e.g., weddings, funerals, purchase of durable goods), because they are typically very large with respect to the total budget of the household (and of the likely displacement they may cause), and that, at least to some extent, they were expected. The more a certain expenditure can be anticipated or planned for, the better is the case for its exclusion, as the observed consumption pattern discounts the occurrence of that expenditure.

Finally, criterion (4) is that of the suitable *valuation* of items to be included in the NCA. What is meant by the *value* of consumption, exactly? In general, consumption of a given item should be evaluated at market prices—ideally, those that the household actually faces when acquiring the item. This is a direct implication of consumer theory (section 2): when the consumer selects the quantities of goods to be consumed (provided she is able to freely choose the quantities, within her budget), her valuation of how much any given good is

[20] A subtler point that may be also filed under the issue of relevance concerns items that are chosen for reasons that are not entirely discretionary on the part of the consumer. Because the analyst's ultimate goal is to measure welfare, one may argue that consumption out of obligation, rather than choice, contributes nothing to well-being and therefore is irrelevant for the purposes of the consumption aggregate. This argument comes up for the case of the so-called *regrettable necessities*—"goods and services that yield no welfare in their own right, but that have to be purchased, for example, in order to earn income. Work clothes or transport to work are obvious examples" (DZ: 21)—together with other considerations, for example, health expenditures. It is a thorny issue, and we will come back to it in section 4.3.

worth *to her* will be compared to the prices that she faces on the market, and quantities will be chosen accordingly (Ravallion 2016, 189).[21] Thus, market prices reflect the consumer's own valuation of the benefit she will enjoy by consuming her choice of goods, which is precisely the objective of money-metric utility (MMU) as a measure of welfare. This principle is easily followed in cases where the actual purchase value of the consumed item is known. This is not the case of items that are *not* acquired through the market, such as subsidized goods and services (the price faced by the household is lower than the one set by the market), or own-produced goods (the household does not pay a price at all). In these and other cases, the analyst finds herself in the position of having to estimate a suitable price in order to place a value on reported consumption.

In the rest of this section, we discuss the implications of these four general principles for the construction of the main components of the NCA, namely food (section 4.2), nonfood nondurable items (section 4.3), durable goods (section 4.4), and housing (section 4.5).

4.2 Food items

That food is a fundamental component of living standards is a given: there is no dispute on the fact that the value of *all* food consumed during the reference period must be included in the NCA. What this means in practice is that the aggregate should include the (annualized) value of food consumed during the reference period, coming from all possible sources: (1) purchased in the marketplace (including meals purchased away from home, for consumption at or away from home); (2) produced by the household itself (food own-production is common among rural households); and (3) received in-kind (as a transfer from other households, charities, or the government, or as payment in exchange for services rendered) (DZ, 27).

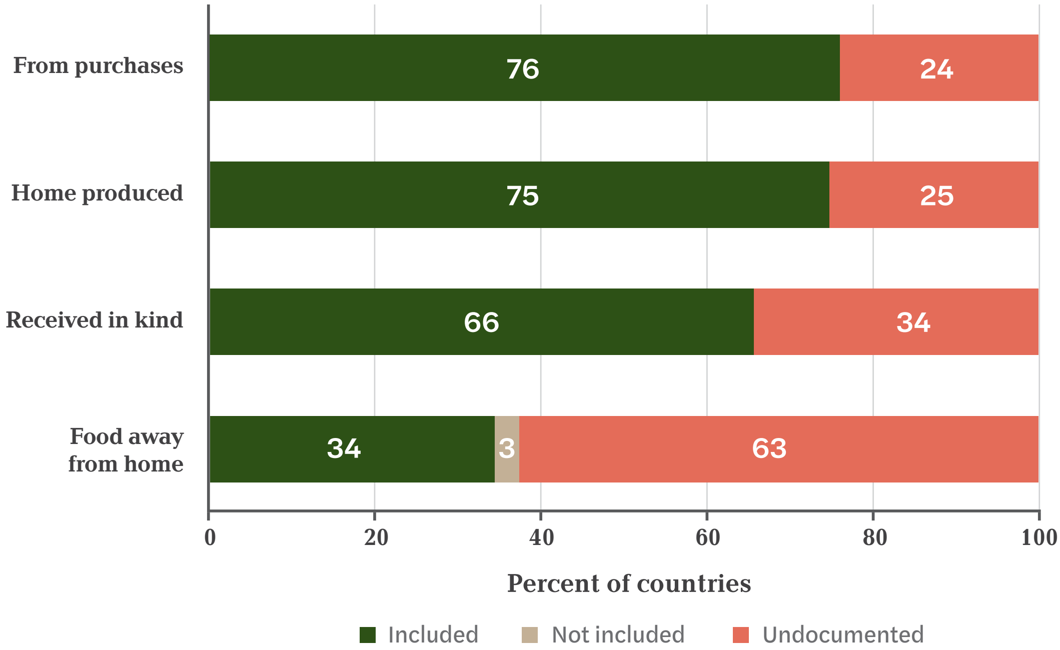
This concludes the conceptual background on the food aggregate. Difficulties in constructing it are mainly empirical in nature, and relate to the availability of all the necessary pieces of information in household survey questionnaires and the quality of the resulting data. In fact, there have been numerous efforts to improve both the quality and comparability of data on food consumption (see FAO and the World Bank 2018, and references therein). Among the topics related to welfare measurement, this is certainly one where the emphasis has shifted from data analysis to data collection during the last two decades.

Figure 4.1 shows that, for the most part, analysts are able to overcome issues of data availability and construct comprehensive food consumption aggregates from survey data (food received in-kind and food prepared away from home seem to be most problematic). But the devil is in the details, and the rest of this section discusses the many analytical challenges that stand in the way of the final aggregate.

---

[21] Technically, the ratio of marginal utilities will be equated to relative prices—this is the condition that identifies the point q^* in Figure in section 2.2.

FIGURE 4.1. The food consumption aggregate: percentage of countries that include or exclude each food source



NOTE: The graph shows the distribution of countries (in percent) for which the consumption aggregate underlying official poverty estimates includes or excludes each component (food prepared away from home is considered as a stand-alone category here, because it is often the subject of a dedicated question or module in household expenditure surveys).
SOURCE: Authors' elaboration of the dataset presented in appendix A.

4.2.1 Acquisition or consumption

One of the first issues encountered when constructing the food aggregate is the distinction between acquisition and consumption of food:

> "In some cases where food can be and is stored over long periods of time, and where the questionnaire permits it, "food consumed' can be distinguished from "food purchased." In principle, it is the value of the former that should go into the consumption aggregate" (DZ, 26).

This fulfils the second criterion for the construction of the consumption aggregate, that of *relevance* (see section 4.1): it is food eaten (consumed), not food acquired (purchased or otherwise received), that increments one's welfare.

The recommendation remains entirely valid in principle, but recent literature clarifies that, in fact, the analyst rarely has a choice at all. More often than not, it is the design of the questionnaire that dictates what goes into the food aggregate (see Appendix E, third point). Smith, Dupriez, and Troubat (2014) document that out of 100 recent household surveys from low- and middle-income countries, as many as 41 record food acquisition alone, and only a small minority record both the amount acquired and the amount consumed *for the same food items.* A common questionnaire design asks households about food *acquired* (purchased) on the market, and then about food *consumed* from own-production and transfers, with no overlap among the two categories (Conforti, Grünberger, and Troubat 2017). As a consequence, most analysts end up computing food aggregates based at least partly on food acquisition, despite this not being the recommended option in theory.

How good is food acquisition as a proxy for food consumption? The literature has not yet reached a firm conclusion. Smith, Alderman, and Aduayom (2006, 10) argue that "because most foods are perishable and consumed with high frequency, and people try to smooth their consumption of food over time, we would expect their acquisitions to match fairly well with consumption"; and when differences due to accumulating or decumulating stocks do manifest, they are likely to be randomly distributed across households, so that population averages of consumption and acquisition should still be close. Evidence from Kenya, the Philippines, and Bangladesh suggests that the average difference between available and consumed calories is less than 5 percent. Kaara and Ramasawmy (2008) and Martirosova (2008) compare estimated food consumption and acquisition for Kenya and Armenia; the first study finds average energy acquired to be about 12 percent higher on average than energy consumed, while a smaller difference is observed in food expenditures (especially for poorer households, who do not make large bulk purchases); the second study finds much smaller differences overall. On the basis of 81 recent surveys, Conforti, Grünberger, and Troubat. (2017) conclude that acquisition data yield estimated calorie intakes that are 10 to 14 percent higher on average than those obtained from consumption data.

Whatever its likely size, there are options to keep the bias in check in case one is forced to construct an acquisition-based measure of food consumption. Specifically, it is good practice to pay special attention to the presence and impact of extreme values in the distribution of both food expenditure and calorie availability, and, if necessary, exclude or impute large bulk purchases (the framework of outlier detection and treatment can help in this regard, see section 7.3).

4.2.2 Choice of the reference period

The use of two different recall periods for the same items used to be routine for Living Standards Measurement Study (LSMS) surveys (Deaton and Grosh 2000, 114). Such a design generates, of course, two competing estimates of food consumption: data were often collected using both bounded recall (value of purchases since the interviewer's last visit, usually two weeks prior) and the so-called "usual month" approach, wherein "the respondent is asked in how many months of the year the household purchased the food item, how often it purchased the item in each of those months, and how much it usually spent each time"

(Deaton and Grosh 2000, 112).[22] DZ argue that, if there is a choice, "analysts should choose the alternative that is likely to provide the most accurate estimate of annual consumption for each household, not for households on average" (DZ, 26). This justifies their preference for the usual month estimate. Since DZ, however, a growing body of research on survey methodology has produced new evidence on "what works," and questionnaire design practices in low- and middle-income countries have changed, putting this recommendation into question.

First, the use of different reference periods for the same items within a single survey does not appear to be as common as it formerly was (Smith, Dupriez, and Troubat 2014, 12): currently, analysts are hardly ever faced with the choice between alternative estimates of food consumption referring to different periods.

Second, fulfilling the wishes expressed by Deaton and Grosh (2000), the "usual month" approach has been assessed against alternative methods in experimental settings. So far, results have cast doubts on its superiority to simple recall questions. Using an experimental design, Beegle et al. (2012) document that the usual month approach led to significant underestimation of household food consumption expenditure with respect to the benchmark chosen for the experiment (a closely supervised individual diary). This was not the case for a "plain" seven-day recall question. The usual month question did not produce estimates with a smaller variability with respect to shorter recalls and was associated with the longest completion times, suggesting greater respondent burden.[23] Another experimental study by Backiny-Yetna, Steele, and Djimal (2017) finds that the usual month approach performed no better and no worse than a seven-day recall method. Gibson (2007, 24) documents the similarity between data gathered via a usual month question and simple monthly averages for the case of Vietnam, indicating that respondents probably answered with reference to the most recent rather than the "usual" month, nullifying the supposed advantage of the usual month approach in capturing seasonality. In a recent overview of the methodological literature, FAO and the World Bank (2018) state that "taken together, this evidence indicates that the usual month may be a lose-lose proposition if it is less accurate and more cumbersome to implement when compared to a seven-day recall. This is possibly the most important single development in the evidence base since the publication of Deaton and Grosh (2000)" (p. 19). Ultimately, the conclusion is that "the 'usual month' approach should not be used" (p. 51).

The upshot for the welfare analyst is that, if indeed she has a choice—which is seldom the case—estimates obtained via short recall periods (one or two weeks), coupled with either a carefully planned out sampling strategy that spreads interviews over the course of one year (as well as geographically), or with multiple household visits in different seasons, are to be preferred to usual month estimates. Appendix E returns on this issue.

---

[22] The "usual month" approach was conceived as a way to maximize the reference period (one year) so that final estimates would not be affected by seasonality, and at the same time, to minimize the recall period (one month) so that interviews would be feasible. While ultimately arguing in favor of the usual month, Deaton and Grosh (2000) acknowledged the weaknesses of the evidence available at the time in support of the approach, and advocated for a stronger support to experimental components in LSMS surveys.

[23] See also Gibson et al. (2015) and de Weerdt et al. (2016).

4.2.3 Food away from home

The shorthand "food away from home" generally indicates food *prepared* away from home, regardless of whether the food is consumed outside or inside the home.[24] This includes meals and snacks purchased from restaurants and other commercial establishments (including takeout meals), received in-kind at school, at work, from other households, and so on. This category of foods deserves special attention because it does not enter the household's pantry as individual items that are later combined into meals. Therefore, unless food away from home is deliberately accounted for on top of elementary food items, its contribution to total consumption gets overlooked. DZ explicitly point out that the value of food away from home must be included in the food consumption aggregate (DZ, 26).

Since the *Guidelines*, evidence of the changing dietary patterns in the developing world has put an even brighter spotlight on food away from home. For example, the percentage of households reporting consuming meals outside the home increased from 23 to 39 percent between 1994 and 2010 in India (Smith 2015). Meals taken at school have been found to contribute to 18 and 40 percent of daily energy intake among children and adolescents in China and Benin, respectively (Liu et al. 2015; Nago et al. 2010). As a consequence, food away from home can have a significant impact on the ranking and profiling of households according to their total consumption expenditure: in a study on Peru, Farfan, Genoni, and Vakis (2017) found that when food away from home is included in the consumption aggregate, 41 percent of individuals change their relative ranking, as measured by consumption decile; among these re-rankings, about one-half are across the poverty line; the poverty profile is altered, as is the poverty line itself, given that the food aggregate is the basis for computing and costing calorie intakes.[25]

Meanwhile, new methodological research has pointed to the inadequacy of standard questionnaire designs in capturing this increasingly important component of food consumption. With a test of different questionnaire designs in Vietnam, Farfan et al. (2019) find that recording food away from home using a single question underestimates consumption of this category of food by about 33 percent with respect to the experimental benchmark (a supervised individual food away from home diary). Yet, the one-line item approach (asking about the total value of all "outside meals" consumed during the reference period by all household members) is the single most common design adopted by recent surveys in low- and middle-income countries, as documented by Smith, Dupriez, and Troubat (2014).

Overall, new evidence reinforces DZ's prescription of including the value of food away from home in the food aggregate, although the shortcomings of standard questionnaire designs often get in the analyst's way. The best course of action is to include whatever is available, acknowledge any faults, and continue to steer statistical institutions toward the inclusion of dedicated food away from home modules in household expenditure and consumption surveys. Appendix E returns on this issue.

---

[24] "Food away from home" may also be taken to mean food *consumed* away from home, irrespective of the origin of the food. While there is no universally accepted definition, there is a general preference toward defining food away from home based on the place of *preparation* (FAO and the World Bank 2018, 36).

[25] See also Borlizzi, Delgrossi, and Cafiero (2017) on Brazil.

4.2.4 Own-production and food received in kind

The consumption of food acquired through channels other than the market is commonplace in many countries, particularly in the developing world. First, there is food own-production: this includes the products usually sold as part of a household's main commercial enterprise, that are instead reserved for consumption (farm households are the typical example), but also noncommercial and subsistence food production, as well as food acquired from hunting, fishing, and gathering. On the other hand, there is food received in kind, in the form of transfers from other households, the private sector, or the government.

In the eyes of the welfare analyst, food items that are own-produced and received in kind share a fundamental characteristic: neither is obtained in exchange for a market price. This implies that their monetary value must be estimated, which boils down to identifying suitable prices for each item, in accordance with the *valuation* criterion (section 4.1). One could see this task as an imputation exercise. The price of any given own-produced (or received) food item is unobservable; if, ideally, an identical good is bought and sold on the market, then the price that prevails in that circumstance should be an adequate proxy for the "missing" one.

The *Guidelines* state that, in principle, when estimating the value of food own-production, farm-gate prices, defined as what households (farmers) would obtain in exchange for their own-produced goods at the location of farm, should be preferred to the price of similar items traded in the market place. This is because the latter prices include transport and distribution costs, and goods that are marketed off-farm might be qualitatively different (DZ, 20, 29). In practice, however, it is rare for the analyst to have access to comprehensive estimates of actual farm-gate prices. While surveys of farm-gate or producer prices do exist, their collection and use are highly problematic, as documented by a recent World Food Programme (WFP) initiative involving El Salvador, Ghana, and Tanzania (Musumeci 2016); in fact, we are not aware of any recent poverty assessments that rely on them to estimate the value of farm households' own production. Most commonly, the information comes from within the household consumption and expenditure survey itself in two forms: self-reported valuations, and unit values from food purchases.

The first scenario is one where the questionnaire asks respondents to state the amount they would expect to receive (pay) if they were to sell (buy) the food from own-production or in-kind receipts that they consumed. The analyst may readily add these self-reported valuations to the food consumption aggregate.

Alternatively, if the food purchases section of the questionnaire allows it, the analyst may compute *unit values* for each food item, defined as the ratio between the amount paid to purchase a given quantity, and the quantity itself. Unit values may then be used to price the quantities of food items that were own-produced and received in kind. The use of unit values as a proxy for prices is a long-standing and much-debated topic—see Prais and Houthakker (1955,113—114) for a seminal discussion, and section 5.2.1 for further details. In this context, knowing that the use of unit values presents a number of empirical complications is enough.

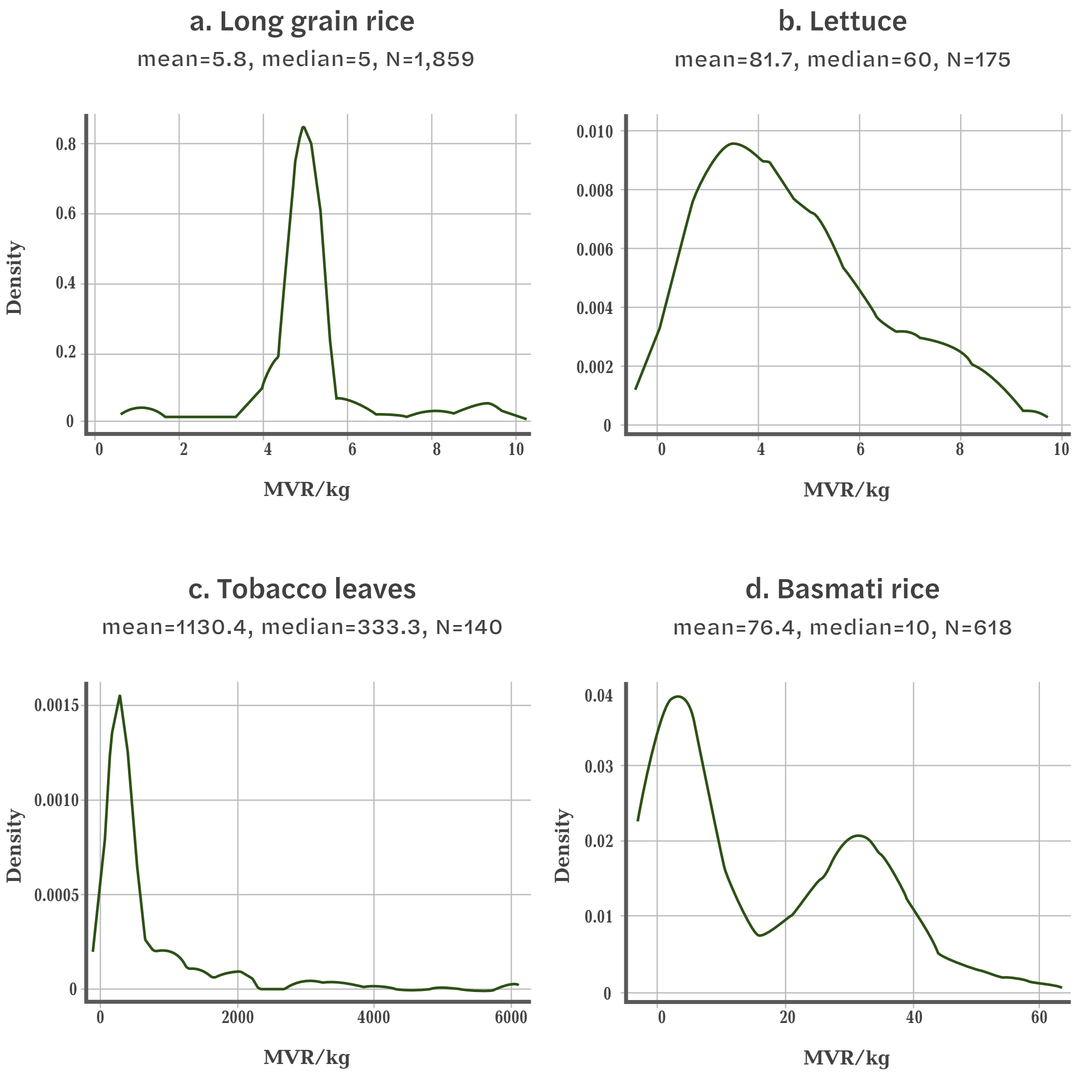
Say the household whose in-kind or own-produced food consumption we need to price is our "target," and the item we are interested in pricing is figs. The question is: how to choose the reference population over which unit values of figs are computed? The *Guidelines* advise in favor of what may be called a "hierarchical" approach (DZ, 30).

First, the analyst computes median unit values from fig purchases reported by households in the same cluster, or primary sampling unit (PSU), as the target household (medians being more robust than means against outliers). If the PSU contains enough transactions, say some 50 or more, then the PSU-level median unit value is used to price figs that were own-produced or received by the target household. Otherwise, the analyst moves up to the next administrative level and computes median unit values in the same subregion (province, district, governorate, or any "fine" territorial unit available in the dataset) as the target household. Once again, if enough observations (50 or more) are found at this level, then the subregion median unit value is used as a proxy for the price of figs in the target household. Otherwise we move one more level up, to the region, and so on, broadening the set of households used for the calculation of the median values of figs, all the way to national medians, if necessary. Equation (4.1) provides a concise description of this algorithm, assuming there are three subnational administrative levels available (say PSU, subregion, and region); j denotes any commodity (figs in our example), h denotes the target household (whose own-production or in-kind receipts we need to price), uv_j^h denotes the unit value of figs assigned to the target household, and E(·) is the expected value operator (or, equivalently, the median):

$$uv_j^h = \begin{cases} uv_j^{L3} = E\left(uv_J \,\middle|\, \text{administrative level 3}\right) & \text{if } uv_j^{L3} \text{ can be computed} \\ uv_j^{L2} = E\left(uv_J \,\middle|\, \text{administrative level 2}\right) & \text{if } uv_j^{L3} \text{ cannot be computed} \\ uv_j^{L1} = E\left(uv_J \,\middle|\, \text{administrative level 1}\right) & \text{if } uv_j^{L2} \text{ cannot be computed} \\ uv_j^{C} = E\left(uv_J \,\middle|\, \text{Country}\right) & \text{if } uv_j^{L1} \text{ cannot be computed} \end{cases} \tag{4.1}$$

The reasoning behind the strategy described in equation (4.1) is that, usually, the value of foods consumed by the target household is better approximated by market transactions that took place in its vicinity, where items of similar quality are likely to be traded. This is not always the case, however, and exceptions risk introducing severe over- or undervaluation (as in the case of the water fetched from a well, valued at the price of the Perrier bottle purchased nearby, or of mangoes that fall off the tree in rural areas valued at prices from an international supermarket in the capital). Small sample sizes at the first stages of the imputation process may also be problematic if unit values are "noisy." Problems can also arise when the list of food items used to collect the data is not detailed or specific enough (e.g., the questionnaire collects data on the consumption of "rice," rather than its individual varieties). In general, the analyst should exercise great care while performing such imputations. One way to monitor the reliability of computed unit values is to plot their distribution, possibly within clusters or regions, to check for anomalies: figure 4.2 shows a few examples. A unimodal, symmetrical, low-variance distribution (panel a) is reassuring; a higher variance distribution may indicate a worrying degree of variability in the quality of the underlying items (panel b), or the presence of outliers (panel c); multimodal distributions often signal gross mistakes, such as misreported units of measurement (eggs measured in units for some households and in dozens for others, or rice measured in kilograms for some households and in bags for others), or different foods that have been lumped together under the same item code (panel d).

FIGURE 4.2. Empirical distributions of unit values for selected food items, Maldives (2016)



NOTE: MVR stands for Maldivian Rufiyaa. All unit values are computed as expenditure per kilogram purchased. Distributions are at the national level. All distributions are trimmed at the top 95th percentile to facilitate the reading of the graphs.
SOURCE: Authors' calculations using the 2016 Maldives Household Income and Expenditure Survey.

Leaving aside issues of implementation, which is the better proxy for the price of nonmarketed food items, between self-reported valuations and unit values from purchases? The *Guidelines* note that, when it comes to own-production, "households' own valuation (...) are likely to be a much better approximation of the true 'farm-gate' value, rather than estimates derived using prevailing market prices from the food purchases section." (DZ, 29). In fact, market prices may be wholly unrepresentative of the value of own-produced food, not only because they include excess costs, but also because of differences in quality—and this applies to in-kind receipts, as well. A market for the same exact items that households produce or receive free of charge may simply not exist. That said, self-reported valuations rely on the existence of market exchanges, too: respondents are asked to guess what would happen if they made a transaction that, in some cases, may be entirely hypothetical, which would result in low-quality data. As Deaton and Grosh (2000) note, the imputation of the value of nonmarket transactions (as any other imputation exercise) "is likely to work best where there is relatively little need for it. (...) Where these markets do not exist, analysts are in effect imposing an accounting framework on the physical data, a framework of dubious relevance to the lives of the people being studied." (p. 117).

The discussion in this section suggests that, when measuring welfare, there is no way around the task of imputing the value of food own-production and in-kind receipts. When self-reported valuations are available, they tend to be preferable to unit values from food purchases—this is not a universal recommendation, as local survey-specific circumstances might suggest a different course of action. Either way, the higher the share of nonpurchased food items in total household consumption, the more care is to be exercised when performing imputations, by checking the reliability of estimated values, with sensitivity analysis being the standard way to proceed.

4.2.5 Food rations

Food rations—the provision of quotas of food items for free or at below-market price—are a type of in-kind transfer of food.[26] Among the largest programs providing food rations in the world is the Indian Targeted Public Distribution System (TPDS). The TPDS targets nearly 800 million people, providing subsidized grain through a network of more than 500,000 fair price shops across the country (Bhattacharya, Faleao, and Puri 2017). Iraq's Public Distribution System (PDS) is of similarly gigantic scale: in 2012, it provided all Iraqi households with food items that account for between two-thirds and three-fourths of total calorie consumption for the poor and the bottom 40 percent (World Bank 2014, 177). Araar, Choueiri, and Verme (2015) document the generosity of food subsidies in Libya: "a family of four is entitled to the following quotas at subsidized prices each month: 8 kg of sugar, 800 gr. of tea, 4 kg of tomato paste, 6 liters of vegetable oil, 10 kg of rice, 12 kg of flour, 4 kg of semolina, and 6 kg of pasta. These quantities (...) can cover well above the total amount of calories necessary for a family of four for a period of one month." (p. 5). Verme and Araar (2017) analyze similar PDSs for seven countries in the Middle East North Africa region.

Whether rations are distributed for free or at subsidized, mandated prices, the problem they pose to welfare measurement is the same: the amount paid by recipients does not represent the benefit from consuming the ration. In the absence of any adjustments to the recorded value of the ration, two serious mistakes would be made. First, the *level* of estimated living standards would be wrong: a household that, thanks to a free food ration moves beyond the food poverty line, would not appear to have improved its condition at all, given that the recorded value of the ration is zero. Second, unless rations are universal (i.e., received by all households in the country), the *ranking* between households would also be wrong: if two

[26] DZ do not mention rations specifically in their recommendations; we choose to discuss the topic separately because of some additional methodological difficulties with respect to "ordinary" food in-kind receipts, and for the relevance of ration programs in some areas of the world, particularly those that are conflict ridden or otherwise suffering food shortages.

households consume an identical diet, but one of them obtains its food via a ration and the other purchases everything at market prices, they would not be ranked as equally well-off, as they should—the former household would appear poorer (Hentschel and Lanjouw 2000).[27]

The issue faced by the analyst is that of finding a price that adequately represents the marginal utility of the ration for the consumer in order to estimate the value of the ration and incorporate it in the food aggregate—in other words, to re-price the ration in accordance with the *valuation* criterion (section 4.1). If *official prices* for ration items exist, one may be tempted to use those: they are prices, after all. However, because these prices are often heavily subsidized, valuation of food rations by means of official prices "would artificially suppress the value of food expenditures stemming from rations" (Iraq 2014a, 9), and is not a recommendable practice. A second possibility is to exploit the information generated by a secondary market for rations. If indeed such a market exists, and the number of transactions recorded in the diary/questionnaire is large enough, then market-equivalent prices can be estimated by calculating *unit values* (the ratio between the proceeds from selling food rations and their quantities). If either of these conditions is unmet, however, unit values cease to be a reliable measure for the value of ration items. This is the case of Iraq, for instance, where less than 2 percent of households report purchases of rice (the most important item in the ration bundle), and less than 0.5 percent for the other items. A third possibility is to identify commodities that are close substitutes to rations and are traded in the market. For example, to the extent that rice received as part of rations is close enough to some variety of rice traded on the market, one can use the price of the latter to proxy the price of the former. A common situation, however, is that significant differences exist between commodities distributed as part of rations and food items found in the market, so that this option may not be of practical help. A fourth possibility is to asks the households' opinion on how much they would pay for ration-equivalent items in the market. If the questionnaire contains such a question, then in principle one could use *self-reported assessments*. The Iraqi experience casts doubts on the accuracy of this type of answer. This is unsurprising: if the secondary market for ration items is thin, few households would be informed, high item-nonresponse would be found in the data, and households that do respond may report inaccurate values. The fifth possibility is a last-resort option: the use of *expert judgment*. In the case of Iraq, "enumerators approached the local ration agent in the cluster, in a manner akin to a price survey. However, there were variations in these prices that may reflect uncertainty, noise and local variations in supply, demand and quality." (p. 10). What ended up being used to value ration items was national median values of prices reported by ration agents.

These five approaches—set out in order of preference, with the exception of the "official prices" route, which is not advisable—exhaust the possibilities at the analyst's disposal when the value of consumed rations must be included in the food aggregate.

[27] The case of rations targeted to specific population subgroups, e.g., to households in a specific region or with income below a certain threshold, will be discussed in detail in section 4.3.

4.3 Nonfood nondurable items

The computation of the nonfood component of the NCA is more involved than that of the food aggregate. With few exceptions, household surveys record nonfood expenditures, which, depending on the item, may be very far from what the analyst is really after, which is consumption.

A convenient way to organize the discussion is to represent the task of constructing the nonfood nondurable aggregate as a two-step procedure. Step 1 consists in *identifying* all elementary household expenditures on nonfood and nondurable commodities and services recorded in the questionnaire, which are typically scattered through different modules. Health expenditures are often collected in a dedicated module, as are housing expenditures, while education expenditures are often collected both at the household and the individual level. Making this sort of inventory helps locate expenditures that may be "hidden" in sections that are not focused on expenditure per se (e.g., in-kind receipts as payment for services rendered may be recorded in the employment section), and keeps potential instances of double-counting under control. Step 2 consists in *selecting* items for inclusion in the NCA: for each and every item in the list drawn in step 1, the analyst must decide whether that particular expenditure is a good proxy for the value of consumption.

The two steps (identification and selection) are facilitated by referencing the *Classification of Individual Consumption According to Purpose* (COICOP), the international reference classification of household expenditures (United Nations, 2018),[28] a strategy already explored in ECASTD (2016) in an effort to harmonize the construction of the NCA for 29 countries in the Europe and Central Asia (ECA) region. The COICOP system has a hierarchical structure, articulated in levels: the highest level, the *division*, is denoted by two digits (e.g., 01 denotes "food and non-alcoholic beverages," 02 is for "alcoholic beverages, tobacco, and narcotics," 03 is for "clothing and footwear," and so on, until category 13, "personal care, social protection, and miscellaneous goods").[29] One level below is the *group*, denoted by three digits, shown in table 4.1.[30] The table shows how the three-digit COICOP classification can serve as a checklist during the process of constructing the consumption aggregate, as the analyst is called to decide whether to include or exclude candidate expenditures. The recommended choice is shown in the last column of table 4.1. The decision is straightforward in some cases, less so in others: the rest of this section discusses the details, group by group.

[28] The first classification under the name COICOP was adopted by the United Nations Statistical Commission in March 1999. In 2018, the commission endorsed a revised version, "COICOP 2018," which we use in table 4.1.

[29] In fact, COICOP 2018 has 15 two-digit categories. We ignore divisions 14 and 15, as they refer to expenditures of nonprofit institutions and expenditures of general government, respectively.

[30] The 2018 revision includes two additional layers (denoted by four and five digits) that classify commodities into finer and finer categories, but the three-digit classification is detailed enough for our purposes in this section.

TABLE 4.1. The COICOP System as a Checklist for the Construction of the Consumption Aggregate

COICOP	Description	Include in NCA?
01.1	Food	Y
01.2	Non-alcoholic beverages	Y
01.3	Services for food and non-alcoholic beverages processing	N
02.1	Alcoholic beverages	Y
02.2	Alcohol production services	N
02.3	Tobacco	Y
02.4	Narcotics	Y
03.1	Clothing	Y
03.2	Footwear	Y
04.1	Actual rentals for housing	Y
04.2	Imputed rentals for housing	Y
04.3	Maintenance, repair and security of the dwelling	S
04.4	Water supply and miscellaneous services relating to the dwelling	Y
04.5	Electricity, gas and other fuels	Y
05.1	Furniture, furnishings, and loose carpets	Y
05.2	Household textiles	Y
05.3	Household appliances	S
05.4	Glassware, tableware and household utensils	Y
05.5	Tools and equipment for house and garden	Y
05.6	Goods and services for routine household maintenance	Y
06.1	Medicines and health products	Y
06.2	Outpatient care services	Y
06.3	Inpatient care services	Y
06.4	Other health services	Y
07.1	Purchase of vehicles	N
07.2	Operation of personal transport equipment	Y
07.3	Passenger transport services	Y
07.4	Transport services of goods	Y
08.1	Information and communication equipment	S
08.2	Software (excluding games)	Y
08.3	Information and communication services	Y
09.1	Recreation durables	N
09.2	Other recreational goods	S
09.3	Gardens and pets	Y
09.4	Recreational services	Y
09.5	Cultural goods	S
09.6	Cultural services	Y
09.7	Newspapers, books and stationery	Y
09.8	Package holidays	Y
10.1	Early childhood and primary education	Y
10.2	Secondary education	Y
10.3	Post-secondary non-tertiary education	Y
10.4	Tertiary education	Y
10.5	Education not defined by level	Y
11.1	Food and beverage serving services	Y
11.2	Accommodation services	Y
12.1	Insurance	Y
12.2	Financial services	N
13.1	Personal care	Y
13.2	Personal effects n.e.c.	S
13.3	Social protection	Y
13.9	Other services n.e.c.	S

NOTE: Y = yes, include in the CA; N = no, exclude from the CA; S = some of the items in this category are to be included, some are not; n.e.c. = not elsewhere classified.
SOURCE: Our elaboration on United Nations (2018: VIII, p. 29).

01 Food and non-alcoholic beverages. The treatment of food expenditures is discussed in section 4.2 of this document. Table 4.1 summarizes the operational recommendation: all food expenditures should be included in the NCA, with the exception of "Services for food and non-alcoholic beverages processing," which are essentially production expenditures in a family enterprise or in the process of producing food and alcohol for the household's own consumption. As such, they are already incorporated in the value of the final product, and their inclusion would lead to double-counting.

02 Alcoholic beverages, tobacco and narcotics. The general recommendation is to include these items in the NCA. There may be an argument against this choice, given that many of these items are "bad for you." That is, why should the consumption of something that is harmful to one's health (and, possibly, to society in general) count toward measured well-being? The case of qat (also spelled "khat") can be used to illustrate the point. Qat is a popular hallucinogen, and in a number of African countries and in the Arabian Peninsula, qat is consumed at parties where people gather and hold conversations while smoking cigarettes and drinking tea and soft drinks. According to the World Health Organization (WHO), however, chewing qat "produces significant toxic effects, including increased blood pressure, tachycardia, insomnia, anorexia, constipation, a sense of general malaise, irritability, reactive depression, migraine and impaired sexual potency in men. Khat is believed to be dependence-producing" (WHO 2003, 18). The inclusion of expenditure for qat in the NCA is not a minor detail. In Yemen, qat accounts for 6 percent of gross domestic product (GDP), 10 percent of private household consumption, and 33 percent of agricultural GDP, and provides employment for one in every seven working Yemeni (Yemen 2007a, 43). The motivation for including it rests on the theory discussed in section 2. The whole premise of a utility-consistent approach to welfare measurement is that it is *not* paternalistic: individuals make their own choices, maximizing their own utility, and are assumed to know better.[31] Ravallion (2016, 189) counts "avoiding paternalism," that is, respecting people's revealed preferences, as a fundamental principle of welfare analysis: "If a person chooses freely to spend some of a meager income on something that is not found on some external observer's favored list, then respect for that person demands we question that list. My prior is that the person concerned is in a better position than anyone else to know what she needs." Based on this argument, expenditures in COICOP division 02, no matter whether aimed at purchasing "goods" or "bads," should be included in the consumption aggregate.[32]

03 Clothing and footwear. These expenditures are unproblematic. It is worth noting that in the System of National Accounts (SNA), many of these items (clothing materials, garments and accessories, footwear and related services including cleaning, repair, and hire) are classified as "semi-durable goods," a distinction also used in the COICOP system. Such a distinction is usually ignored when constructing the consumption aggregate; the international practice takes clothing and footwear as nondurable goods, and the corresponding expenditure simply needs to be annualized and included in the NCA.

---

[31] Note that Yemeni are well aware of the issues mentioned in the WHO's report. In a recent survey, more than 70 percent of the respondents describe qat chewing as a "bad habit" that is also bad for the economy and bad for the nation's image. Users want to "kick the habit" but they cannot. (Yemen 2007b, 17).

[32] Milanovic (2008) provides further details on how welfare analysts should think of qat (taken as a paradigm of any "bad," under the presumption that each country has its own qat-equivalent).

04 Housing, water, electricity, gas, and other fuels. This category includes goods and services for the use of the dwelling, its maintenance and repair, the supply of water and other miscellaneous services related to the dwelling, and energy used for heating or cooling. Regarding rent, both actual and imputed rent (codes 04.1 and 04.2 in table 4.1), are discussed in detail in section 4.5 because of the empirical complexities of including them in the NCA. COICOP group 04.3 (maintenance, repair, and security of the dwelling) only includes expenditures on materials and services for *minor* maintenance and repair, while *major* maintenance and repair do not even appear in the COICOP classification, given that "only expenditures which tenants and owner-occupiers incur on materials and services for interior decoration, minor maintenance and repair, that would normally be seen as the responsibility of a tenant, are part of individual consumption expenditure of households." (UN 2018, 81). This distinction wholly applies to the NCA, as well: expenditures for minor repairs and maintenance should be included, while expenditures for major repairs should not. The rationale for this recommendation is provided by two of the four criteria discussed in section 4.1: expenditures for major maintenance can be interpreted as investment (increasing the value of the property) rather than consumption, and they are often lumpy expenditures.

Expenditures on utilities—water, electricity, gas, and other fuels—are a straightforward inclusion in principle but can be problematic in practice. A recurrent problem is the fact that markets for utilities are often subsidized, rationed or subject to pricing schemes that vary according to the amount consumed (e.g., the rate increases as customer usage increases), or other factors. The consequence of households facing different prices is that expenditure ceases to work as a proxy for consumption. If households living in a given region have access to electricity at a subsidized price, then their expenditure will command more electricity than other households who purchase it at market prices. In fact, any welfare comparisons based on unadjusted expenditures will be biased. Hentschel and Lanjouw (2000) is a most useful read, as it illustrates a number of adjustment methods: repricing is a necessary step before including any subsidized or otherwise nonmarket priced consumption of utilities in the NCA.[33]

05 Furnishings, household equipment and routine household maintenance. According to the COICOP system, this group "covers a wide range of products for the equipment of the house or dwelling and the household durables, semi-durables and nondurables, as well as some kind of household services. It includes all kinds of furniture and lighting equipment (05.1), household textiles (05.2), major and smaller electric household appliances (05.3), glassware, tableware, and household utensils (05.4), tools and equipment for house and garden (05.5), and goods for routine household maintenance (05.6.1). (...) It also includes repair, installation and rental services. (...) Domestic services by paid staff in private service, supplied by enterprises or self-employed persons are included, as well as window cleaning and disinfecting services, dry-cleaning and laundering of household textiles and carpets (05.6.2)" (UN 2018, 89). In general, these expenditures simply need to be annualized, aggregated, and included in the CA. In practice, according to the criteria in section 4.1, there is often the need to exclude selected items, mainly because they qualify as "lumpy" expenditures.

[33] It should be noted that the complexity of the adjustments needed in these cases has worked against their regular inclusion in applied work—the database presented in appendix A contains no documented examples of such adjustments being made.

The identification of lumpy expenditures is inevitably discretional, due to the fact that a given expenditure can be judged lumpy only in relative terms, that is, only when its size is judged "large" compared to typical consumption. In this group, major household appliances have the potential to be considered lumpy, and in fact some of them qualify as durable goods whose purchase value should always be excluded from the NCA (see section 4.4 of this document).

06 Health. Whether to include or exclude health expenditures is one of the most controversial decisions among those listed in table 4.1. For this reason, we postpone our discussion to section 4.3.1, where we clarify the nature of the problem (why are health expenditures problematic?) and provide an updated picture of the current practice.

07 Transport. Transport-related expenditures fall into four main categories, none of which turns out to be problematic from the standpoint of the construction of the NCA. First is the purchase of vehicles (group 07.1). Because vehicles are durable goods, their purchase value should be excluded from the NCA (this is explained in section 4.4 of this document). All other categories of transport expenditure (operation of personal transport equipment, passenger transport services, and transport services of goods) should be included in the NCA.

08 Information and communication. Most expenditures belonging to this category should be included in the NCA. There might be a few exceptions, typically for items in group 08.1 (information and communication equipment), either because they qualify as durable goods, or because the corresponding expenditures tend to be lumpy. In the absence of specific guidelines, the four criteria from section 4.1 of this document help to make the appropriate decisions.

09 Recreation, sport and culture. Major recreation durables (e.g., camper vans, boats, yachts, and the like) should be excluded from the NCA, following the same argument that applies to all durable goods (see section 4.4). All other expenditures, as long as they are typical and not clashing with the criteria discussed in section 4.1, should be included. Some discretion on the part of the analyst is required.

10 Education services. The general recommendation is to include education expenditures in the NCA. There are, however, two theoretical arguments to the contrary. DZ note, for example, that education expenditures "(...) are located at a particular point in the life cycle, so that, even if all households paid the same for education and had the same number of children, some *(those with kids attending school at the time of the survey)* would appear better-off than others simply by virtue of their age." (pp. 31—32, our italics). Counteracting this effect is hardly feasible, as the analyst would need to spread each household's education spending across the life cycle of its members, and cross-sectional surveys simply do not allow for such an adjustment. A second argument for excluding education-related expenditures is that education qualifies as an investment, not as consumption, and as such it should be included in savings, not in the consumption aggregate. This point is debated in a few countries, for purposes other than poverty measurement (e.g., MacLeod and Urquiola 2018): as a matter of fact, our review of the international practice (see appendix A) has turned up no exceptions to the rule of including education expenditure in the NCA: *all* countries do so. This follows

DZ's recommendation: "we follow standard national income accounting practice and recommend that educational expenditures be included in the consumption aggregate" (p. 34).[34]

11 Restaurants and Accommodation Services. Both group 11.1 (food and beverage services provided by restaurants, cafés and other facilities) and 11.2 (accommodation services for visitors and other travelers away from their permanent principal or secondary residence) represent fairly typical consumption, and their inclusion in the NCA is not debated. Note that ready-made meals are part of the food group, and are also included.

12 Insurance and financial services. Insurance services (12.1), irrespective of the type of insurance, should be included in the NCA: they represent consumption of a service, which is routine, discretionary, and welfare-enhancing. In contrast, because fees for financial services, repayments of debt, interest payments, and so forth, are related to the purchase and management of financial assets, which are savings (or investment), they are usually not included in the NCA.

13 Personal care, social protection and miscellaneous goods. In general, expenditures for personal care (group 13.1, including small electric appliances for personal care, hairdressing, etc.) should be included in the NCA. Group 13.2, "personal effects not elsewhere classified," is heterogeneous and includes a variety of items whose inclusion can be more questionable. For instance, expenditures on jewelry and watches are arguably to be excluded—they are lumpy, presumably occasional, and might be interpreted as investment rather than consumption. The group of expenditures related to devotional articles contains items that range from votive candles, likely less expensive, to much more expensive coffins and gravestones. The four criteria in section 4.1 should help to guide the most minute choices. The concern for heterogeneity also applies to expenditures in group 13.9. The group labeled "social protection" (group 13.3) covers nonmedical assistance and support services provided to elderly and disabled persons and other categories in need, as well as families and children (retirement homes, rehabilitation centers, child-minding, daycare, etc.): these expenditures should be included in the NCA, paying mind not to double-count any health expenditures if they are also included.

While the COICOP system (as in table 4.1) represents a useful tool for organizing the construction of the nonfood nondurable aggregates, there are a few more general considerations that deserve to be mentioned before concluding this section.

First, any expenditures for *taxes and levies* should be excluded from the NCA because these are not consumption expenditures, but rather a deduction from income (DZ, 31). Local taxes that are closely linked with services received are a potential exception to the rule (they may be seen as payment for a service received by the government, where more taxes paid translate to more services received).

[34] Note that the presence of a universal (and affordable) public education system does not affect the general recommendation: the decision to spend additional resources on education (e.g., private school) would be based on quality considerations, strengthening the case for including education expenditures in the consumption aggregate. Oseni et al. (2018) provide an updated discussion of the best practices for collecting information on education expenditure in household surveys.

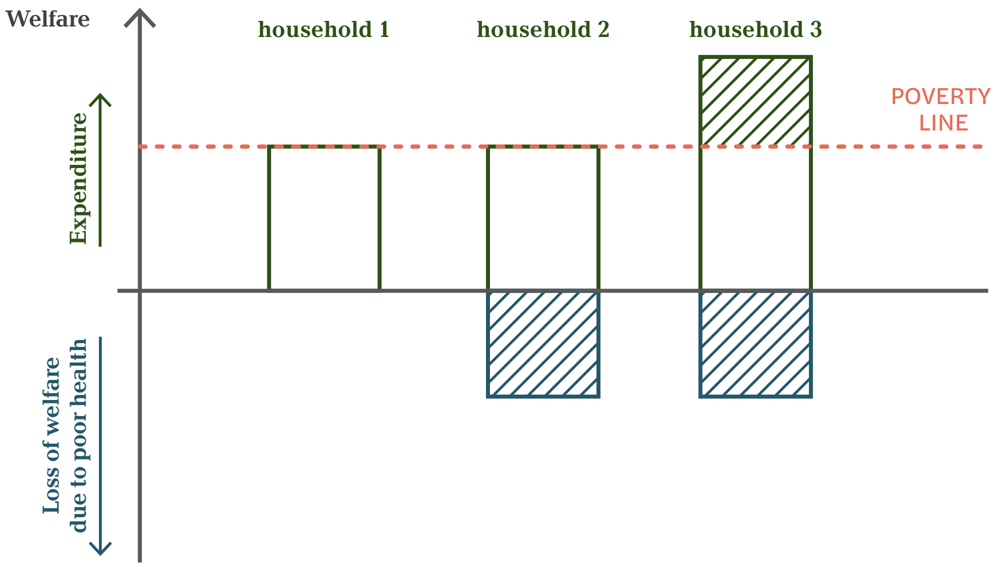
Second, there is the case of the so-called *regrettable necessities*, expenditures that go toward items that do not yield any direct satisfaction (utility) to the consumer, but have to be made regardless. A classic example are expenditures related to commuting (i.e., travelling regularly by train, scooter, car, or bus between one's residence and place of work). DZ recommend to include all regrettable outlays in the NCA because, while in theory their exclusion is justified (if they do not arise from a discretional consumption choice, but are forced by the circumstances, they are not welfare-enhancing), it is close to impossible to determine a priori which expenditures are "regrettable." The proposed solution is clearly a compromise and is open to debate—a long-standing one, in fact. Interested readers may benefit from revisiting the elegant discussion in Nordhaus and Tobin (1973).

The last issue worth mentioning concerns *gifts*, remittances, transfers to other households, and charitable contributions. On the one hand, a household choosing to employ its resources by giving them away is maximizing its own utility, which can certainly include concerns about the well-being of others: in this sense, gifts given may be considered consumption. However, DZ cite a simple and convincing counter-argument: for the household on the receiving end of the transfer, the gift is *also* consumption, because it can either be enjoyed (if it is in-kind) or used to purchase other commodities (if it is cash). This would amount to counting the value of the good as consumption twice, both for the giver and for the receiver, which is obviously not desirable. A paradoxical scenario extending this argument is one where a cash amount is passed around to each household in the sample, in turn given and received, forming a kind of "money pump" in measured welfare that is entirely the result of double-counting. The recommendation of excluding gifts given from the NCA is consistent with the *Canberra Group Handbook on Household Income Statistics* published in 2011. According to the international experts, known as Canberra Group, "current transfers of cash, goods and services to other households such as gifts, remittances, alimony, child support, etc." are excluded from the definition of household consumption expenditure (UNECE 2011, 19). In conclusion, DZ's recommendations stand.

4.3.1 Health expenditures

The inclusion of health expenditures in the nominal consumption aggregate is among the most debated choices among those reviewed in this section. What is the debate about? One of the main points of contention, as the *Guidelines* put it (DZ 2002, 33), is that we can observe and measure the increment in welfare due to receiving (consuming) health care, but not the loss in welfare determined by the need for care, arising from a deterioration of one's health status. This asymmetry seems especially troubling when we consider welfare comparisons. Figure 4.3 illustrates.

FIGURE 4.3. The problem with health expenditures and welfare measurement



SOURCE: Authors' elaboration.

In figure 4.3, the space of *measured* welfare, represented by expenditure, is above the horizontal axis. We use a vertical bar to represent a household's consumption aggregate—the taller the bar, the higher the expenditure, and the better off the household. Below the horizontal axis is what we cannot observe and measure: the loss of welfare due to a compromised health status. Pictured are three households: household 1, whose consumption aggregate is right at the poverty line (the bar above the horizontal axis is the same height as the poverty line) and is healthy (there is no bar below the horizontal axis, indicating a loss of zero). Household 2 experiences some health problems (this is represented by the blue bar below the horizontal axis), and is also at the poverty line in terms of total expenditure: for the sake of our example, let us assume that household 2 cannot afford any health care, and that its expenditure is exactly equal to that of household 1 (bars above the axis for households 1 and 2 are equal). Household 3 experiences the same health issues as household 2 but is able to spend some of its (greater) total resources on health care (the shaded portion of the bar), which puts its total expenditure above the poverty line.[35] Comparisons between these households clarify the conundrum related to health expenditures.

Let us compare households 2 and 3. Their health status is equally compromised, but household 3 is getting health care, unlike household 2. If health expenditures were included in

[35] An alternative (and likely) scenario is that household 3 is unable to increase its *total* expenditure and simply gives up some of its non-health expenditures in order to afford health care (i.e., health expenditure *displaces* other spending). The total expenditure bar would be the same height as the others in figure 4.3, but part of it would be shaded. For the sake of simplicity, we omit this case from the figure because it does not change the conclusions we draw from the example. The issue of the displacement of current spending is relevant when considering the arguments for or against excluding "atypical" expenditures from the aggregate, a topic discussed in section 4.1.

the NCA, we would see that household 3 is better off than 2, thanks to its ability to consume more of the goods and services it desires. Her expenditure is revealing exactly the fact that she is benefitting from that spending. If we were to exclude health expenditures from the NCA, we would lose this distinction and the comprehensiveness of our measure of welfare would deteriorate, not to mention that if we did not allow a person coping with a health issue to spend money on health care she would be very unhappy. Overall, in this situation, inclusion appears to be a better strategy than exclusion.

Let us now turn to households 1 and 3. With health expenditure included in the NCA, household 3 would appear to be better off than 1, despite the fact that it only consumes more goods and services because of a health crisis. As Deaton and Zaidi (2002, 32) put it, "by including health expenditures for someone who has fallen sick, we register an increase in welfare when, in fact, the opposite has occurred." If we disregarded health expenditures, households 1 and 3 would appear equally as well off. This is also problematic because the household whose members are sick is likely to be worse off, overall, than the healthy one.

The bottom line from the discussion of figure 4.3, so far, is that each choice—including or excluding health—has its drawbacks. If we exclude health expenditures from the NCA, we miss the welfare-enhancing value of health care: keeping health status fixed, we would ideally like to capture the increased welfare of households that enjoy better care. If we include it, we attribute higher living standards to households that are actually struggling: health expenditures have the peculiar characteristic of being associated with a reduction in welfare rather than an increase, as is the case with other expenditures.

Is there some argument that would allow the analyst to pick the "lesser of two evils," and resolve the uncertainty? In fact, there is: we will argue that the two sides of the trade-off are not on the same level. While the first argument, including health for comprehensiveness, is consistent with the conceptual framework outlined in section 2, the second, excluding health because it is linked to a loss of welfare, is not. Consider the comparison between households 1 and 2 in figure 4.3, based on expenditure only. Household 2 is sicker, and so, intuitively, worse off than household 1. But both households' health expenditure is zero, so the choice of including or excluding it would not make a difference in this case. The issue with ranking households with identical expenditures and different health status persists, regardless of how health expenditures are treated. It is an inevitable shortcoming for a monetary measure of well-being, which is unable to account for some facets of welfare, such as the state of one's health. In fact, those who argue that health expenditures should be excluded often do so because they implicitly have in mind the healthy state as the counterfactual (Blinder 1985; Wagstaff 2019).[36] But is this the right counterfactual? When we compute a consumption aggregate, we are *not* trying to measure health status (much better measures are needed for that); we are trying to measure total consumption (consistently with the theory reviewed in

[36] Wagstaff (2019, 2): "we implicitly think of out-of-pocket medical expenses as involuntary and being incurred in response to a health shock, allowing the individual to return to their previous utility level but not conferring any utility per se; indeed, the receipt of medical care per se likely confers disutility."

section 2). In this perspective, failing to capture the extent to which households can afford health care is a worse outcome than failing to measure health status—we already lost the second battle the moment we opted for consumption as a measure of well-being. The natural conclusion of these considerations is that the blue bars in figure 4.3 (those below the horizontal axis, indicating loss of health) are ultimately irrelevant, and therefore, health expenditures should be included in the aggregate.

However, there is another argument in favor of excluding health expenditures from the NCA, and it has to do with their irregular and unpredictable nature. It should be stressed that, as noted by DZ, this reasoning does not apply to all health expenditures. Some items, such as preventative care, dental care, cosmetic procedures, and so on, are discretionary and disjointed from a concurring health crisis. This makes them entirely similar to all other "uncontroversial" expenditures that we examined so far and justifies their inclusion in the NCA. On the other hand, some components of health expenditure may be classified as "lumpy." Households tend to consume health care in response to negative shocks, and in some contexts, this means having to spend large sums. Health payments can be catastrophic to individual welfare (Wagstaff and van Doorslaer 2003).[37] Hospitalization is a typical example: it is a relatively rare event, and it can imply a considerable disbursement on the part of the household. If a large hospitalization expenditure is recorded during, say, a three-month reference period, annualizing it and adding it to the NCA amounts to assuming that whichever household member was recently hospitalized does so on average every year, four times per year—which is patently absurd. The "lumpiness" and infrequency of these health expenditures would suggest exclusion from the NCA, in accordance with the principle of having to proxy typical consumption (section 4.1).

But there is more. Following Ravallion (1988) and Lanjouw and Lanjouw (1997), DZ develop a theoretical framework that helps assess the extent to which the inclusion of a "noisy" expenditure component into the CA, while improving the comprehensiveness of the aggregate, can bias both poverty and inequality estimates (see also Lanjouw and Lanjouw, 2001). The advantage of a comprehensive CA, one that is inclusive of health expenditures, can be offset by the fact that health expenditures are plagued by measurement error, which is taken to mean not only that they can sometimes be recorded incorrectly,[38] but more to the point, that they are irregular: "Transitory expenditure around a longer run mean is effectively the same as measurement error" (DZ 2002, 56). It is exactly the concern for the trade-off between comprehensiveness and precision that leads DZ to recommend health expenditures be included in the CA provided (1) they do not add to much measurement error, and (2) their *elasticity* to total expenditure is high. More precisely, DZ provide the analyst with a useful result—equation (6.9) in the original paper. In particular, they show that the bias of the poverty count will be smaller when a noisy component (health expenditure, in the present context) is included in the welfare aggregate, as opposed to when it is excluded, if:

[37] Out-of-pocket health expenditures absorb, on average, between 2 to 5 percent of total household expenditures, depending on the region and the country, but aggregate budget shares mask the importance of health expenditures to households who incur health care payments. The incidence of "catastrophic" health expenditures is highest in South Asia, especially India and Nepal (Wagstaff, Eozenou, and Smitz 2019).

[38] An example is what happens in the presence of insurance schemes—individuals who have an insurance typically report only out-of-pocket expenditures or insurance copayments (both could be a small fraction of the total value), while individuals without insurance report the whole cost.

$$\varepsilon_e < \frac{\sigma_e/e}{\sigma_c/x} \tag{4.2}$$

In equation (4.2) notation is as in DZ but the interpretation is adapted: ε_e is the total expenditure elasticity of the CA after excluding health expenditures, while the ratio at the right-hand side of (4.2) is a measure of noisiness (σ denotes the variance, e denotes total expenditures net of health expenditures, while c is for comprehensive and denotes total expenditure inclusive of health expenditures); more precisely, it is the noise of non-health expenditures relatively to that of total expenditure. Equation (4.2) can be made more transparent by noting that the (weighted) sum of ε_e and $\varepsilon_{\text{health}}$ (the elasticity of health expenditures to total expenditure) sum up to one. This implies that (4.1) can be rewritten in terms of health elasticity as follows:

$$\varepsilon_{health} > f\left(\frac{\sigma_{health}/x_h}{\sigma_c/x}\right) \tag{4.3}$$

where σ_{health} is the variance of health expenditures x_h. Equation (4.3) is somewhat easier to interpret than (4.2), and provides a useful guideline: include health expenditures only if their elasticity to total expenditure ($\varepsilon_{\text{health}}$) is large, larger than a value that critically depends on the relative measurement error of health expenditures. The intuition is that the advantage of adding health expenditures to the CA in terms of coverage should not be offset by the fact that by so doing we are adding too much measurement error, where the latter is measured by the ratio $(\sigma_{\text{health}}/x_h)/(\sigma_c/x)$. The practical recommendation emerging from equation 4.3 is that health expenditures should be included only if their elasticity to total expenditure is higher than their "noisiness." DZ only use 10 words to summarize the above discussion: "The higher the elasticity, the stronger the case for inclusion" (DZ, 33).

Table 4.2 collects a number of estimates of health elasticities. Above the horizontal separator line are the estimates reported in DZ; below the line we show a number of estimates for other countries, either produced by us or available from recent publications. DZ found that health elasticities were relatively low (with the exception of South Africa). While the selection of countries in table 4.2 is mainly driven by how easily data are publicly accessible, the results in the table tend to confirm DZ's finding. There are a number of exceptions, though (Guatemala, Namibia, Panama, and Albania), and cases where elasticity are neither clearly "high" or "low" (Iraq, Maldives).

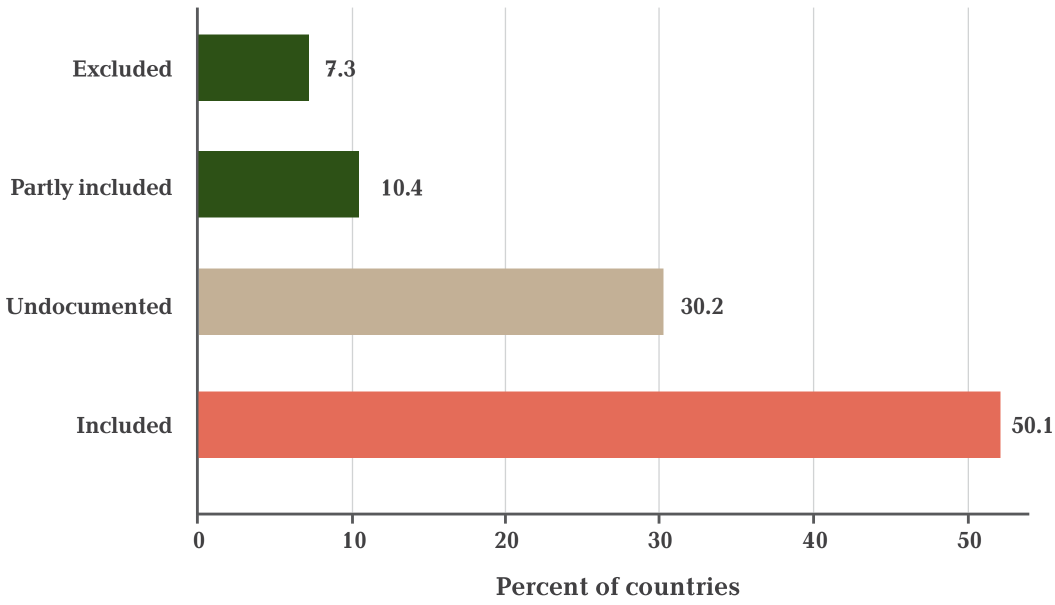
Overall, the extent to which the requirements set by equation (4.3) are met is not entirely clear, and the evidence summarized in table 4.2 does not appear to be conclusive. More importantly, it probably lacks the statistical power to prove what we hope it does—it is common to experience the fragility of regression results to the econometric specification of the model underlying table 4.2. Nevertheless, estimating health elasticity can still be useful. Analysts should start from the presumption that health expenditures should, in general, be included and then consider whether some of these are exceedingly large and infrequent: the higher the health elasticity, the stronger the case for including health expenditures. In the presence of "large" health elasticity the benefit in terms of comprehensiveness would not be offset by the measurement error that comes with their inclusion.

TABLE 4.2. Elasticity of Health Expenditures

Country	Year	Estimated elasticity	t-statistic	R^2
Deaton and Zaidi (2002, 33)				
Vietnam	1992–93	0.86	33.2	0.19
Nepal	1996	0.75	20.9	0.15
Kyrgyz Republic	1996	0.74	14.3	0.14
South Africa	1993	1.14	58.7	0.40
Panama	1997	0.80	29.2	0.25
Brazil	1996–97	0.85	31.0	0.26
New estimates				
Guatemala	2014	1.10	40.3	0.17
Tanzania	2010	0.71	19.0	0.11
Namibia	2015	1.28	67.7	0.41
Iraq	2012	1.02	72.0	0.18
Malawi	2016	0.78	28.0	0.10
Panama	2008	1.29	62.5	0.46
Albania	2008	1.28	38.4	0.23
Bosnia and Herzegovina	2007	0.72	46.7	0.24
Macedonia	2008	0.75	45.2	0.27
Montenegro	2011	0.78	19.6	0.30
Serbia	2010	0.80	39.1	0.27
Maldives	2019	0.97	3.8	0.10
Myanmar	2015	0.73	10.5	—
Palestine	2016	0.89	—	—
Lebanon	2011	0.80	—	—

NOTE: The top panel (countries above the line) is from DZ (p. 33); estimates in the bottom panel are either published figures or our own calculation. Countries have been selected based on public accessibility of their data; regressions use the inverse probabilities of selection as weights—see Solon, Haider, and Wooldridge (2015)—but results are not significantly different from unweighted ordinary least squares (OLS) estimates.
SOURCES: Guatemala: our estimates based on Encuesta Nacional de Condiciones de Vida (ENCOVI) 2014. Tanzania: our estimates based on the National Panel Survey 2010–2011. Namibia: our estimates based on 2015/16 Namibian Household Income and Expenditure Survey (NHIES). Iraq: our estimates based on the Household Socio-Economic Survey 2012. Malawi: our estimates based on the Fourth Integrated Household Survey 2016–2017. Panama: our estimates based on Encuesta de Niveles de Vida 2008. Myanmar: Myanmar poverty assessment report 2017. Albania, Bosnia and Herzegovina, Macedonia, Montenegro, and Serbia: Europe and Central Asia Team for Statistical Development (ECATSD) (2016, p.21, table 3.3). Maldives: our estimates based on 2019/20 Household Income and Expenditure Survey (NHIES). Palestine: Palestine poverty assessment report 2018, p. 18. Lebanon: Lebanon poverty assessment report 2015, p. 14. See Appendix A for full references to poverty assessment reports.

FIGURE 4.4. Including or excluding health expenditures?



SOURCE: Authors' elaboration of the dataset presented in appendix A.

Figure 4.4 shows that more than 50 percent of countries examined in our database do include *all* health expenditures in the CA. Almost 20 percent of countries can be assumed to follow DZ's original guidelines, including only a selection of health expenditures or none at all, with the rest choosing not to document the choice.

One last remark is in order. The debate on whether health expenditures should be included in the consumption aggregate does not bear on the usefulness of health expenditures per se for welfare analysts. In fact, analysts need data on visits to medical facilities to analyze the use of health services by different socioeconomic groups; they need data on health expenditures to estimate the costs to households of obtaining health care. For this reason, health expenditure data should be collected by level of care (primary, secondary, or tertiary), by type of provider (public, private, or traditional), by purpose of the visit (preventative, curative, or prenatal care), and by kind of services received (see Gertler, Rose, and Glewwe 2000; Lu et al. 2009; Heijink et al. 2011).

4.3.2 Leisure and public goods

If the measure of total consumption is to be truly comprehensive, then the opportunity to include the value of leisure and public goods should be discussed. These are "commodities" that are consumed by most households, and that certainly matter for individual welfare. All other things being equal, a household who consumes more leisure has a higher standard of living than a household with less leisure time. DZ explain why treating leisure as any other good, including its value in the CA (computed by using wage as a price, for instance) along with expenditures on other goods, is *incorrect.* Essentially, the empirical difficulties of such an attempt are prohibitive, and the assumptions required too arbitrary. DZ conclude that "the attempt to value leisure introduces more problems than it is likely to solve, and may compromise the integrity and general credibility of the welfare measures produced from the survey data." (p. 18)

Regarding public goods, in principle there is no doubt that the presence of public goods and/or of publicly provided goods such as a safe environment, good health facilities, and the like have a positive impact on the welfare of households and individuals with access to these services. In practice, the difficulties of measuring the value of these services lead DZ to recommend, once again, to "not include any valuation of public goods in the calculation of the household consumption aggregate" (p. 24).

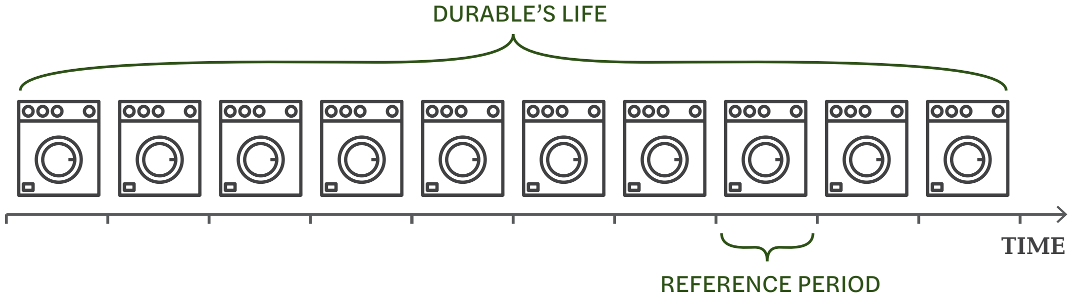
4.4 Durable goods

Durable goods, such as automobiles, household appliances, electronic devices, and the like, need to be discussed separately from other components of consumption. This is due to their defining characteristic, durability, which "is more than the fact that a good can physically persist for more than a year (this is true for most goods): a durable good is distinguished from a nondurable good by its ability to deliver useful services to a consumer through repeated use over an extended period of time" (ILO 2004, 419). Simply put, a durable good delivers utility to the consumer for an amount of time that exceeds the survey period (figure 4.5).

This is problematic from the point of view of the welfare analyst, because the price paid when a durable good is purchased reflects the value of the durable for its entire life: in fact, it is the discounted value of the flow of services that the consumer will benefit from in the future. As a consequence, a wedge is driven between the *consumption* of a durable good during the reference period—say, using a washing machine to do laundry during the year—and the recorded *expenditure* for it—the price paid to purchase the washing machine. In line with the idea of relevance (the second criterion in section 4.1), the consumption aggregate ought to include only the value of using the good during the survey year, rather than the whole purchase price. The challenge lays precisely in understanding which fraction of the purchase value is used up during the reference period, an amount that is rarely, if ever, directly observed. In technical terms, the analyst's task is to estimate the so-called *consumption flow* from durable goods.

DZ discuss a number of methods to estimate the consumption flow from durable goods, and they do so by outlining a simple theoretical framework that has been used extensively ever since. In what follows, we summarize DZ's discussion, by way of Amendola and Vecchi (2014, 2021).

FIGURE 4.5. The fundamental problem with durable goods



SOURCE: Adapted from World Bank training material.

Notation is a bit tedious, but is essential to understand the implementation of each method. Let t denote survey year. CF_t is the *consumption flow* from a durable good owned by the household during the survey period: we will use a washing machine as our stand-in for any durable good in this section. We denote by v the "vintage" or age of the good, that is, the number of years since it was manufactured (if $v = 3$ this means that the household owns a washing machine that was produced three years ago). We denote by s the number of years since the household acquired the good (if $s = 0$ it means that the washing machine was purchased during the survey year, if $s = 1$ it was purchased 1 year ago, and so on). It follows that s must be lower than or equal to v (if $s = v = 0$ then the household has purchased a new washing machine during the survey year).

In the rest of this section, we focus on three methods: the acquisition approach, the rental equivalence approach, and the user cost approach. To anticipate the conclusions, we will argue that the acquisition approach is wrong and should be abandoned, the rental equivalence approach is theoretically correct but practically unfeasible, while the user cost approach is both theoretically sound and reasonably easy to implement.[39]

The acquisition approach consists in ignoring the problem of distributing the initial cost of the durable over the useful life of the good and in allocating the entire charge to the period of purchase. This corresponds to the following estimator for the consumption flow:

$$CF_t = \begin{cases} p_{v,t} & \text{if } s = 0 \\ 0 & \text{if } s > 0 \end{cases} \tag{4.4}$$

where $p_{v,t}$ is the current market value of the good (more precisely, $p_{v,t}$ denotes how much a good produced v years ago is worth on the market at the beginning of the survey period t). According to equation (4.4), the consumption flow is zero for households who did not purchase a durable good during the survey year ($s > 0$), irrespective of whether they already

[39] A fourth approach is the *opportunity cost approach*, originally proposed by Diewert (2008) and recently suggested by Diewert and Shimizu (2019, section 5): "the opportunity cost approach to pricing the services of a consumer durable is equivalent to taking the maximum of the rent and user cost that the durable could generate over the period under consideration." To our knowledge, there has been no implementation of the opportunity cost approach, with the exception of a few studies meant to price the services of owner-occupied housing. More on this in section 4.5.

own one or not. This is clearly an undesirable solution. Even in low-income countries, households own (or have access to) a number of durable goods, while only occasionally purchasing them. Households that own and use a washing machine purchased before the survey period would be considered "as well off as" households that do not own one at all. On the other hand, according to this method, the consumption flow corresponds to a durable good's purchase price $p_{v,t}$ for households who did purchase it during the survey year (s = 0). This assumes that households that purchased a washing machine in the survey period enjoy its entire value (use it all up) by the end of the year. This is, once again, undesirable. To recap, according to the acquisition approach, items that are owned but were purchased before the survey year do not contribute to the household's well-being, while items purchased during the survey year contribute to the household's well-being for their *full* value. Both solutions are in stark contrast with the very definition of a durable good. As DZ conclude, the acquisition approach is incorrect and not recommended.

A second method to estimate the consumption flow is the so-called *rental equivalence approach*. The idea is to estimate the utility that derives from using a durable good during the survey period by collecting information on how much it would cost to rent it for a year. Assume that consumers can rent a v-year-old washing machine; in that case, the consumption flow would correspond to its current market rental value, denoted by $R_{v,t}$:

$$CF_t = R_{v,t} \tag{4.5}$$

While such a solution is simple and makes perfect economic sense, in practice it presents a number of difficulties. The single most important concern is with the existence of rental markets for each of the durable goods owned by households. Is it safe to assume that households can rent a washing machine for a year? What about other durable goods? Even if the answer were positive, how confident could the analyst be in the quality of washing machines available for rent being similar to owned ones? These and other concerns discourage from using the rental equivalence approach, at least as a first choice.

A third method to estimate the consumption flow is the *user cost approach*, which we introduce with a conceptual experiment. Consider a household that owns a durable good (we disregard the age of the good for the sake of simplicity, but with no loss of generality). Let p_t denote the market value of the good at the beginning of the survey year t. The household faces two options: (1) to *sell* the durable good, or (2) to *use* the durable good. If the household sells the durable good and invests the revenue on the financial market, at the end of the year the household receives $p_t\,(1 + i_t)$, where i_t denotes the general nominal interest rate. If, on the other hand, the household uses the durable good and sells it at the end of the year, the household obtains $p_t(1+\pi_t)(1-\delta_t)$, where π_t is the inflation rate specific to that particular durable good during year t, and δ_t is the annual depreciation rate of the good (due to both physical deterioration and loss of market value). The consumption flow can be calculated as the difference between the value of the two options at the end of the year, which is the cost that the household is willing to pay for using the durable good for one year:

$$CF_t = p_t(1 + i_t) - p_t(1 + \pi_t)(1 - \delta_t) \tag{4.6}$$

If we assume that $\pi_t\delta_t \approx 0$, this last expression can be approximated as follows:

$$CF_t = p_t(i_t - \pi_t + \delta_t) = p_t(r_t + \delta_t) \tag{4.7}$$

where $r_t = (i_t - \pi_t)$ denotes the *real* interest rate. Equation (4.7) says that the value of the services that a household receives from a durable good is the sum of two cost components: $p_t r_t$ is the foregone real interest, i.e., the interest one could have earned if one had invested the money corresponding to the value of the good in a bank account, instead of the good itself. This is what economists refer to as the *opportunity cost* of the durable good. The second component is the depreciation $p_t \delta_t$, that is, the drop in the value of the good during the year, which is, again, foregone money for the household that did not sell the good.[40]

Of the two "ingredients" needed to compute CF_t in equation (4.7), r_t is the easiest to obtain: it is usually available from sources external to the survey. Instead, the depreciation rate δ_t, which measures the loss (or gain) in value that durable goods experience with age due to physical deterioration and market value change, must be estimated. How does a washing machine depreciate over time? Do washing machines depreciate at the same rate as refrigerators? How is the depreciation rate (δ_t) estimated in practice? The answer is quite simple, provided one is ready to make a few assumptions (Amendola and Vecchi 2014, 24—31). We start by modelling the *depreciation pattern*, that is, the process that describes how durable goods lose value year after year. Suppose we start with a brand new washing machine, and denote by $p_{0,t}$ its market value in t. We denote by δ_1 ($\delta_1 \leq 1$) the deterioration rate for the first year of its life. The market value of the washer after one year, $p_{1,t}$, can be expressed as follows:

$$p_{1,t} = (1 - \delta_1) p_{0,t} \tag{4.8}$$

Following the same notation and the same interpretation, we can write the value of the good after two years as follows:

$$p_{2,t} = (1 - \delta_2)\, p_{1,t} \tag{4.9}$$

Equation (4.9) expresses the price of the washer when it turns 2 years old as a fraction of its value when it was 1 year old. Substituting equation (4.8) in equation (4.9). we obtain:

$$p_{2,t} = (1 - \delta_2)\,(1 - \delta_1) p_{0,t} \tag{4.10}$$

Proceeding iteratively for v periods gives:

$$p_{v,t} = \prod_{i=1}^{v} (1 - \delta_i) p_{0,t} \tag{4.11}$$

Equation (4.11) shows how the depreciation pattern depends on the sequence of deterioration rates $\delta_1, \delta_2, \ldots, \delta_v$. The challenge here is to estimate this depreciation sequence. To simplify the task, we may *model* the depreciation rate, and many different ways to do so have been suggested by the literature recently reviewed in Diewert and Shimizu (2019, 18—28). One model in particular, the *geometric depreciation model*, has been used extensively. It consists in assuming the depreciation rate to be constant over time:

$$\delta_i = \delta \tag{4.12}$$

which, in turn, implies that:

$$p_{v,t} = (1 - \delta)^v p_{0,t} \tag{4.13}$$

[40] Equation (4.7) corresponds to equation (3.1) in DZ (p. 35).

The last step consists in working out the depreciation rate δ in equation (4.13):

$$\delta = 1 - \left(\frac{p_{v,t}}{p_{0,t}}\right)^{\frac{1}{v}} \tag{4.14}$$

Equation (4.14) allows to estimate δ based only on information on the market values of homogeneous durable goods of different age, $p_{v,t}$ and $p_{0,t}$ It is worth noting that in practice, questionnaires often fail to record the exact pieces of information that appear in equation (4.14). This problem can usually be overcome quite easily. As a practical approximation for v (the age of the durable), the analyst may use the years of ownership of the good (this assumes that no used goods are ever purchased). Rather than asking for $p_{0,t}$ (the current market value of a new item), many questionnaires record the price originally paid for the good when it was purchased, formally $p_{v,t-v}$. If this is the case, the analyst can use an average of the inflation rate, π, in order to approximate $p_{0,t}$, as follows: $p_{0,t} \approx (1 + \pi)^v p_{v,t-v}$.

Plugging in the depreciation rate in equation (4.14) into equation (4.7) we obtain the consumption flow according to the user cost approach:

$$CF_t = p_t(r_t + \delta) \tag{4.15}$$

A second option worth considering is the *economic life* depreciation model. While the geometric depreciation model implicitly assumes that durable goods last an infinite amount of time, and their value tends to zero asymptotically, the idea behind the economic life model is that a durable good depreciates over time at a constant rate δ (the same as the geometric model), but at some point, it reaches the end of its life; at that moment, it still embodies some positive economic value, which can be thought of as a scrap value. The economic life depreciation model assumes that the scrap value equals some (small) proportion α of the good's original purchasing value: $p_{T,t} = \alpha p_{0.t}$, where T is the maximum duration of the economic life of the durable. If $\alpha = 0.05$, for instance, the analyst is assuming that when the good reaches the end of the period when it is able to be used for its original purpose, the value of the good still equals 5 percent of the purchasing price. By substituting $p_{T,t} = \alpha p_{0,t}$ in equation (4.14), we obtain a neat expression for calculating the depreciation rate:

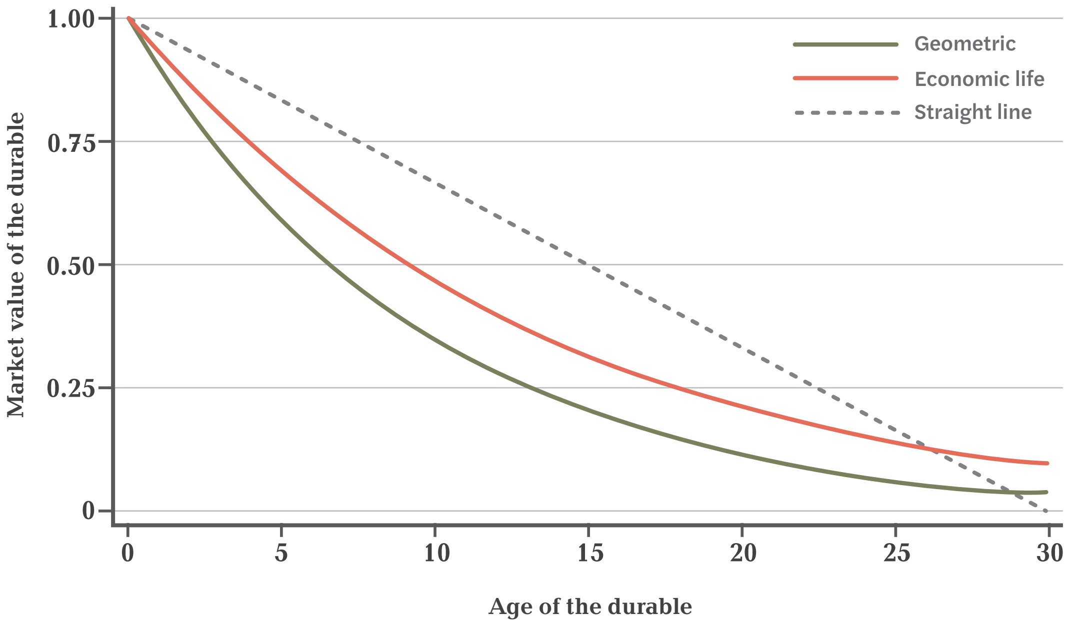
$$\delta = 1 - \alpha^{\frac{1}{T}} \tag{4.16}$$

Equation (4.16) facilitates the estimation of the consumption flow tremendously, not only for its analytical simplicity, but also because it is less demanding in terms of data, and can work when the survey only provides very limited information on durable goods, such as a mere inventory of goods owned by the households, and an estimate of their current value.[41]

[41] A drawback associated to equation (4.16) is that the estimate of δ depends on the parameter α, which is arbitrarily chosen by the analyst (any number between 1 and 10 percent appears to be a reasonable choice). The elasticities provided in Amendola and Vecchi (2021) show that the consumption flow is quite insensitive to the choice of alpha: a 5 percent increase in α is associated with a 1 percent decrease in the consumption flow.

Figure 4.6 provides a visual representation of the depreciation pattern implied by the geometric model (green line), and compares it with two alternative models, the economic life model (red line), and the straight line model (broken line), which is the simplest possible depreciation model, where the market value of the good decreases following a linear pattern; see Amendola and Vecchi (2021) for a complete discussion.

FIGURE 4.6. The geometric depreciation model compared to other models



SOURCE: Authors' elaboration.

Due to its analytical simplicity and empirical robustness, the geometric model has become popular among analysts and many statistical agencies. It does not work equally well for all durable goods—Diewert and Wei (2017), for instance, illustrate the case of computers, whose value remains more or less constant for a period and then drops suddenly—but all in all, equation (4.14) is the one to recommend for estimating depreciation rates. If data availability is limited, then equation (4.16) provides a useful alternative.

Once the depreciation rate in equation (4.14) has been estimated for all durable goods, the corresponding user costs can be computed as in equation (4.7), provided that the current market value of the durable ($p_{v,t}$) and the current real interest rate (r_t) are also known.[42] Table 4.3 illustrates the results obtained by applying the user cost method with the geometric depreciation model to the case of Maldives.

[42] DZ recommend to calculate r_t "(...) as an average over several years, and to use that real rate for all durable goods" (p. 35).

TABLE 4.3. The consumption flow from durable goods

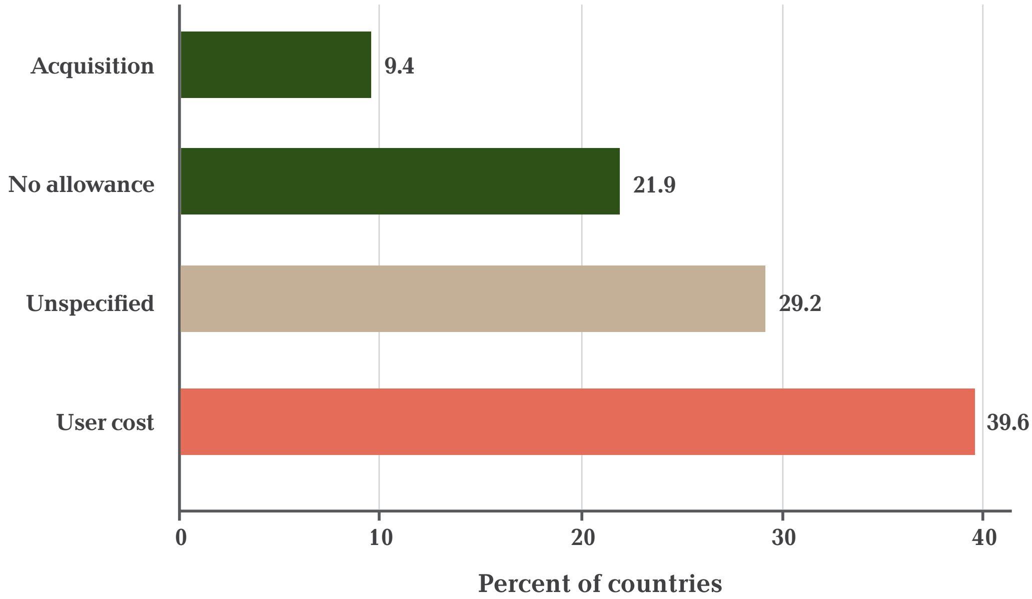
Consumer durable	Depreciation rate (eq. 4.13)	Average consumption flow (MVR per person per month) (eq. 4.14)	Number of observations
Air conditioner	0.1188	1,441.6	1,701
Bicycle	0.2083	365.7	1,905
Car/jeep	0.0742	12,082.7	142
Computer/laptop	0.1148	1,091.0	2,349
Fan	0.2083	362.5	3,720
Iron	0.2240	53.6	3,934
Mobile phone	0.1976	1,920.0	4,308
Motorcycle	0.0787	4,942.7	1,708
Radio	0.1731	43.5	2,028
Refrigerator	0.1190	535.5	3,502
Rice cooker	0.2387	106.1	3,162
TV	0.1188	749.1	3,553
Telephone	0.1541	44.7	101
Washing machine	0.1678	455.0	3,923
Water pump	0.1256	225.6	3,117

NOTE: MVR stands for Maldivian rufiyaa. Figures are obtained using the user cost method and the geometric depreciation model.
SOURCE: Authors' elaboration based on microdata from Maldives HIES 2016.

One caveat that DZ hint at, but perhaps bears repeating explicitly, is that in the absence of complete or reliable enough data for the estimation of a user cost consumption flow, it is best to omit the contribution of durable goods to household welfare entirely, rather than to include purchase values in the consumption aggregate. The former solution preserves the correct rankings among households, while the latter leads to incorrect comparisons.

Figure 4.7 shows how common each of the approaches discussed in this section are, according to recent poverty assessments around the world. The user cost method is the single most used approach, but consumption aggregates including no allowance for durable goods at all, and reports that fail to document the underlying choices are even more common when combined. The use of the acquisition approach is less frequent, but not negligible: 9.4 percent of consumption aggregates include the purchase value of durable goods in the aggregate, a practice that should be discontinued.

FIGURE 4.7. Methods to estimate the consumption flow from durable goods



SOURCE: Authors' elaboration of the dataset presented in appendix A.

4.5 Housing

Most people, if challenged to guess how well off someone is by looking at just one of the goods they own, would want to take a peek at their house. Indeed, given the size of the investment it requires, housing is, for a majority of families, the most valuable among the durable goods they consume, and is a crucial component of any consumption aggregate that aspires to be comprehensive (Malpezzi 2000; Ceriani, Olivieri, and Ranzani 2019b).

Conceptually, housing is identical to other durables when it comes to computing its contribution to the consumption aggregate. A home, once purchased, delivers utility to the consumer for an amount of time that exceeds the typical survey period of one year. The purchase value of the house represents the present value of inhabiting it for a long period of time, and according to the fundamental criteria of consumption aggregation (section 4.1), is not *relevant* as a measure of current consumption, in addition to being far from *typical*: "House purchase is such a large and relatively rare expenditure, under no circumstances should expenditures for purchase be included in the consumption aggregate." (DZ, 35). Instead, the goal is to measure the value of occupying the house for just the length of the survey period. This is equivalent to the concept of consumption flow from durable goods discussed in section 4.4, although it is more often called the *flow of housing services* in this context.

Differences between housing and other durable goods become apparent once we move into estimation territory. Unlike most other durable goods, housing generally has a rental market, and household expenditure surveys almost always record rent paid. Because rent is precisely the market value of occupying a house for a given period of time, it is a theoretically correct, and in many cases, empirically viable estimate of the flow of housing services (this is the *rental equivalence* approach described in section 4.4, equation 4.5).

But there is a problem: households who own their dwelling do not pay rent, and neither do households who occupy a dwelling provided free of charge by an employer, a relative, the government, or any other entity. Households may also rent their home at below-market rate, thanks to subsidies or other special arrangements, in which case rent paid is not zero, but still not representative of the value of services enjoyed from the dwelling. We will call these households *homeowners* and *nonmarket tenants* (which lumps together tenants paying subsidized rent and those paying no rent at all). If we were to compare them to *market tenants* (households renting their home at the market rate) based solely on annual housing expenditures, we would place the living standards of homeowners and nonmarket tenants at a systematically lower level: in fact, all else being the equal, the owner of a mansion purchased before the survey period would appear to be poorer than someone who rents an apartment (the consumption aggregate for the former would be zero in the rent category, while the latter would include a positive amount). The homeowner does "consume" housing, but we fail to account for its value because we do not observe the rent that the owner would *hypothetically* pay if she were renting her home. To rank households correctly, the analyst must estimate, or impute, *an implicit rental value*—more commonly known as *imputed rent*—for owners and nonmarket tenants, and capture the value of their consumption of housing services.

In summary, the estimation of the flow of housing services is, at least in principle, easy for market renters, for whom it corresponds with actual rent paid, and more complex for nonmarket tenants, for whom imputed rent must be estimated. Thus, we will discuss rent imputation methods at length. DZ's *Guidelines* mention four main approaches: self-reporting (or self-assessment), hedonic regression, rent-to-value, and user-cost methods. These have remained standard in the following two decades, except for a few extensions, according to a review by Balcazar et al. (2017). In the remainder of this section, we draw from this and other recent works to summarize the state of the art for rent imputation methods in the context of welfare measurement (sections 4.5.1 – 4.5.3) and summarize their advantages and drawbacks depending on the context at hand, offering practical recommendations on the best course of action for the analyst (section 4.5.4).

4.5.1 Self-reported imputed rent

Most household consumption and expenditure surveys record the dwelling occupancy status of households, sorting them into (at least) tenants, owners, subsidized tenants, and households occupying dwellings provided to them for free by the owner. It is common for the questionnaire to ask all categories, with the exception of actual tenants, to estimate the amount they would have to pay (receive), if they were to rent (lend) the dwelling they are currently occupying on the market. Such estimates are called *self-reported (or self-assessed) imputed rent*.

If these self-assessments are reliable, the analyst may simply treat them as if they were actual rents, for both homeowners and nonmarket tenants—a simple solution indeed for the rent imputation problem. The "if," however, is a crucial one in practice: this approach rests entirely on the assumption that respondents are both informed and objective about the value of their dwelling, and about the rent they would pay (receive) for a home with similar quality and location attributes. In many common situations, this assumption is questionable.

In many countries, rental markets are concentrated in urban areas, while rural populations typically own their homes, with no actual rent exchanges happening around them at all. If rental markets are "thin" (i.e., small, with few transactions), respondents may simply not be informed enough to come up with a realistic estimate of the rental value for the house they occupy. For instance, in the Maldives virtually all dwellings outside the capital are owner-occupied; self-reported imputed rents from the most recent household income and expenditure survey were found to be unreasonably high because respondents would base their answers on the rate they expected to receive if renting out their property as a guesthouse, seeing as no rental market for residents existed in their surroundings (Maldives 2018, 27).

Another possible reason why self-reported imputed rent may be distorted is lack of objectivity on the part of respondents: the so-called "owner pride" effect is the tendency for owners to "place above market values on special features of their dwellings, especially if they created those themselves" (Heston and Nakamura 2009).[43]

For these reasons, the reliability of self-reported imputed rent should always be carefully assessed before the variable is used. This is easier said than done, as there is no tried and true method to test for the presence of a bias of the self-reported measure. Section 4.5.4 offers a few recommendations.

4.5.2 Hedonic rent imputation methods

A second approach to rent imputation, or rather a family of approaches, is that of *hedonic rent imputation methods*. In this context, the term "hedonic" (which literally means "pertaining to pleasure") refers to the idea that the price of a good is determined by its characteristics, those that consumers appreciate and consider valuable. In the case of housing, this means that "a household's rent is a function of the characteristics of its dwelling, including location, structural attributes (e.g., type of construction, number of rooms, age of the building, etc.) and neighborhood characteristics" (Balcazar et al. 2017, 884). Once a specific form for the function linking rent to measurable housing characteristics (a model) is selected, imputed rents for owners and nonmarket tenants can be estimated based on the features of their homes.

The simplest such model is standard linear regression. The typical specification is actually log-linear:

$$\ln rent_h = \beta x_h + \varepsilon_h = \beta_0 + \beta_1 x_{1h} + \ldots + \beta_k x_{kh} + \varepsilon_h \tag{4.16}$$

[43] On the difference between willingness to pay (WTP) and willingness to accept (WTA), see Hanemann (1991), Fehr, Hakimov and Kübler (2015), and Tunçel and Hammit (2014). The overall indication is that measures based on WTA are usually superior to those based on WTP.

where ln $rent_h$ is the logarithm of actual rent paid by household h, x_h is a vector of k characteristics of household h's dwelling (x_1 being the first regressor, all the way to the k-th regressor, x_k), and ε_h is the error term. The model is estimated via OLS on the renter population, and regression coefficients are then used to predict rent out of sample, i.e., for homeowners and nonmarket tenants.

The choice of regressors is to a large extent determined by the information available in the housing module, and local knowledge helps to select relevant variables. In particular, if rental markets are segmented (i.e., if the same dwelling characteristics are likely to be valued differently in different locations, say densely packed urban areas versus suburban areas), the analyst should consider adding dummy variables and interactions to control for the segments.

Whatever the model chosen by the analyst, the computation of predicted values deserves a brief remark (which closely follows Wooldridge 2012, 212—213). Because the regression in (4.15) is in semi-log form, predicted rents must be "retransformed" into levels (that is, from logarithms back to currency units) before they can be added to the consumption aggregate. The seemingly natural way of doing so is to compute predicted rent as the exponent of predicted values, to "undo" the logarithm (we remove the subscript h to simplify the notation):

$$\widehat{rent} = \exp\left(\hat{\beta}\mathrm{x}\right) \tag{4.17}$$

where $\widehat{rent}$ indicates predicted rent, x is the set of covariates from equation (4.15), and $\hat{\beta}$ is the vector of estimated coefficients. However, the "naïve" retransformation in equation (4.17) is incorrect, and in fact, it systematically *underestimates* rent on average. To see why, consider the expected value E (i.e., the mean) of *rent*:[44]

$$E[rent] = E[\exp(\beta\mathrm{x} + \varepsilon)] = \exp(\beta\mathrm{x}) \cdot E[\exp(\varepsilon)] \tag{4.18}$$

The mean of exp(ε) is greater than 1, and its value depends on the distribution of ε. Therefore, the mean of equation (4.17) is always smaller than $E[rent]$.

Under standard linear regression assumptions, and when $\varepsilon \sim N(0,\sigma^2)$, the expected value of exp(ε) is $\exp(\sigma^2/2)$. Thus, the correct retransformation turns out to be:

$$\widehat{rent}_{adj} = \exp(\hat{\beta}\mathrm{x}) \cdot \exp(\hat{\sigma}^2/2) \tag{4.19}$$

where $\hat{\sigma}$ is the standard error of the regression.[45]

In practice, it is desirable to be able to compute a retransformation that does not rely on the rather strong assumption of normality of the error term. In this case, it is advisable to use the more general Duan's (1983) smearing estimator:

$$\widehat{rent}_{duan} = \exp(\hat{\beta}\mathrm{x}) \cdot \frac{1}{\mathrm{n}} \sum_{h=1}^{n} \exp(e_h) \tag{4.20}$$

[44] All means are conditional to the covariates: a more rigorous notation would be $E[rent|\mathrm{x}]$. We chose to use a looser notation for simplicity—the interested reader is directed to Wooldridge (2012, 212—213).

[45] The expression for the standard error of the regression is $\hat{\sigma} = \frac{\sum_{h=1}^{N} \varepsilon_h}{n-2}$ (Wooldridge 2012).

where e_h are the regression residuals. Equation (4.20) is easily estimated, which makes it a strong candidate for carrying out the imputation of rent. The use of Duan's estimator, while largely undocumented in the materials accompanying official poverty estimates, is gaining traction—see Palestine 2018, Mauritania 2016, The Gambia 2017.

Regarding the specification of the model in equation (4.16), a number of refinements to the simple linear model are available to the analyst, should the correction of specific biases be a priority in the context at hand. A common threat to the identification of regression coefficients is the endogeneity of housing tenure status: renters may be a nonrandom subset of the population, one with specific characteristics, that lead them to self-select into renting rather than owning (or occupying for free). If this is the case, then the coefficients estimated by OLS may reflect the peculiar features of this selected sample, rather than the market value of dwelling characteristics, and generalizing to the whole population would be a mistake. To correct for this effect—called *selection bias*—some analyses focusing on European countries and North America (Frick et al. 2010; Törmälehto and Sauli 2013) employ a two-stage Heckman correction, that models tenure status before estimating the hedonic regression (Heckman 1979). A few studies compare predicted imputed rent pre- and post-correction: for instance, Norris and Pendakur (2013) find that the Heckman correction "increases estimated household consumption by 5 percent over uncorrected estimates and decreases estimated poverty rates quite dramatically" in Canada, and Ceccarelli, Cutillo, and Di Laurea (2009) find that the correction causes imputed rent to be about 6 percent *lower* than the OLS prediction in the case of Italy. Table 4.4 illustrates the Italian example.

TABLE 4.4. Average imputed rent by area of residence, Italy

		North West	North East	Center	South	Islands	**Italy**
Self-assessment	(1)	595	627	633	365	323	**530**
OLS	(2)	448	533	460	322	253	**419**
Heckman	(3)	420	499	431	301	237	**393**
Percentage change (1)-(3)		−29	−20	−32	−18	−27	**−26**
Percentage change (2)-(3)		−6	−6	−6	−7	−6	**−6**

SOURCE: Our elaboration on Ceccarelli, Cutillo, and Di Laurea (2009).

Other extensions, such as unconditional quantile regression, and semi-parametric and nonparametric models (which account for nonlinearities in the relationship between dwelling characteristics and rent), as well as spatial models (which allow for complex patterns of spatial dependency) are available but not widely used in the rent imputation literature.

No matter the econometric specification, all hedonic rent imputation methods are vulnerable to the issue of thin rental markets, similarly to self-reported valuations. If the sample of renters is small in size, or highly segregated (that is, if the characteristics of rented dwellings are widely different than those of owned ones), or both, the information on which regression-based predictions are based may just be too poor for out-of-sample rent imputation to be reliable.

4.5.3 Other rent imputation approaches

The so-called *rent-to-value approach* can be implemented without directly leaning on survey data about market rents or self-reported imputed rents, as long as some information about the current sale value of the dwelling (property value) is available. The method relies on the estimation of a *capitalization rate*, defined as the ratio of rental value to property value. The estimated capitalization rate may be applied to property values reported by owner-occupiers to obtain an estimate of implicit rental values. One way to compute the capitalization rate is to divide the value of gross imputed rent for owner-occupiers, derived from the national accounts, by an estimate of the gross value of the owner-occupied housing stock, derived from the household survey (Yates 1994; Saunders and Siminski 2005). However, such a procedure hinges on the reliability of the national accounts measure of imputed rent, which may also be threatened by a segregated, thin rental market. It also underestimates the likely variability of actual capitalization rates, given that, depending on the available data, it may produce but a single estimated rate for the whole country (Balcazar et al. 2017, 888)

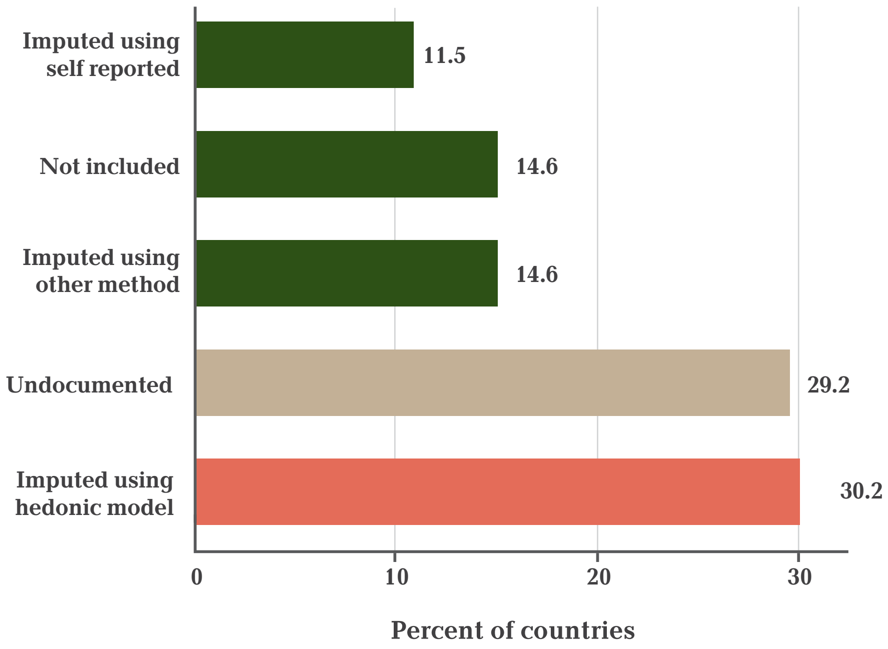
The *user-cost approach*, whose theory was discussed in section 4.4 for "ordinary" durable goods, can be applied to housing as well. However, the need to account for maintenance expenses, property taxes, insurance payments, and so on when computing the financial opportunity cost of maintaining ownership of a house during the survey period (Diewert and Shimizu 2019, 60) makes this approach rather demanding in terms of data. It is also not uncommon for housing to appreciate over time, rather than depreciate, which makes for a *negative* estimated consumption flow. All things considered, the implementation of the user-cost method for housing is problematic in practice.

4.5.4 Discussion

As the literature reviewed in this section shows, there is no lack of options when it comes to measuring the housing component of a welfare aggregate. In the current practice of welfare measurement, hedonic rent imputation appears to be the most popular method, as shown in figure 4.8, although the majority of countries do not document their choice (this is likely to hide many instances in which self-reported rent is simply used without mention). Reassuringly, no countries document the inclusion in the consumption aggregate of the purchase value of the dwelling, which is a sure mistake.

Among the available approaches to estimate rent, none is preferable a priori: the best choice is determined by the data collected by the survey, their quality, and by the characteristics of local rental markets. However, the literature does offer some pointers to help the analyst make sensible decisions.

FIGURE 4.8. Approaches to Including rent expenditures in consumption aggregates



SOURCE: Authors' elaboration of the dataset presented in appendix A.

If actual rent and self-reported rent are both available in the questionnaire, it is best to proceed with caution before simply adding both variables to the consumption aggregate; the analyst should always carry out some preliminary analysis to test the accuracy of self-assessments. Is the distribution of self-reported values reasonable, given the context? Are there systematic differences between self-reported and actual rent? No test can give a definitive answer, but it is important to at least gather some evidence: even basic summary statistics on actual and self-reported rents computed by population subgroups (e.g., urban and rural) can go a long way toward that end. Evidence of systematic differences between actual and self-reported rent indicates that non-renters regularly over- or underestimate the market value of their dwelling, and that the self-reported variable may not be reliable. Recently, Ceriani, Olivieri, and Ranzani (2019a) have explored the use of matching estimators to assess the accuracy of homeowner self-assessed rents, suggesting a procedure that has the advantage of only requiring the information usually available in household budget surveys. With reference to Peru, they find that owners tend to live in considerably larger dwellings with superior features such as bathrooms, higher quality materials, and so forth compared to tenants. This is one of the reasons why owners report what could be considered as an excessively high self-assessed rental price, possibly compounding owner's pride. Ensuring homeowner and tenant dwellings exhibit similar characteristics is therefore a first, key step for an accurate comparison. If self-reported rents are deemed implausible, and if the questionnaire gathers enough information on dwelling characteristics, then hedonic rent imputation should be considered as an alternative.

The biggest threats to the credibility of hedonic imputation are thin rental markets (the subsample of actual renters is so small that the model cannot credibly be estimated) and segregated markets (renters are concentrated in a specific area of the country) in such peculiar conditions with respect to non-renters that any generalization to the whole population is problematic. It is not uncommon for these two issues to occur simultaneously in real-life situations. Unfortunately, there is no surefire solution to the problem of thin markets: "It is unlikely that implicit rents can be predicted with any great accuracy when the share of market tenants is small" (Balcazar et al. 2017, 893). When there is simply not enough information to go by for implementing hedonic methods, the analyst may be better off turning to non-hedonic methods, provided the housing module collects data on property values.[46]

When rental markets are segregated, but the size of the renter subsample is not prohibitively small, a Heckman correction to the basic hedonic model can help mitigate selection bias in the renter population. Analysts are encouraged to experiment with the correction, with a caveat: the simplicity of the Heckman model, which can be implemented with one line of code by most statistical applications, can be deceiving. Identification of coefficients in the second steps relies on justified exclusion restrictions, which are far from straightforward in any context. A naïvely implemented correction may be worse than no correction at all. Again, non-hedonic methods may be a viable alternative, if the questionnaire allows it.

If all else fails, the analyst should seriously consider omitting housing from the consumption aggregate. The reasoning is similar to the case of durable goods: rent expenditure is an important component of the total, but including a noisy measure introduces serious error in the aggregate and is likely to cause mis-ranking of households in the final distribution. This option may be the most reasonable in cases where rental markets are very underdeveloped or nonexistent. The literature on the implications of rent imputation for welfare measurement clues the analyst in on the distributional impact of excluding housing from the aggregate. In general, it seems that the omission of rent would imply an increase in estimated inequality, given that rents are usually found to have an equalizing effect on total household expenditure; the effect on poverty is debated, although subsidized housing tends to re-rank poor households upward (Balcazar et al. 2017). Ceriani, Olivieri, and Ranzani (2019b) show that the estimated effect of including housing in the consumption aggregate is rather small for poverty and inequality estimates, although the ranking of households (and thus the poverty profile) may be more affected.

[46] In situations where the renter population is small, some analysts have adopted a sort of "reverse-hedonic" approach: a hedonic model is estimated on the non-renter population, and estimated coefficients are used to predict self-reported rent for actual renters. The underlying argument is that while self-reported rent is likely biased, at least the bias would be consistently applied to all households, preserving their rank in the expenditure distribution. This reasoning assumes that the self-reported variable is reliable, save for a constant over- or underestimation across all households (e.g., an "owner pride" effect). If rental markets are thin, which motivates the approach in the first place, it is difficult to argue that non-renters evaluate their dwelling based on objective characteristics: they have no information on which to base such judgments.

5. Adjusting for Price Variation

In virtually all real-life applications, different households face different prices. The price level of commodities and services varies both over time (inflation) and across the national territory (geographical cost-of-living differences), and with it, the purchasing power of money: the higher the level of prices, the lower the household's purchasing power. Consequently, the *nominal* consumption aggregate, that is the measure of total household consumption expenditure unadjusted for price differences, is *not* a correct measure of living standards. Two households who report the same nominal expenditure but face different prices are not able to afford the same quantities of goods: therefore, their nominal expenditures cannot be thought of as indicating the same level of welfare. Instead, nominal amounts conceal differences in the cost of living, making the household facing higher prices "worse off," all other things being equal, than the one facing lower prices.

In the rest of this section, we address the conceptual and practical questions associated with the computation of a *real* consumption aggregate (real is the opposite of nominal, a real expenditure being adjusted for purchasing power), which we shall simply refer to as "consumption aggregate." Section 5.1 provides definitions for price indices and true cost-of-living indices. The use of these indices is then discussed in the following sections: section 5.2 deals with *spatial* price variation, while section 5.3 with *temporal* within-survey price variation. Section 5.4 provides a summary of the main lessons.

5.1 Price deflators

The theoretical framework in section 2 of this document is devoted to illustrating the first recommendation from DZ's *Guidelines,* which says that welfare should be measured by total household consumption expenditure divided by a Paasche price index, that is, what we called money-metric utility (MMU). It is worth noting that (1) the idea of evaluating living standards at a reference set of prices is intrinsic to the concept of MMU, and (2) this requires the use of a price deflator. This fundamental tool has not yet been discussed in detail, and this is what the present section sets out to do.

We start our discussion by noting that the number of different goods and services available in a modern market economy is simply enormous—a single supermarket may contain tens of thousands of differently priced items. In Africa, the number of prices collected by the national statistical offices varies considerably across countries, from 1,150 price quotations in São Tomé and Príncipe to 65,000 in South Africa (ILO 2013; Gaddis 2016). China collects more than 1.1 million prices every month, while other countries reach hundreds of thousands of monthly observations, from Brazil (more than 400,000) to the Philippines

(300,000), and from Iran (200,000) to Egypt (150,000). To measure the overall price level and its changes over time and across space, these multitudes of item-level price quotations must be aggregated into one number, an *index*. The two main ways to do so consist in computing a *consumer price index* (CPI), or a *true cost-of-living index* (TCLI). Both tools are commonly referred to as price deflators (or simply deflators), and they can be used to convert nominal consumption expenditures (or incomes) into real terms; but they are conceptually different, and have different applications.

5.1.1 Price indices

A CPI measures the difference in the cost of a fixed basket of goods and services in two price situations, that is, when prices are at some reference level (a certain month or year for a temporal price index, a certain region or national average for a spatial price index) versus when they are at some other level (a different month or year, or a different region). The use of CPIs dates back to the nineteenth century: the Laspeyres and Paasche indices were first introduced in the 1870s and are now among the most popular indices used around the world (ILO 2004; Berry et al. 2019). Table 5.1 summarizes the key formulae for both indices, mostly using DZ's notation.

TABLE 5.1. Paasche and Laspeyres price indices

	Paasche		Laspeyres	
Textbook formula	$P^h = \dfrac{p^h\, q^h}{p^0\, q^h}$	(5.1)	$L^h = \dfrac{p^h\, q^0}{p^0\, q^0}$	(5.2)
DZ formula	$P^h = \left(\sum w_k^h \left(\dfrac{p_k^0}{p_k^h}\right)\right)^{-1}$	(5.3)	$L^h = \sum w_k^0 \left(\dfrac{p_k^h}{p_k^0}\right)$	(5.4)
DZ log approximation	$\ln P^h \approx \sum w_k^h \ln\left(\dfrac{p_k^h}{p_k^0}\right)$	(5.5)	$\ln L^h \approx \sum w_k^0 \ln\left(\dfrac{p_k^h}{p_k^0}\right)$	(5.6)

NOTE: The "dot-product" $p \cdot q$ in equations (5.1) and (5.2) denotes $\Sigma_k p_k q_k$. w_k^h denotes the share of commodity k in the household's budget when prices are p^h, that is: $p_k^h q_k^h / \sum_k p_k^h q_k^h$. Similarly, $w_k^0 = p_k^0 q_k^0 / \sum_k p_k^0 q_k^0$.

SOURCE: All formulae are from chapter 4 in DZ.

The first row, textbook formula, reports simple definitions of the Paasche (denoted with P^h) and Laspeyres (L^h) indices, using compact notation. We use p and q to denote price and quantity vectors, respectively (the set of prices and quantities for any number of goods). For instance, the Laspeyres index (equation 5.2) is defined as the ratio between the cost of a fixed basket of commodities (q^0), valued at the prices faced by the household (p^h), to the cost of the same basket, valued at the reference set (or base period) market prices (p^0).[47]

[47] DZ's use q^z and p^z as the base period quantities and prices, where z refers to the consumption patterns of households close to the poverty line. The interpretation of the index is the same, regardless of the choice of the base level.

A similar definition applies to the Paasche index (equation 5.1). The key difference between the two indices lies in how prices are *weighted* by the quantities consumed: while both indices account for price relatives, that is, the prices faced by the household relative to the reference prices (p^h/p^0), the Paasche index also accounts for the specific consumption pattern of each household (q^h), something that is not true of a Laspeyres index, which instead uses a set of reference quantities, q^0, that is the same for all households.

The second row, DZ formula, in table 5.1 reports alternative formulae for Paasche and Laspeyres, as in the *Guidelines*: this is the formulation that is most useful in practice when it comes to computing the indices, as it is expressed in terms of variables that are collected directly by surveys. Here, price relatives (p_k^h/p_k^0) are used and weighted not by quantities but by expenditure shares (w_k^h), which denote the share of commodity k in the total budget of household h. Budget shares are easiest to observe and compute, as consumption is most often recorded in terms of, rather than quantities (this is certainly the case for most nonfood items).

The third row, DZ log approximation, in table 5.1 reports yet another formula for the two indices, which has the advantage of expressing both as a weighted mean of price relatives, with expenditure shares as the weights.

The Paasche and Laspeyres indices are not the only available options, though. As early as the end of the nineteenth century, economists began to look for a compromise between the two. Two proposals have gained prominence in the literature. Irving Fisher (1867—1947), an American economist and statistician, proposed a "superlative price index" defined as the geometric average of the Laspeyres and Paasche indices (table 5.2, equation 5.7). Leo Törnquist (1911—1983), a professor of statistics in Finland, proposed another superlative index, defined as the geometric average of the price relatives (p_k^h/p_0^h), weighted by the arithmetic average of the expenditure shares (w_k^h *and* w_k^0) of commodity k for the two periods (table 5.2, equation 5.8). By taking the logs of this index, as in equation 5.9, we can express it as a simple arithmetic average of the (log) price relatives (p_k^h/p_k^0) with average weights, a form similar to equations (5.5) and (5.6). While fixed-weight price indices such as the Paasche and Laspeyers have been used primarily in the *temporal* context to measure price changes between two time periods (section 5.3), the Fisher and Törnquist superlative indices have gained increasing prominence in the *spatial* context (section 5.2), where the aim is to estimate subnational purchasing power parities (PPPs), with the whole country treated as the reference region (Ray 2018, 41).

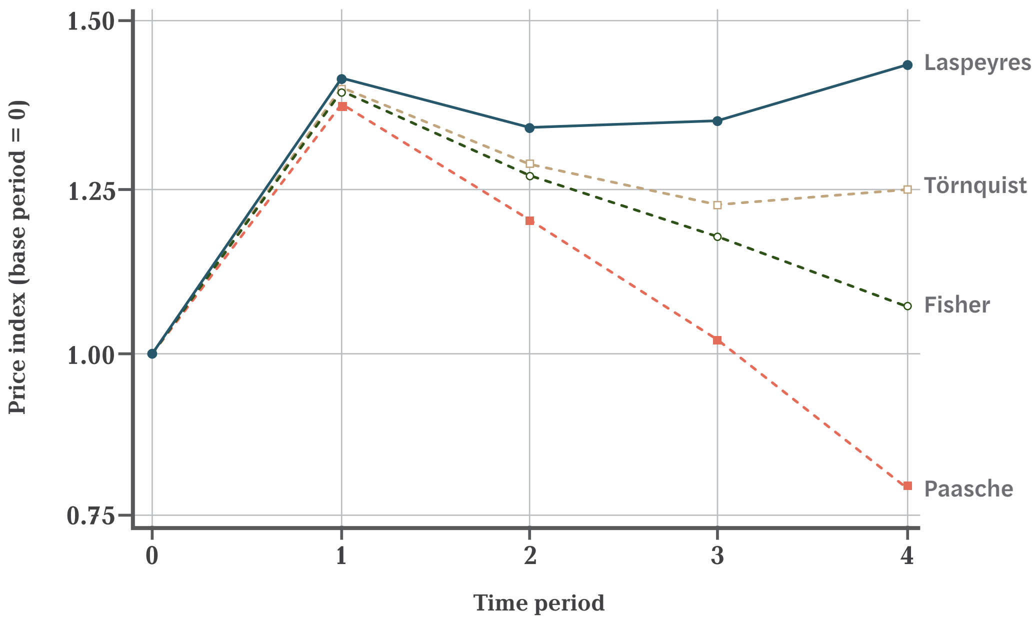
TABLE 5.2. Fisher and Törnquist superlative price indices

	Fisher		Törnquist	
Textbook formula	$F^h = \left(L^h \times P^h\right)^{1/2}$	(5.7)	$T^h = \prod_{k=1}^{n} \left(\frac{p_k^h}{p_k^0}\right)^{\left(\frac{w_k^0 + w_k^h}{2}\right)}$	(5.8)
Log formula	–		$\ln T^h = \sum \left(\frac{w_k^0 + w_k^h}{2}\right) \ln\left(\frac{p_k^h}{p_k^0}\right)$	(5.9)

SOURCE: Deaton and Muellbauer (1980, ch. 7).

What makes Fisher and Törnquist "superlative" price indices? The term is statistical jargon, to indicate a technical property enjoyed by a small circle of price indices (Diewert 1976): a superlative index is expected to provide a good approximation of the true cost-of-living index (TCLI), that the economic literature on price deflation regards as the appropriate target for any consumer price index (Diewert, Greenless, and Hulten 2010, 2).[48] This brief explanation exposes a question that has been lingering around our discussion: how is one to choose among the different available indices? Understanding what a TCLI is, exactly, is essential for grasping some of the arguments for and against each candidate. Before turning to this task, we turn to a related question: how much does the choice between consumer price indices matter, empirically?

FIGURE 5.1. Same data, different price indices



SOURCE: Authors' elaboration based on ILO (2004, ch. 19).

Figure 5.1 illustrates the different pattern of the price indices considered in table 5.1 and table 5.2. Underlying figure 5.2 is a "toy" economy, as designed by the ILO's (2004) *Consumer Price Index Manual*. Data are artificial but realistic (details available in appendix C), and the key result is general enough to be worthy of a comment. In the figure, the *y*-axis measures the value of each index after setting the reference set of prices equal to 1; we can interpret the *x*-axis either as measuring different time periods (in which case the graph represents the dynamics of inflation) or different regions (in which case the graph represents geographical

[48] A short historical note to explain the origin of the term "superlative." Fisher (1922), a landmark in the history of index number theory, listed some 180 bilateral price index formulae—in principle, 180 candidate formulae to contend the role played by the Laspeyres and Paasche indices. Fisher tested the properties of 134 of these formulae and then graded them according to how far away numerically they were from his ideal index. He classified the formulae into seven groups, in increasing order of merit: worthless, poor, fair, good, very good, excellent and *superlative* (Diewert 2013, 22). Laspeyres only received a "very good."

differences in the average level of prices). Either way, the takeaway from figure 5.2 (Laspeyres and Paasche can take very different trajectories, while Fisher and Törnquist are more consistent with each other) is that the choice of index number formula matters. Indeed, it might matter a great deal.

5.1.2 True cost-of-living indices

A TCLI is defined as the ratio of the minimum expenditure needed to obtain a certain standard of living, again in two price situations. The TCLI is rooted in the consumer theory reviewed in section 2.2: if we ask "what is the minimum amount that the consumer needs to spend in order to achieve utility u at prices p?", the answer is "the cost function." that is $c(u, p) = x$, introduced in equation 2.2.[49] If we denote two vectors of prices by p^0 and p^1—0 and 1 can be thought of as two different years, or two different regions, then the ratio between two cost functions evaluated at p^0 and p^1 defines a TCLI:

$$TCLI\left(\bar{u}, p^0, p^1\right) = \frac{c\left(\bar{u}, p^1\right)}{c\left(\bar{u}, p^0\right)} \tag{5.10}$$

According to equation (5.10), the TCLI is a measure of the change in the amount a household would have to spend in order to maintain a given standard of living (represented by $\bar{u}$, which denotes a certain utility level) as a result of a change in the price level. The denominator is the minimum expenditure to achieve utility $\bar{u}$ given the set of prices p^0, while the numerator is the minimum expenditure to achieve the same utility level $\bar{u}$, but at prices p^1: the ratio is a measure of the extent to which the cost of living has changed due to the change in prices from p^0 and p^1.

The key difference between a TCLI as in equation (5.10) and any of the price indices seen in table 5.1 and table 5.2 is that, when prices (and only prices) change, the TCLI does not assume that consumers continue to purchase the same basket, as implicitly assumed in the indices in table 5.1 and table 5.2, be it a household-specific basket (Paasche) or one that is common for the whole population (Laspeyres). Rather, the TCLI allows consumers to shift their purchases away from goods whose relative prices have increased, and toward goods whose relative prices have fallen (National Research Council 2002). Simply put, the TCLI explicitly incorporates the effects of these substitutions in, say, reducing the expenditure required by a consumer to maintain a given standard of living when prices increase.

In general, the TCLI and the CPI formulas illustrated above give different results (Deaton and Muellbauer 1980, 172—173. Take the Laspeyres index: one of its key features is that it tends to *overstate* cost of living differences by not allowing any substitution between goods to occur when prices vary (Diewert 2001). This is a consequence of holding quantities q^0 fixed at the base (reference) level: in reality, when faced with a price change, households tend to

[49] The TCLI was originally introduced by Russian economist Alexander A. Konüs (1895—1990) in 1924, with a paper that was then translated into English in 1939. See also Deaton and Muellbauer (1980), Triplett (2001), and Ray (2017, 2018).

reallocate their resources away from relatively pricier and toward relatively cheaper goods. Consequently, the Laspeyres index provides an upper bound to the cost of living faced by a household. The opposite problem arises with the Paasche formula, because the weights q^h are set at the level that is optimal at prices p^h after all substitutions have taken place. To the extent that price and demanded quantity are negatively correlated, the Paasche index provides a lower bound to the cost of living faced by the household.[50]

To sum up, one can argue that neither Laspeyres nor Paasche indexes capture consumer substitution adequately: when faced with differences in relative prices, consumers are likely to adjust their consumption patterns toward relatively cheap goods and away from relatively more expensive ones (Deaton and Tarozzi 2005). In this respect, other CPIs, like the Fisher or Törnqvist index, do a better job—in fact, we emphasized that they are "superlative" indices: they are expected to approximate well the TCLI in equation (5.10) and, under special conditions, they are exactly equal to it.

The question of whether to prefer one of the available CPIs or a TCLI to deflate the nominal consumption aggregate, however, is complex, and goes beyond the issue of consumer substitution. Each option has advantages and disadvantages, and no general recommendation has been agreed on. For instance, the adoption of a TCLI is riddled with some major empirical difficulties. First, utility is not observable: any utility specification is necessarily ad hoc, and if the results are sensitive to the specification of the utility function, then theory does not provide much guidance—see Ray (2017); second, the TCLI approach requires more data than the CPI approach—see, for instance, Oulton (2012). More in general, the price index literature is divided into what Triplett (2001, F332) describes as "two different worlds": on one side are academics, who are mainly concerned about the properties of index number formulas and for whom "substitution bias" is at the forefront; on the other side are statisticians and statistical agencies, who are more likely to appreciate the technical and empirical difficulties that arise when constructing an actual price deflator. As Triplett remarks, "both the typical academic and the typical [statistical] agency positions are partly right and partly wrong," meaning that, in practice, the choice is a matter of balancing trade-offs. In fact, we are about to add a third perspective into the mix: that of welfare analysts, who are interested in which price deflator is best *for the specific purpose* of deflating the nominal consumption aggregate.

The goal of this section is not to reach a conclusion on which of the candidate indices is best in general, but rather, to familiarize the reader with the "machinery" that welfare analysts have at their disposal when faced with the need to compute a real welfare indicator. Having a good grasp of the pros and cons of CPIs and the TCLI is essential to understand the debates

[50] This is a technical note that can be skipped without affecting the understanding of the argument in the main text. The TCLI defined in equation (5.10) depends on the reference utility level $\bar{u}$. If $\bar{u}$ is chosen to be that corresponding to the standard of living in the base price situation, say u^0, then the TCLI will correspond to the Laspeyres price index. If $\bar{u}$ is chosen to be that in the comparison situation, say u^h, then the TCLI will correspond to the Paasche price index (Ray 2018, 41). Gaddis (2016, section 3.2.2) provides a concise and accessible discussion of the so-called "CPI bias" that is the difference between the CPI and a (conditional) TCLI. See also Deaton (1998), and Hausman (2003).

that have developed around the issue of cost-of-living differences in the decades following DZ's first recommendation, that is, "you shall measure welfare using MMU."[51]

5.2 Spatial deflation

The computation of a welfare indicator consistent with utility theory calls for a specific adjustment for cost-of-living differences: division by a Paasche price index (section 2). This is the advisable choice if poverty is to be measured by ranking individuals according to their MMU.

While theory delivers clear instructions, a number of practical difficulties often interfere with the use of Paasche (as well as with all other price index formulae examined in section 5.1.1). As is apparent from equation (5.1), a Paasche index suitable for welfare measurement is household-specific (it is a function of q^h, the bundle of goods and services consumed by each household h) and requires a full set of market prices for each and every item in each household's consumption basket. Unfortunately, household surveys are typically not able to collect prices for all consumption items think, for example, of nonfood items or services, whose prices are clearly hard to measure (Gibson 2007, 5). This raises the issue of how to carry out the estimation of a Paasche index in practice.

Information on prices is, of course, paramount. Deaton and Grosh (2000) discuss at some length the pros and cons of different sources of price data. DZ built on that discussion, and narrowed the list down to three main sources, namely (1) a dedicated price questionnaire (e.g., a price survey administered in each cluster as part of a community questionnaire), (2) other price sources (e.g., government price surveys), and (3) the household purchases reported in the survey, which allow for the computation *unit values*, that is, the ratios of reported purchase expenditures to quantities (Houthakker and Prais 1952). Dedicated price questionnaires and other sources focused specifically on market prices are still the exception in a majority of low- and middle-income countries—surveys prioritize collecting expenditure or income data over price data (Gibson 2013; Gibson and Rozelle 2011)—so the information that the analyst typically has at her disposal is of type (3). The next section discusses this source in detail.

5.2.1 Unit values

Unit values have a long list of advantages and have been used extensively in the past. Their fortune seems to have turned, however, as they are increasingly accused of being a poor proxy of market prices. In an influential article published in the *American Economic Review*, Deaton (1988) identified the nature of the problem. A first issue is due to the fact household surveys typically collect data using a 5- or 6-digit Classification of Individual Consumption According to Purpose (COICOP) system: with this level of detail, analysts can count on

[51] See DZ (2002, 23). Further insight can be gained reading van Veelen (2002, 2009), and van Veelen and van der Weide (2008).

data relative to the consumption of, say, "meat." The unit value of "meat" is not a price, not in technical terms. Why? Because "meat" is not a homogeneous commodity, but a collection of elementary commodities, a "commodity group," containing commodities of varying qualities (in this case, different cuts of meat). This implies that a decrease in the unit value of meat could reflect either a price decrease or a shift toward the consumption of cheaper commodities within the group.[52] This issue has two main consequences. First, since better-off households tend to buy higher-quality goods, unit values tend to be positively correlated to incomes, and consequently, price indices based on unit values are likely to be artificially higher in areas where wealthier households live. Second, it seems reasonable to assume that households respond to an increase in the price of a type of meat by switching to cheaper cuts of meat, with the consequence that the change in the unit value for meat will understate the original price increase. In short, when the quality of goods consumed varies over time (or across regions), unit values can be a flawed indicator of prices: changes in unit values may not capture genuine changes in prices, but rather changes in the quality of consumed goods, as chosen by consumers.

Deaton (1988) demonstrates that unit values can still serve as a useful proxy for prices in some circumstances (McKelvey 2011, 157). In a recent series of studies John Gibson and co-authors have tested a key assumption, one that should hold if we are to trust unit values to proxy market values correctly: prices of each elementary good in a commodity group should move in fixed proportions across locations (technically, this property is known as *Hicksian separability*). "For example, pork loin is an expensive cut while shoulder is not. If the ratio of loin to shoulder prices is lower in one town than elsewhere, consumers there buy relatively more loin, giving a higher unit value than under fixed price ratios (the unit value is weighted more towards loin)." (Gibson and Kim 2015, 34). When this happens, relying on unit values would lead to the wrong conclusion that prices are higher in the region where more loin is bought (because it is relatively cheaper than in other regions), while in fact what is happening is that we are comparing commodities of different quality. Gibson and Kim (2015) have argued that within-group prices are unlikely to be constant across regions due to the *Alchian-Allen effect* (also known as the "shipping the good apples out" hypothesis). The idea is simple: in the presence of transport, storage, or processing costs, there are incentives to trade high quality goods, while holding on to low quality goods for local, domestic consumption.[53] For the case of Vietnam, they find that unit values do change consistently with the Alchian-Allen hypothesis: the unit value for the "rice" commodity group in the north is 6 percent higher than in the rest of the country, purely due to the quality mix effect, irrespective of any genuine spatial price differences. This has potential empirical relevance in the context of poverty measurement: in Vietnam, "rice was 36 percent of the value of the food basket used for the cost-of-basic-needs poverty line prior to 2010 so this quality mix effect spuriously lifts the food poverty line by 2 percent in the north. In effect, a higher standard of living (eating higher quality rice) is being confused with a higher cost

[52] "Even for a narrowly defined commodity such as beef, there are more and less expensive cuts, and there are lean, scrawny (and cheap) agoutis (a large rat), as well as fat, sleek, and tasty ones" (Deaton 1988, 420).

[53] Imagine high quality apples sell for $2 a kg and low quality apples sell for $1 a kg. Transportation cost to ship apples in a different region is $0.50 per kg. In the place where apples are produced the relative price of high quality to low quality apples is 2:1, that is, high quality apples are twice as expensive. At the place of destination, the relative price of high quality to low quality apples is: 2.5:1.5, that is, high quality apples are only 67 percent more expensive. Incentives are to ship high quality apples. The presence of a per unit transaction cost lowers the relative price of high quality goods (equivalently, it increases the demand for high quality goods).

of living. The upward bias in the poverty line in the north due just to the rice quality mix raises the head count poverty rate there by almost 5 percent in 2010" (p. 438). Omoniyi, Vundru, and Kilic (2021) obtain similar results for Malawi, with a well-designed analysis based on a national market survey combined with the Fifth Integrated Household Survey (IHS5) 2019/20 and the Integrated Household Panel Survey (IHPS) 2019.

In sum, despite their many advantages, the use of unit values as a substitute for market prices in welfare analysis is being increasingly criticized. Many solutions have been explored. Early attempts aimed to adjust unit values with regression-based methodologies and suggested that the upward bias of unit-value—based price indices was real but "small" (Deaton 1987, 1990), and this explains their popularity during the 1990s and 2000s.[54] By contrast, a number of recent studies found that the income (or quality) effect has a significant impact on spatial price differences, even in the presence of corrections. Comparing adjusted with unadjusted unit values, Deaton and Dupriez (2011) found differences up to 50 percent for certain food items. Gibson and Rozelle (2005) designed a unique experiment during a survey in Papua New Guinea and found that "unit value overstates the average market price by about 30 percent for sweet potatoes and bananas, the two most common locally produced foods" (p. 77). Their conclusion was that "there are substantial biases when unit values are used as a proxy for market price, even when sophisticated correction methods are applied" (p. 69). A similar conclusion is reached by Capéau and Dercon (2006) for Ethiopia, and Gibson and Romeo (2017) for Melanesia.

We are now better placed to comment on DZ's final recommendation, namely to estimate a Paasche index by using "within-survey prices supplemented by prices from the price questionnaire, if available" (p. 54), as in the following formula:

$$\ln P^h = \sum_{k \in \text{food}} w_k^h \ln\left(\frac{p_k^h}{p_k^0}\right) + \sum_{k \in \text{non food}} w_k^h \ln\left(\frac{p_k^c}{p_k^0}\right) \tag{5.11}$$

where the first summation refers to the set of goods for which unit values are available at the household level, hence the suffix h for household (typically, food items); the second summation refers to goods for which unit values are not available (typically, nonfood items). In the latter case, "price relatives will come from community questionnaires or even other regional sources, and will not be available at the household level" (hence the suffix c, for cluster, in the notation).[55] Certainly, unit values cannot be used unless careful consideration is devoted

[54] See Deaton (1997, 283-ff) for a general discussion. Deaton and Tarozzi (2005) carried out a large-scale study on prices in India showing the great potential of unit values, which they used to track inflation over time and also to compare price levels across regions. For Rwanda, Muller (2008, 40) found that market prices and unit values were very close for elementary-level food items. The *Harmonization Guidelines* developed by the team responsible for Europe and Central Asia do trust unit values: "ECAPOV adopts, whenever information on purchased quantities is available, the Paasche index. We use the unit values from food consumption (whenever available from the diary of consumption) and we apply it to all consumption items." (ECASTD 2016, 31).

[55] Equation (5.11) in the text corresponds to DZ's equation (4.7), page 44, where p_k^h are referred to as within-survey prices, or equivalently as household-level unit values. The fact that price data may not exist at all for a sizable group of items (those for which unit values cannot be computed) clearly poses problems to the coverage of the index. Gibson (2016, 437) points to the "growing share of food consumed, as meals and in the catchall categories at the end of lists of types of ingredients, [that] has no quantity data so unit values cannot be calculated." In terms of equation (5.11), this implies that the number of food items in the first summation gets smaller and smaller. See also Gibson and Kim (2013).

to the issue of *quality bias*. Depending on the context, this might simply consist in using the primary sampling unit (PSU) medians (to gain protection against extreme values and excessive heterogeneity in household-level unit values), or more sophisticated regression-based adjustments. For instance, one can consider purging the unit values from quality bias by estimating a simple linear regression:

$$\ln(uv_{ih}) = \alpha + \beta_i ln(x_h) + \delta_i D_h + \gamma_i X_h + \varepsilon_i \tag{5.12}$$

In equation (5.12) the dependent variable $\ln(uv_{ih})$ is the log of the unit value for the *i*-th commodity consumed by the *h*-th household, while the covariates include the log of per capita total household expenditure $ln(x_h)$, a vector of dummy variables (D_h) for the geographical location of the households, and other socio-demographic controls (X_h), such as number of adults, number of children by age group, and so forth. Following the discussion in Chen and Ravallion (1996, 36) and Deaton (1997, 288) the regression coefficients on the location dummy variables can be interpreted as differences in prices purged of quality differences.

In conclusion, we might describe the current situation as an ongoing transition between the pre-Deaton era, where the use of unit values was the rule, and a new paradigm, where reliable price data will hopefully become widely available. Gibson and Kim (2019) are but the last in a string of studies that express skepticism on the use of unit values (and more generally on household survey data), many of which were cited in this section: the *pars construens* of this argument consists in urging economists "to apply enough pressure on statistical agencies to implement household surveys that gather both market prices and unit values" (p. 18). In the same vein, Gaddis (2016, 11) concludes her review on prices for poverty analysis in Africa as follows: "Even though unit values may be considered a useful source of price data in developing countries today, they are expected to lose relevance in the future." In the meantime, the recommended strategy is to explore the extent to which quality bias is a concern in the context at hand, for instance by summarizing median unit values by item and by per capita expenditure decile (the presence of a steep gradient would be taken as evidence of a non-negligible quality bias), and consider implementing regression-based empirical adjustments as required.

5.2.2 Poverty line ratios and CPI-based methods

When the estimation of a Paasche price index as advised by the *Guidelines* and illustrated in equation (5.11) is deemed unfeasible with the available data, what are the alternatives? One option is the use of *implicit* (or *indirect*) *spatial deflators*, of which *poverty line ratios* are a special case with special relevance in the context of welfare measurement. There are countries where poverty lines are estimated at the household level, for example, Botswana, Pakistan, or Myanmar; other countries estimate poverty lines separately by region or other population subgroups (provinces, districts, governorates, etc.). In such cases, the comparison between these "local" poverty lines is precisely an indicator of spatial price differences. Suppose to calculate median values of household-level poverty lines by region: if we denote by PL_r the median poverty line in region *r,* then the ratio of a regional poverty line to the national poverty line (say, $E[PL_r]$) corresponds, under certain conditions, to a true cost-of-living index:

$$SPI_r = \frac{PL_r}{E\left[PL_r\right]} \tag{5.13}$$

where SPL_r denotes the spatial price index in region r. The intuition behind equation (5.13) is simple and rooted in the theory reviewed in section 2. In abstract terms, a poverty line can be interpreted as a cost function $c(u, p)$, evaluated at a "minimum" utility level u_z (where the suffix z indicates the poverty line)—see Ravallion (2008, 556) in the *New Palgrave Dictionary of Economics*. Thus, equation (5.13) is the ratio of two cost functions, which is exactly the definition of the TCLI in equation (5.10). Note that both cost functions are valued at the reference utility level u_z : if SPL_r is greater (lower) than 1, then the interpretation is that the cost of achieving the utility level u_z in region r is greater (lower) than average. Table 5.3 illustrates for the case of Botswana. The first two columns report the poverty lines (at constant prices) for 2003 and 2009, by region; columns 3 and 4 show the regional poverty line ratios defined in equation (5.12).

TABLE 5.3. Implicit spatial price indices (SPI) in Botswana, 2003–2009

	PDL 2003	PDL 2009	SPI 2003	SPI 2009
Gaborone	579	1,068	76.4	81.0
Francistown	721	1,133	95.2	86.0
Other cities/towns	728	1,094	96.1	83.1
Rural southeast	734	1,296	96.8	98.4
Rural northeast	815	1,428	107.6	108.4
Rural northwest	808	1,479	106.6	112.3
Rural southwest	863	1,388	113.9	105.3
Cities/towns	661	1,091	87.8	82.2
Urban villages	806	1,532	107.1	115.5
Rural areas	760	1,269	100.9	95.7
Rural	760	1,269	100.0	95.6
Urban	760	1,373	100.0	103.4
Botswana	758	1,318	100.0	100.0

NOTE: PDL (poverty datum line) indicates regional poverty lines expressed in Pula per household per month; SPI (spatial price index) indicates the spatial price deflator derived from the poverty lines.
SOURCE: Botswana (2015), Annex C, p. 180.

Joliffe, Datt, and Sharma (2004, 561) estimate implicit spatial deflators for Egypt based on five cost-of-basic-needs (CBN) poverty lines, finding remarkably large geographical differences in prices (36 percent difference between rural and metropolitan areas).

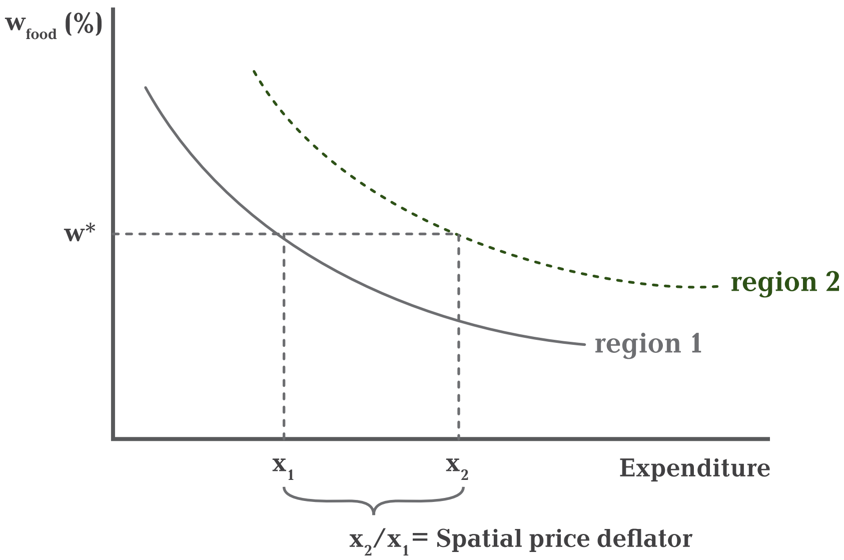
Gibson (2007, 5) notes that this method provides a solution to situations where reliable price data are not available—equation (5.12) does not require any price data. The analyst should pay attention, however, to the fact that poverty line ratios account not only for differences in the cost of living, but also for differences in a number of other factors, such as different consumption patterns or differences in needs. For instance, household-level poverty lines, by accounting for different caloric needs and housing costs due to the different demographic composition of households, automatically incorporate an allowance for differences in needs and economies of scale. This can be seen as an advantage or a hindrance, depending on the context, so that there is not one specific prescription to use or avoid poverty line ratios based on this. If the analyst trusts the method underlying the construction of the regional poverty line, equation (5.13) is a strategy worthy of attention.

A second type of implicit spatial deflator is based on temporal CPIs. CPI price data exist in almost all countries (Berry et al. 2019). The problem with the CPI is that it is a tool designed for tracking inflation, and the transposition from intertemporal to spatial price measurement is not straightforward. For instance, Gibson (2007, 6) argues that "in cases where information on price levels across regions are lacking, analysts may be tempted to estimate these regional price levels at a given point in time by applying a local CPI to some base period when cross-sectional price levels were known. For instance, a baseline household survey may enable poverty lines and other deflators to be estimated for each region while subsequent surveys lack a price collection module (or lack quantity information to derive price movements from unit values). But if there is a CPI available for each region (or for key cities within or near to each region) this might be used by a poverty analyst to estimate current price levels across regions. The available evidence suggests that such a procedure is biased." In the same paper, Gibson cites a study on Russian data carried out by Gluschenko (2006) that compares a spatial deflator computed directly (using local prices) with an indirect one obtained by applying local CPIs to a pre-existing spatial deflator, finding large differences: "the direct method gives a spatial price index for each province whose range is 44% of the national mean price level, but the indirect method gives a much wider range, of 72%" (Li and Gibson 2014, 96). The discussions in Gaddis (2016) and Chen et al. (2020, 6) lead to similar conclusions.

5.2.3 Engel curves

Another strategy to sidestep the lack of adequate information for estimating CPIs might be to use "no-price" methods: Engel curves, for example, provide a way to estimate spatial deflators in the absence of any information on prices. This approach follows a tradition that dates back to Arrow (1958, 78), recently revived by Costa (2001) and Hamilton (2001), and followed by a number of other authors, for example, Coondoo, Majumder, and Chattopadhyay (2011) or Almås (2012). The idea hinges on the so-called Engel's Law, a well-known economic empirical regularity, according to which the percentage of total household expenditures spent on food declines with the standard of living. The idea of using Engel curves to compute an indicator of spatial differences in prices can be explained by means of a syllogism. The major premise is that the Engel curve, a function that describes how the food budget share varies with total expenditure (or income), exists and is unique. The minor premise is that identical households with the same budget share devoted to food have the same real expenditure. The conclusion is that systematic differences in measured nominal expenditures of households with the same food budget shares reveal price-level differences, figure 5.2 illustrates. The curves labelled region 1 and region 2 are Engel curves estimated for two different regions; the horizontal distance between the two curves is the difference between the total nominal expenditures of households that have the same food budget share w^*, but live in different regions. If Engel's Law holds, this difference is entirely attributable to a difference in the cost of living and can be interpreted as a spatial price deflator.

FIGURE 5.2. Illustration of the Engel curve method



SOURCE: Authors' elaboration.

Deaton and Dupriez's (2011) assessment of the Engel curve method is negative, their argument being that the Engel curve is *not* unique and is *not* stable (as assumed by the major premise in the above syllogism), and that the Engel curve varies with a number of factors, namely preferences, lifestyle, and so forth.[56] Gibson, Le, and Kim (2017) also advise against the Engel curve method. For the case of Vietnam, they argue (and document) that spatial deflators estimated from a food Engel curve are a poor proxy for deflators obtained from multilateral price indexes. Similarly, after an extensive literature review, Gaddis (2016, 22) concludes that "empirical studies confirm that results from Engel curve estimations are not necessarily robust, especially in spatial contexts."[57] More recently Almås, Beatty, and Crossley (2018) have argued that the Engel curve method is "internally inconsistent."

Overall, the use of "short-cut" methods such as that based on Engel curves is not recommended for poverty measurement (Ravallion 2016, 166).

5.2.4 Other strategies

Alongside the different strategies reviewed so far, a number of other approaches have been explored. Kakwani and Hill (2002), for instance, developed an axiomatic approach for constructing spatial cost of living indices and applied it to Thailand. While the method is appealing for theoretical reasons, the difficulties associated to its implementation are likely to hinder its diffusion. Li and Gibson (2014) suggested the use of specific-price indices, given

[56] A similar argument was put forward by Ravallion and Bidani (1994) in the context of their critique of the food energy intake method for estimating poverty lines.

[57] See also Dabalen, Gaddis, and Nguyen (2016).

a Balassa-Samuelson framework: the idea is that in a context where traded goods prices converge rapidly, the main source of price dispersion across space should come from nontraded items, hence the suggestion to construct spatial indices based on house prices. While the focus of the paper is on China, a growing international literature suggests that this argument has a broader applicability: housing costs, in particular, account for a large share of the total inter-area price differences (Joliffe 2006; Moretti, 2010; Pittau, Zelli, and Massari 2011).

A recent development worthy of consideration is in Chen et al. (2020). The authors apply the Country Product Dummy (CPD) method—originally developed by Summers (1973) to deal with missing price information on cross-country data, and subsequently used in the International Comparison Program (ICP)—to a CPI price database, that is to data collected for measuring inflation, and obtain a (multilateral) *spatial* price deflator. To the extent that CPI prices do not suffer from classical problems, for example, coverage limited to urban areas (urban bias), or a weighting scheme that is not consistent with the consumption pattern of the poor (plutocratic bias), then the combination of regional CPI prices and the CPD method delivers spatial deflators. At the time of writing, this new method has only been tested in Ghana, Rwanda, and South Africa, producing encouraging and insightful results (the implementation of the method is simple, and we learn, for instance, how large the bias in spatial price measurement is when housing expenditures are not properly accounted for, when considering only food prices, or when rural areas are not well covered by the survey).

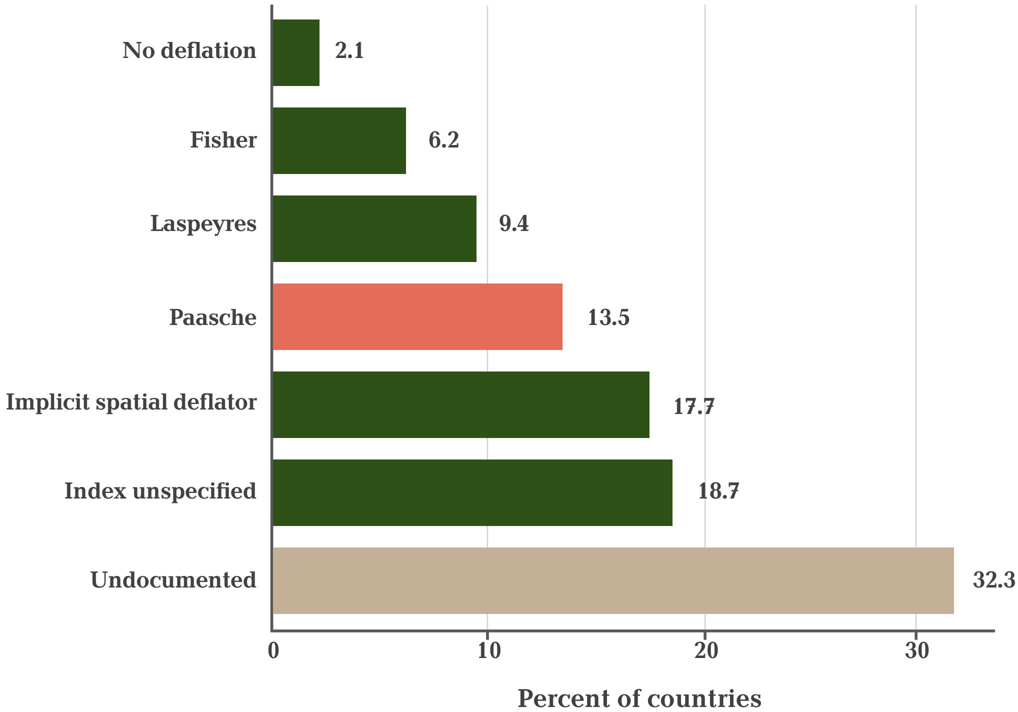
Finally, there are also arguments in favor of a "do nothing" strategy, that is, avoiding adjusting for spatial price differences altogether. Issues of estimation of the price index may be entirely redundant if one is convinced that spatial deflation does more harm than good to the ultimate goal of making welfare comparisons between households. A recent study on Italy, a country marked by a sesquipedalian North-South divide, argued that because low prices are usually associated to lower levels of public goods provision, and other environmental aspects whose negative impact is not necessarily captured by nominal total household expenditure, failing to account for these differences in welfare, while adjusting for spatial price differences instead, leads to underestimating actual disparities in living conditions (D'Alessio, 2017, 18). In general, to the extent that the cost of living (the price level) and the presence of local amenities are positively correlated—see Brueckner, Thisse, and Zenou (1999)—the biases caused by failing to account for each factor go in opposite directions and offset each other, at least to some extent. Accordingly, one strategy to curb the bias would be to find a way to include the value of public goods in the consumption aggregate (see section 4.3) if adjusting for price differences; another would be to not adjust for spatial variation in prices at all.[58] However, more research is needed on this point before it can be said whether and in which cases avoiding any adjustment is reasonable.

[58] For the US, Glaeser (1998) notes that "mobility of labor generally means that higher prices are compensation for something else, so that additional compensation for those prices does not increase equity." Bodvarsson, Simpson, and Sparber (2015, 16) indicate that "persistent regional differences in wages, rents, and prices represent a compensating differential for regional differences in amenities."

5.2.5 Current practice

Figure 5.3 summarizes the current international practice in terms of spatial deflation of the consumption aggregate. About one-half of the countries examined do not document the methodological choices underlying the computation of a real welfare aggregate at all, or they leave out important details (such as the index used). Among those that document their choices, there is no convergence on one single method: Paasche is the single most popular consumer price index, but Laspeyres and Fisher indices are also used, and the indirect approach based on regional poverty lines also seems extremely popular. Several countries also indicate that they do not adjust the welfare aggregate for spatial differences in prices.

FIGURE 5.3. Current international practice on spatial deflation



SOURCE: Authors' elaboration of the dataset presented in appendix A.

Earlier on, it was argued that the choice of a Paasche price index as the recommended way to adjust the nominal consumption aggregate for spatial differences in prices stands the test of time. The discussion in this section, however, has raised a number of caveats. First, there are barriers to the full implementation of the recommendation. Our review of the current practice cannot point to the reasons why Paasche is not as prevalent as it could be—any combination of data constraints and lack of capacity may be responsible—but the practical difficulties in estimating it are likely part of the problem. Second, reliance on unit-values — based price indices is increasingly criticized. The arguments range from well-known issues (quality bias, measurement error, etc.) to newer insights (failure to capture important factors, most notably housing costs). Third, while no obvious solution has emerged, the demand for price data other than those currently available is common to many recent studies. The United Nations Statistics Division *Handbook on Poverty Statistics* (2005, 184) states, for instance, "Given the need for price data and the concerns about both unit values and relying on existing price collection efforts, it would be worthwhile for statistical agencies to invest more effort in gathering prices from local stores and markets and opinions about prices when their household surveys are fielded." Fourth, in countries where public goods provision is positively correlated with the cost of living, giving up deflation entirely may cause *fewer* distortions to welfare comparisons in practice. This argument is only reinforced by the first caveat: using a badly estimated deflator when none is needed can only make matters worse. This is still a debated topic, but it should be kept in mind in some contexts.

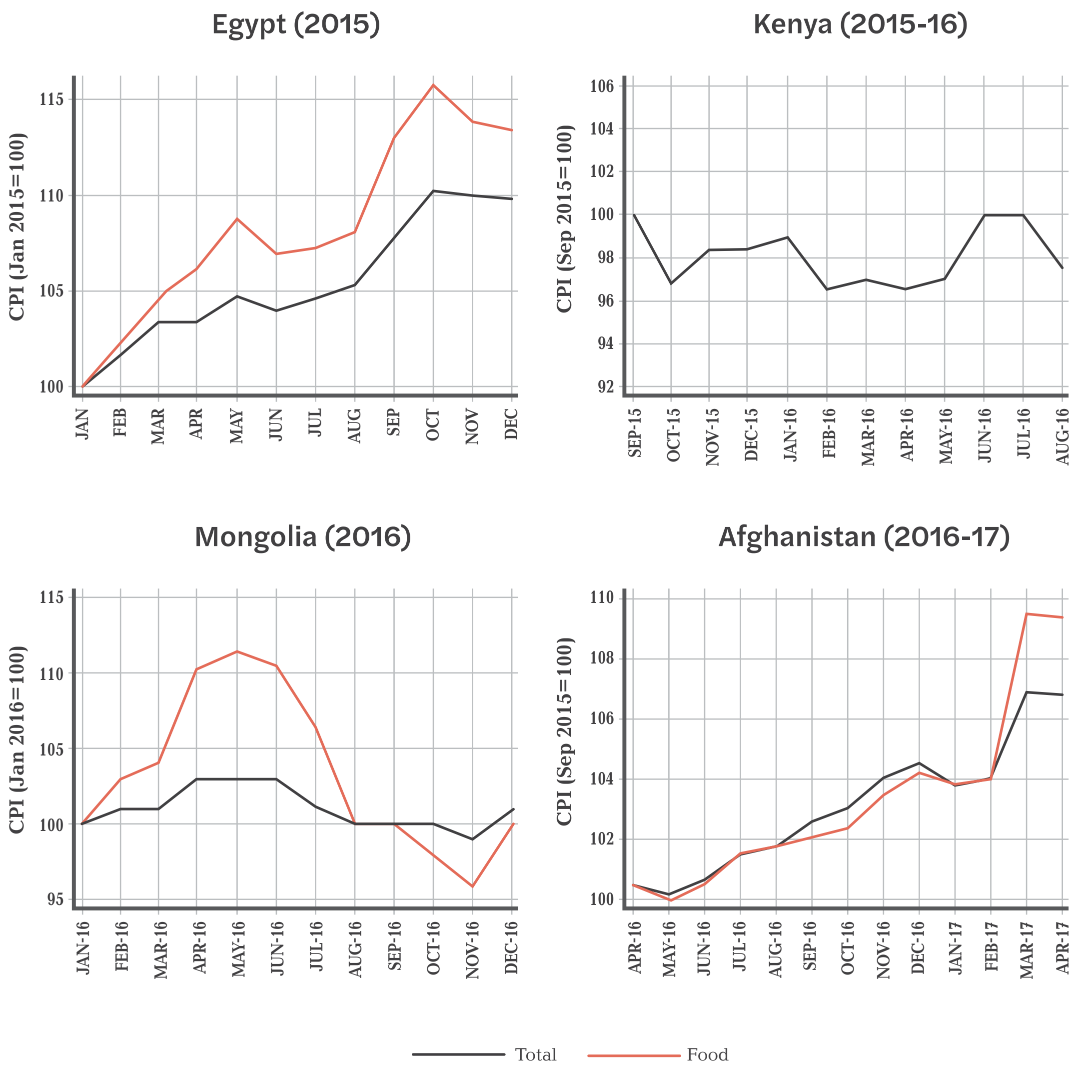
5.3 Temporal deflation

Price variation over time comes into play both *between* surveys, when comparing household expenditures over several years, and *within* the survey period, when ranking the living standards of households interviewed at different times during fieldwork. In the presence of inflation, all other things being equal, households interviewed at the beginning of the survey year, for instance in January, have greater purchasing power than households interviewed later on, say in December: as time goes by, the price level increases due to inflation, and the purchasing power decreases. To get interpersonal comparisons right, there is a need to adjust for intra-survey inflation.

A few examples of the dynamics of prices in countries that recently released poverty estimates, and for which monthly CPI data are easily accessible, are shown in figure 5.4. Examples range from the price soar throughout the whole survey period observed in Egypt and Afghanistan, where households interviewed toward the end of the survey period would have faced prices in the range of 10 to 15 percent higher than at its start; to sudden surges, as in Mongolia where interviews conducted in one particular month would have pictured a rather unique situation with respect to the rest of the year to short-term fluctuations throughout the year, as in Kenya. In each case, part of the differences between nominal consumption aggregates computed over the survey period would not be due to genuine differences in living standards, but to price variation.

Among DZ's recommendations is a mention of the need for temporal deflation when comparing two surveys some years apart (DZ, 40), while within-survey temporal deflation is not emphasized: "When we are working with a single cross-sectional household survey, the price variation is less *temporal* than *spatial*" (DZ, 9). Perhaps this is part of the reason why the available documentation for recent poverty estimates around the world is frequently silent on this methodological aspect: more than 56 percent of the countries examined as part of the database described in appendix A fail to document whether and how within-survey temporal deflation is implemented. However, especially in contexts where price volatility over time is significant, it is good practice to divide computed nominal consumption aggregates by a monthly temporal price index, according to the interview date of each household.[59]

FIGURE 5.4. Within-survey temporal price variation in selected countries



SOURCES: Authors' elaboration based on Egypt Statistics CPI series extracted on 07/10/2019 (http://www.capmas.gov.eg/Pages/StaticPages.aspx?pageid=5089); Kenya (2018, 32) based on KiHB 2015/16; Mongolia (2017, 63) based on HSES 2016; Afghanistan (2018, 110) based on ALCS 2016/17 and Central Statistics Organization CPI series (http://cso.gov.af/en/page/ict/5555/5353) extracted on 07/10/2019.

[59] In low- and middle-income countries, in particular, the volatility of food prices has caused concern after world food price spikes in 2008 and 2011 (World Bank 2012; Minot 2014), and food price seasonality is also significant in some contexts (Kaminski, Christiaensen, and Gilbert 2016)

One last remark is in order. There is a concern regarding whether official general price indices do a good job in capturing the inflation faced by the poor. The definition of a TCLI in equation (5.7) implies that, in general, the cost of living index number varies with the standard of living (the level of utility). The concern for the choice of the specific population group which the price index refers to, however, applies to all methods discussed in this section. More research is needed on this front before offering clear-cut recommendations.

5.4 How to deflate the consumption aggregate

In this section, we follow through to the final step of the discussion of price deflation and address the issue of how to actually perform the conversion of the *nominal* consumption aggregate (discussed in section 4) into *real* terms, once a suitable price deflator is available (be it spatial, temporal, or both).

Let X denote the nominal consumption aggregate—we can think of it as the sum of J components (e.g., two-digit COICOP expenditure groups, or even more aggregated expenditure components):

$$X = \sum_{j=1}^{J} X_j = X_1 + X_2 + \ldots + X_J \tag{5.14}$$

For instance, X_1 could be food expenditures, X_2 housing expenditures, and so on. Let CPI denote a deflator (could be spatial, temporal, or both), and let us assume that price indices for as many categories as specified in equation (5.14) are available: CPI_1 is the price index for the first category (in our example, food items), CPI_2 is a price index for the second category (housing), and so forth. In general, the overall CPI can be thought of as a function of the sub-indices CPI_j , which we write as follows $CPI = f(CPI_1,...CPI_n)$, typically $f(\cdot)$ is some arithmetic or geometric average of the elementary indices. To deflate the consumption aggregate, the analyst has two options:

$$\widehat{X} = \frac{X}{CPI} \tag{5.15}$$

$$\tilde{X} = \sum_{j=1}^{J} \frac{X_j}{CPI_j} = \frac{X_1}{CPI_1} + \frac{X_2}{CPI_2} + \ldots + \frac{X_n}{CPI_n} \tag{5.16}$$

Both $\widehat{X}$ and $\tilde{X}$ qualify as *real* consumption aggregates. In equation (5.15), $\hat{x}$ is obtained by dividing the total nominal household expenditure by the (aggregate) CPI. By contrast, in equation (5.16), $\tilde{X}$ is obtained by adding up each component of the consumption aggregate, deflated by the corresponding subindex.[60] In general, the two methods give different answers.

[60] For the sake of keeping the discussion simple, we are assuming away potential problems due to the inconsistency between CPIs and the consumption aggregate. The issue is important, as explained by ILO (2004, 35): "The data collected on prices and the data collected on household expenditures must be mutually consistent when measuring real consumption. This requires that both sets of data should cover the same set of goods and services and use the same concepts and classifications. Problems may arise in practice because the price indices and the expenditure series are often compiled independently of each other by different departments of a statistical agency or even by different agencies." The details of this discussion are beyond our scope here.

Surprisingly, the literature does not provide specific guidance on which one is the better approach. Our recommendation is to use $\widehat{X}$. This is based on two main arguments. First, while $X/\widehat{X} = CPI$, in general $X/\tilde{X} \neq CPI$. In other words, only the method in equation (5.14) ensures the "reversibility" of the deflation of the nominal aggregate, using the kind of data that are typically available to the general public—a desirable property for official statistics, if only for its transparency and ease of communication. More importantly, the use of equation (5.15) implies that all budget shares, defined as X_j/X in nominal terms, would be distorted, that is, they would be different from the budget shares calculated in real terms ($\tilde{X}_j/\tilde{X} \neq X_j/X$). This is a major shortcoming, given the role that budget shares play in welfare analysis—they describe the households' consumption patterns and are used for a broad number of econometric exercises.

To illustrate the unwelcome consequences of item-specific deflation, as in equation (5.15), we will use economic historian Stefano Fenoaltea's ingenious argument.[61] Fenoaltea (2020) imagines a school where class photographs are being taken. Each class can be likened to a household budget, the height of each student representing that household's various expenditures. In each class picture, the ranking of students by height (the relative magnitudes we are interested in) is clear, as they are neatly arranged from tallest to shortest in front of the camera: we can easily pick out, for instance, who is the tallest boy or girl in the group. Now imagine we want to reconstruct a picture of the entire student body by piecing together the individual class pictures. As it happens, each class has had their picture taken at a different distance from the lens, so that our reconstruction of the school-wide picture will require some "rescaling" in order to come out right. We can take one class as our reference point, but all other class pictures must be zoomed in or out, according to each group's distance from the camera. However laborious the process, the final student body picture should preserve the within-class rankings (each class' tallest boy or girl should still appear taller than their classmates in the school-wide picture); it will do so if an entire class is scaled in the same proportion (all the household's expenditures are deflated with the same price index), and not if the students are scaled in different proportions (if the household's various expenditures are deflated with category-specific deflators). In summary, unless all components of the nominal consumption aggregate are deflated by the same deflator, the deflated household-level budget shares (the relative heights of classmates) turn out to be wrong.

One final remark is on the relationship between temporal and spatial deflation. There is an obvious problem with determining the sequence in which the temporal and spatial adjustments should be performed: the order in which the two operations are performed matters for the final result (the real consumption aggregate). The most logical solution—calculating a *monthly* spatial price index to perform both adjustments simultaneously—is unlikely to be applicable in practice due to data limitations. In the absence of guidance from the literature on this point, or contextual factors that make a specific sequence more convenient, the solution adopted is ultimately a matter of convention.

[61] Fenoaltea is concerned with industrial production rather than with consumption, but faces the same problem as the welfare analyst does. Taking x_j in equation (5.13) to represent the contribution to total value added of industry *j*, Fenoaltea (1976) argues that every x_j, and so industry's total value added, should be deflated by the same ("general") deflator (*CPI*), not industry-specific value-added deflators CPI_j. Fenoaltea (2020), the source of the quote, illustrates the same point in less technical terms.

6. Adjusting for Household Size and Composition

Welfare, living standards, and poverty are *individual* attributes (Deaton 1997, 223; Ravallion 2016, section 3.3). However, expenditure and consumption data are typically collected at the *household* level. Most household surveys are simply silent on how total household expenditures are allocated among household members.[62] The mismatch between the theoretical unit of analysis (the individual) and the statistical unit (the household) has important implications for the work of analysts interested in interpersonal welfare comparisons, both *within* the household and *across* households.

The first topic, the intra-household distribution of living standards, is of great interest to both academics and policy makers, given its implications for gender inequality and child well-being. It is an active field of research that draws on consumer theory to infer individual consumption based on household-level data, often in ingenious ways. Summarizing this vast literature would be beyond the scope of this section (useful reviews are found in Chiappori and Meghir 2015 and World Bank 2018). Suffice to say that the available methods to estimate intrahousehold resource shares have yet to be integrated in routine poverty monitoring. In the context of poverty and inequality measurement in the general population, the issue of intra-household inequalities is essentially assumed away, and household members are conventionally assigned an equal share of household resources.[63] Recently, attempts to estimate the extent of intra-household inequality using a relatively straightforward methodology that relies on widely available individual income data have been successful (Ponthieux 2017). While a definitive best practice has yet to emerge, this is an area where the "status quo" of inequality and poverty measurement is expected to change in the near future.

The second topic, comparing individuals across households, cannot be eluded. Even assuming household members get an equal share of the household's resources, there is still a need to rank households consistently, and total expenditure is not a satisfactory metric for doing so. Take household size: the number of "mouths to feed" obviously matters when comparing the welfare of different households. A one-person unit spending 3,000 rupees per month is clearly better off than a two-person unit spending the same 3,000 rupees per month. A slightly subtler point concerns the way in which needs increase with household size: in order to be as well off as a single-person household, would a two-person household need to spend exactly two times as much? One can argue that the answer is negative: goods like

---

[62] A few recent exceptions are found in De Vreyer and Lambert (2018), D'Souza and Tandon (2019), Mercier and Verwimp (2017), and Santaeulàlia-Llopis and Zheng (2017), although the surveys used by these studies limit individual measurement to a few components of consumption (often food).

[63] Ignoring intra-household inequality leads to underestimation of overall inequality, while the effect on poverty is ambiguous (Haddad and Kanbur 1990; Kanbur 2018; De Vreyer and Lambert 2018).

housing, vehicles, and other household amenities do not need to be duplicated in order to be consumed by more than one person, and can instead be shared at little or no additional cost (much like certain public goods). If this is the case, economists say that there are economies of scale to living together, which should be accounted for when making comparisons. A final issue concerns household composition. Consider two households with the same total expenditure and the same size, but whose members have different demographic characteristics: for instance, two adults as opposed to one adult and one child. Should the two households (and the individuals in them) be seen as equally well off? Again, there are good reasons to respond in the negative, given that children and adults have different needs.

Conceptually, adjusting for differences in household size and composition can be seen as just another type of deflation, not dissimilar from an adjustment for price variation (Ravallion 1994, 20; DZ, 47). In this context, the "deflator" is some indicator of household size, which may be as simple as the number of household members, yielding expenditure per capita as the welfare indicator, or as elaborate as a measure of both economies of scale and household composition. In the latter case, the measure is called an equivalence scale.

An *equivalence scale* is a device to convert a household's specific demographic profile into a number of "equivalent adults." Instead of simply counting for one, irrespective of age, gender, and other characteristics, each individual of the household is assigned a weight. Typically, a (male) adult is chosen as a benchmark, his weight is set to 1; other individuals receive a weight that is less than 1 according to how much lower their cost is, in terms of consumption expenditure, relative to a male adult. Children, for instance, might receive a weight of 0.25, or 0.33, or 0.50, implying that a child is expected to "cost" 1/4, 1/3, or 1/2 of the reference adult male, respectively. A similar reasoning may be applied to age groups and genders: females may consume less than males, older people less than younger ones, and so on. On top of this, an equivalence scale may include an adjustment for economies of scale, which can be implemented in various ways, but that in general entails that household size is made to increase non-linearly with every additional member (regardless of whether they are adults or not). The sum of weights assigned to individual members of a household, plus any adjustment for economies of scale, yield an equivalent household size (ES). Its interpretation is straightforward: a household with an equivalent size of, say, 3.5 needs to consume 3.5 times as much as a single (male) adult in order to enjoy the same living standard.

Unfortunately, there is no objective way to determine how individuals should be weighed, nor is there a consensus on a scientific way to account for economies of scale (Ferreira and Ravallion 2011, 608). A number of widely used methods are available, each with its own drawbacks. DZ identify three main approaches to the specification of an equivalence scale—behavioral, subjective, and arbitrary—and this is still a useful categorization for a review of what is available. According to the behavioral approach, the analyst can estimate an equivalence scale based on data collected from observing consumer behavior, that is, by investigating how household consumption varies with household size and composition (FAO 2005c). Despite the theoretical appeal of this approach, experts tend to advise against its use (see, for instance, Coulter, Cowell, and Jenkins 1992a; Deaton 1997, 241 – 70; Haughton and Khandker 2009: 29 – 30), and its complexity has arguably hindered its popularity. A second approach, initiated by van Praag (1971), is based on directly questioning households to construct an equivalence scale based on self-reported assessments.[64] DZ, together with other authors (see, for instance, Coulter, Cowell, and Jenkins 1992a), do not recommend this approach, as it suffers from a number of measurement issues and econometric drawbacks, and it requires data collected through ad hoc questionnaires. Twenty years later, it seems safe to conclude that the subjective approach has not gained much traction. What is left is the *arbitrary* approach, which is the one recommended by DZ. The term "arbitrary" may conjure the idea of capricious discretion, but this would not be a fair characterization. This approach does use normative weights, but these hinge, at least to some extent, on objective, non-arbitrary parameters.[65]

Many of the arbitrary (per DZ terminology) equivalence scales present in the literature are described by the following general formula, originally suggested by Cutler and Katz (1992, 548), and subsequently adopted by the US National Research Council (1995, 161):

$$ES = (A + \alpha K)^{\theta} \tag{6.1}$$

where *ES* stands for "equivalent household size," that is, the number of adults that the members of a household are equivalent to; A denotes the number of adults (however defined) in the family; K is the number of children; α is the parameter, typically lower than 1, that captures the relative cost of a child to an adult; and θ is the parameter that captures economies of scale ($\theta = 1$ implies no economies of scale, $\theta = 0$ represents a purely hypothetical scenario of complete sharing, where each individual is assumed to consume the total consumption of the household). Equation (6.1) is easy to interpret: it first calculates the number of adult equivalents ($A + \alpha K$) and next raises the results to the power of θ to account for economies of scale for larger families. The choice of the values for α and θ is critical, as we shall see, and ultimately discretionary. Nevertheless, DZ suggest that "a case can be made for the proposition that the current best practice is to use equation (6.1) for the number of adult equivalents, simply setting α and q at sensible values" (p. 52).[66]

The so-called *Organisation for Economic Co-operation and Development (OECD) equivalence scales* are another popular incarnation of the arbitrary approach:

$$ES_{OECD-I} = 0.3 + 0.7A + 0.5K \tag{6.2}$$

$$ES_{OECD-II} = 0.5 + 0.5A + 0.3K \tag{6.3}$$

where A is for adults (individuals ages 14 or more) and K is for children (individuals ages 13 or less).[67] Equation (6.2) is commonly referred to as "Oxford scale" or "old OECD scale" or "OECD-I" scale; it gives a weight equal to 1 to one adult (regardless of gender), of 0.7 to any other additional adults, and of 0.5 to each child. This formula accounts for both household

[64] As the approach was mainly developed in Leyden (in the Netherlands), the "subjective approach" is also called the "Leyden approach" (van Praag and Warnaar 1997, 260).

[65] Take, for instance, the equivalence scales that have been suggested since the 1980s—see, among others, Buhmann et al. (1988), Coulter, Cowell, and Jenkins (1992b), Hagenaars, De Vos, and Zaidi (1994), Atkinson, Rainwater and Smeeding (1995) —based on norms set by outside experts (an expert might be a scholar familiar with the country or an institution).

[66] Interested readers are referred to Ravallion (2016, 168) who provides a more general formula for equation (6.1).

[67] Both equations (6.2) and (6.3) are subsumed under equation (6.1). Set $\theta = 1$, and replace A by $1 + \beta(A - 1)$. For $\beta = 0.7$ and $\alpha = 0.5$ we get equation (6.2); for $\beta = 0.5$ and $\alpha = 0.3$ we obtain equation (6.3).

composition (it treats children and adults differently) and for economies of scale (one adult counts for one, any additional adult counts for less than one). Hagenaars, De Vos, and Zaidi (1994) propose the so-called "OECD-modified equivalence scale," equation (6.3), currently used by the Statistical Office of the European Union (Eurostat) in its calculations of monetary poverty for the EU.[68]

A number of recent publications have turned to a simpler formula and tend to use the *square root scale*, or Luxembourg Income Study (LIS) scale, which takes the square root of the number of household members *N*:

$$ES_{LIS} = \sqrt{N} \tag{6.4}$$

Equation (6.4) is a version of equation (6.1), where $\alpha = 1$ and $\theta = 0.5$: it makes no allowance for differences in household composition but adjusts for economies of scale (according to it, a household of four people needs to spend twice as much as a single person).

Table 6.1 provides a summary of the discussion so far and compares equivalent household size (panel a) and household expenditure per adult equivalent (panel b) computed using different equivalence scales, for a few typical household configurations. Comparing numbers *row-wise* helps to gauge the sensitivity of results to the choice of different scales. To illustrate, take the row corresponding to "2" adults: the equivalent size ranges from 2 (column 1, per capita adjustment) to 1.4 (column 6, corresponding to the square root scale). Similarly, assuming a total household expenditure equal to 2,000 rupees, panel b shows that per-equivalent household expenditure ranges from 1,000 rupees (column 1) to 1,429 rupees (column 6). Reading the table *column-wise* is informative on how household equivalent size varies as household composition varies, given the choice of a specific scale. For instance, adding an adult to a single-person household increases household equivalent size by 1 if using a per-capita adjustment (column 1), and by 0.4 if using the square-root scale (column 6). Similar added-member effects can be explored for the addition of the first, second, third child, and so on.

The choice of the equivalence scale affects both poverty levels (how many people are poor), and the poverty profile (who are the poor). For example, equation (6.2), which weighs children more heavily than equation (6.3), will lead to a greater proportion of children being classified as poor (O'Higgins and Jenkins 1990; de Vos and Zaidi 1997: table 3). As consequential as it is, there is simply no right choice: no scale can be deemed superior to all others. This has two main implications. First, any recommendation on the "best" scale to use is necessarily open to criticism. DZ conclude their analysis with a two-pronged recommendation, namely (1) "continue using per capita expenditure" (p. 54), and (2) improve on the per capita measure using the equivalence scale in equation (6.1), whereby the choice of the parameters depends on the country. For poor economies, DZ recommend setting α in the range between 0.25 and 0.33, and θ =0.9 (table 6.1, column 2). The rationale is that in poor, agricultural countries children are relatively "cheap" compared to adults, and given the high incidence of rival goods (food, first of all) in the expenditure patterns, there is little room for economies of scale. For richer economies, DZ recommend to set $\alpha = 0.5$,

[68] The OECD-modified scale has been criticized by researchers from Eastern Europe, for example, Éltetõ and Havasi (2002) and Szulc (2006) because it implies too large of economies of scale in consumption.

and $\theta = 0.75$ (table 6.1, column 3). A second implication is the need to carry out *sensitivity analysis*: if economic theory fails to order the different scales, it is important to check the empirical robustness of results. The analyst should make sure that the poverty profile is not too sensitive to the choice of the scale, or of its parameters. This will be discussed further in Section 8.

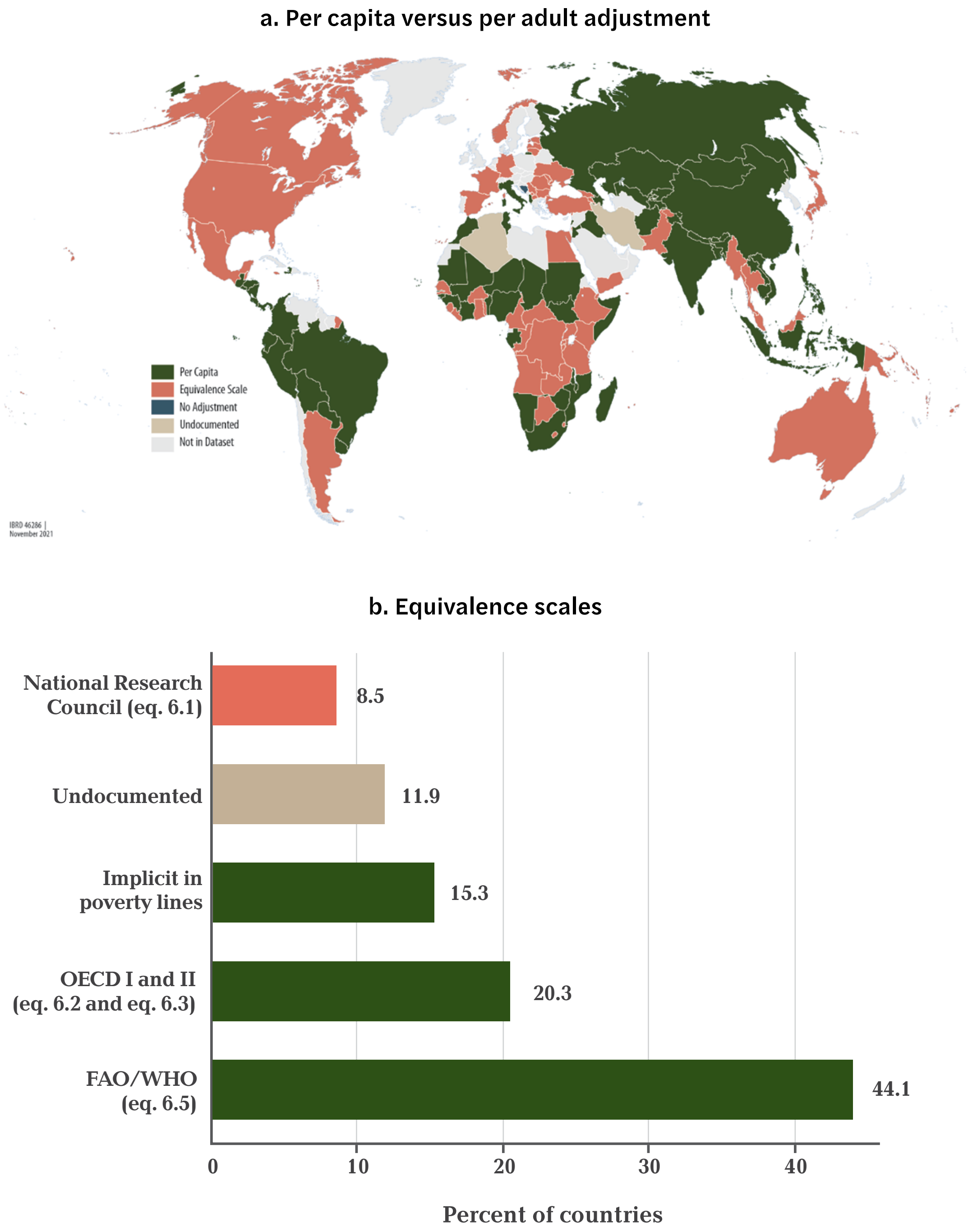
TABLE 6.1. Household size, household equivalent size (ES), and household equivalent expenditure

Household composition	eq. 6.1			eq. 6.2	eq. 6.3	eq. 6.4
	α = 1, q = 1	α = 0.25, q = 0.9	α = 0.5, q = 0.75			
	Per capita	DZ recommendation for poor economies	DZ recommendation for rich economies	OECD-I	OECD-II	Square root
a. Equivalence scale						
Equivalent household size (ES)						
1 adult	1	1	1	1	1	1
2 adults	2	1.9	1.7	1.7	1.5	1.4
2 adults, 1 child	3	2.1	2.0	2.2	1.8	1.7
2 adults, 2 children	4	2.3	2.3	2.7	2.1	2
2 adults, 3 children	5	2.5	2.6	3.2	2.4	2.2
b. Household expenditure per adult equivalent						
1 adult	2,000	2,000	2,000	2,000	2,000	2,000
2 adults	1,000	1,053	1,176	1,176	1,333	1,429
2 adults, 1 child	667	952	1,000	909	1,111	1,176
2 adults, 2 children	500	870	870	741	952	1,000
2 adults, 3 children	400	800	769	625	833	909

NOTE: The columns labelled "DZ recommendation for poor and rich economies" refer to the ranges for α and θ mentioned on page 52 of DZ's Guidelines.
SOURCE: Our elaboration.

Figure 6.1 describes the current international practice in terms of adjustment for household size and composition (regardless of whether the welfare measure is based on consumption or income). Panel a shows a prevalence of equivalence scales in North America and Europe, while the rest of the world does not seem to be neatly divided into "camps" of subscription to either approach. In panel b, we zoom in on the more than 50 countries that use equivalence scales and examine the extent to which the scales reviewed in this section, equations (6.1) — (6.4), are used in practice. Two results stand out: first, DZ's recommendation (the red bar, corresponding to equation (6.1) in this section) turns out to be the *least* practiced among the countries covered in our database. Why we do not know. Second, the equivalence scale that is most widely used in practice, labelled FAO/WHO, is not explicitly discussed in DZ, nor is it among the scales adopted by major international organizations: the reference to FAO and WHO simply indicates that the scale is based on nutritional requirements and does not seem to univocally identify any one formula—see the examples presented in table 6.2, as well as Waid et al. (2017) for Bangladesh.

FIGURE 6.1. Different countries, different equivalence scales



SOURCE: Authors' elaboration on the dataset presented in appendix A.

The so-called FAO/WHO methodology consists in computing the adult equivalent weight of any given age-gender group as the ratio of the energy requirement of an individual belonging to the group, and that of an adult male:

$$ES_{FAO/WHO} = \sum_i \sum_j N_{ij} \frac{ER_{ij}}{ER_{adult\ male}} \tag{6.5}$$

where N_{ij} denotes the number of household members in age range i and of sex j, and ER_{ij} denotes the corresponding energy requirements. Table 6.2 illustrates. For the case of Argentina in 2016, we count 24 age brackets ($i = 1,...,24$), and energy requirements are calculated separately by gender ($j = 1,2$). Note that the ER_{ij} coefficients are *estimated* based on the FAO/WHO technical tables, which provide the minimum calorie intake for individuals of different age, gender (with further distinctions for pregnant or breastfeeding women, working children, etc.), body size (height and weight), and physical activity level.[69] The energy requirement for the reference male adult ($ER_{\text{adult male}}$), indicated with an asterisk in table 6.2, serves as a *numeraire*, that is, it normalizes each individual ER_{ij} to a number in the interval [0,1]. According to the Argentinian methodology, the cost of a 6- to 9-month-old infant is 28 percent that of a 30—60-year old adult. The corresponding figure for adult women is about 77 percent. The figures reproduced at the bottom of table 6.2 for countries in Africa and South Asia are derived in the same way as the Argentinian coefficients, but energy requirements are not documented.

TABLE 6.2. Examples of equivalence scales based on FAO/WHO nutritional requirements

	Male		Female	
Age	**Energy requirement (kcal/person/day)**	**Equivalence scale**	**Energy requirement (kcal/person/day)**	**Equivalence scale**
Argentina 2016				
6–9 m	776	0.28	776	0.28
9–12 m	952	0.35	952	0.35
1	1,030	0.37	1,030	0.37
2	1,277	0.46	1,277	0.46
3	1,409	0.51	1,409	0.51
4	1,518	0.55	1,518	0.55
5	1,643	0.60	1,643	0.60
6	1,760	0.64	1,760	0.64
7	1,813	0.66	1,813	0.66
8	1,865	0.68	1,865	0.68
9	1,910	0.69	1,910	0.69
10	2,192	0.79	1,918	0.70
11	2,255	0.82	1,986	0.72
12	2,347	0.85	2,051	0.74
13	2,472	0.90	2,089	0.76
14	2,650	0.96	2,100	0.76
15	2,760	1.00	2,116	0.77

[69] Weissel and Dop (2012) illustrate the methodology with a simple artificial example.

TABLE 6.2. (Continued)

Age	Male: Energy requirement (kcal/person/day)	Male: Equivalence scale	Female: Energy requirement (kcal/person/day)	Female: Equivalence scale
16	2,828	1.03	2,111	0.77
17	2,881	1.04	2,124	0.77
18–29	2,826	1.02	2,106	0.76
30–45 (*)	2,758	1.00	2,111	0.77
46–60 (*)	2,750	1.00	2,090	0.76
61–75	2,288	0.83	1,860	0.67
75+	2,050	0.74	1,750	0.63
Burundi 2013, Liberia 2016, Republic of Congo 2011				
< 1		0.27		0.27
1–3		0.45		0.45
4–6		0.61		0.61
7–9		0.73		0.73
10–12		0.86		0.73
13–15		0.96		0.83
16–19		1.02		0.77
20–50 (*)		1.00		0.77
51+		0.86		0.79
Micronesia 2013, Fiji 2008, Kiribati 2006, Nauru 2012, Palau 2006, Samoa 2013, Seychelles 2013, Tuvalu 2010, Vanuatu 2010				
0–14		0.50		0.50
15+ (*)		1.00		1.00

NOTE: Shaded cells indicate that energy requirements were not reported in the official documentation.
SOURCES: Argentina (2016, 2); Burundi (2015, 36); Liberia (2017, 9); Republic of Congo (2017, 112); Micronesia (2017, 78); Fiji (2011, 78); Kiribati (2010, 19); Nauru (n.d., 24); Palau (2008, 15); Samoa (2016, 30); Seychelles (2016, 2); and Vanuatu (2013, 9).

To recap, the choice of an equivalence scale is ultimately based on value judgements, on which differing opinions are expected (Lanjouw and Ravallion 1995). This was true in the late 1980s when the discussion on equivalence scales gained momentum, was confirmed during the 1990s when many international agencies began to take sides and to adopt specific equivalence scales, and still holds true today. Our reading of the literature is that DZ's first recommendation is still valid: analysts are advised not to abandon the use of per capita consumption, given that "20 years of experience with per capita expenditure has given analysts a good working understanding of its strengths and weaknesses, when it is sound (in most cases), and when it is likely to be misleading (for example, in comparisons of the average living standards of children and the elderly)" (DZ, 48). The number of years has now increased to 40, but this is the only required amendment to DZ's conclusion. Another argument in favor of per capita expenditure is the consensus for using it as *the* adjustment for international comparisons. Ferreira et al. (2015, 11) explain why all consumption and income measures in *PovcalNet* are in per capita terms.[70] Similarly, Batana and Cockburn (2018, 2): indicate that "if the use of equivalence approaches is suggested to provide a more accurate measure of poverty, the global monitoring of poverty still mainly uses per capita consumption as the core measure of welfare."[71]

The second of DZ's recommendations—using equation (6.1) with parameters set at reasonable values, as indicated in table 6.1—also remains valid. In fact, it seems that a large proportion of countries did embrace the idea of adjusting for household size and composition using the "arbitrary" approach, but that many of them have opted for different parametric specifications, namely based on nutritional energy requirements. While the multiplicity of formulations falling under the FAO/WHO label represent an attempt to adopt a more "scientific" approach to the computation of equivalent size, they have some important drawbacks (FAO 2005a). One is that relative caloric requirements do not necessarily correspond to relative costs (the fact that an adult needs three times as many calories as a child does not imply that an adult also costs three times as much; that depends on diet composition and relative prices).[72] Another problem is coverage: the relative needs of two individuals do not depend on food alone. To be sure, the literature has not yet examined the merits of "FAO/WHO scales" in detail, but what seems clear is that they do not settle the issue of choosing the best adjustment for household size and composition.

In his last book, Tony Atkinson did not hesitate to recommend the use of equivalence scales: "it seems impossible to ignore the differing needs of households made up of different numbers and people of differing ages, even if we cannot agree on just how large those differing needs are." (Atkinson 2019, 78—79). His argument includes a call for attention to aspects not usually considered by adjustments for individual needs: "One issue that should be better explored is the way in which equivalence scales can reflect the fact that people with disabilities face higher costs of living than other people to achieve the same quality of life." (Atkinson 2019, 79). Atkinson's plea to focus on the barriers to a life well lived confirms, though only implicitly, that an exclusively food-based approach to equivalence scales is debatable, despite its popularity. Overall, the choice of adopting equation (6.1) remains sound, both conceptually and empirically. It is good practice to use sensitivity analysis to explore the consequences of choosing a specific scale (see section 9 for a detailed discussion).

---

[70] PovcalNet is a computational tool that allows analysts to replicate the World Bank's calculations of absolute poverty in the world (http://iresearch.worldbank.org/PovcalNet).

[71] See also the special issue hosted in volume 13 of the *Journal of Economic Inequality*, 2015.

[72] Any equivalence scale can be seen as an extended TCLI (section 5.1.2), that is, as the ratio of the cost functions of two households enjoying the same level of utility, facing the same level of price, but with different demographic compositions (see, for instance, FAO 2005a, Lewbel 2006, or Ray 2018). The FAO/WHO scales are not defined in the metric of expenditures, but in that of calorie consumption. Additionally, FAO (2005b, 7) notes that equivalence scales based on nutritional requirements (1) "rule out economies of scale by definition, as they only take into account nutritional differences among household members with respect to good (food) which is 'private' within the household," and (2) "[are] based on the determination of 'subsistence' levels there is no guarantee that the same scale would prevail at higher levels of well-being."

7. Data Issues

No dataset is flawless; even when working with high quality data, one must usually contend with issues such as missing data, extreme values, inaccurate or implausible records, and more. What to do in these recurring situations is not always clear. One might be tempted to try and "fix" all perceived data flaws, based on some prior knowledge of the plausibility of observations. In fact, the practice and scholarship of the last 20 years point the opposite way, and lean toward preventing data issues ex ante, rather than correcting them ex post. This is made apparent by the efforts poured over the years into the improvement of data collection methods, resulting in the expansion of the survey methodology literature (see Appendix E), and in the disruption of interviewing technology, among other things. It is becoming more and more frequent for statistical institutions to embed data quality safeguards into the early stages of data collection; for instance, checks for out-of-range values and flags for missing data can be hard-coded into Computer Assisted Personal Interviewing (CAPI) systems. With the widespread adoption of these methods, measurement error is expected to decrease significantly (Caeyers, Chalmers, and DeWeerdt 2012; Fafchamps et al. 2012). In this light, ex post corrections are to be seen as a last resort, one the analyst may be forced to turn to when something major has gone wrong somewhere up the chain.

This consideration, while necessary, brings us back to square one. When all is said and done, and final datasets are passed on to the analyst, they are typically at least inspected for flaws, and often—though certainly not always—some adjustments are considered necessary. Mishandling data at this stage has the potential to do great damage. In actuality, data cleaning—a catch-all term describing the identification and treatment of all sorts of data imperfections—is commonly framed as a "pre-analytical" task and is easily overshadowed by topics that are seen as more technical and consequential. As a result, it is often poorly documented. On the contrary, it can have a massive impact on the statistics of interest in welfare analysis, and it should be seen as part and parcel of the construction of a welfare aggregate, on par with the other methodological choices discussed in this document. This is especially true when comparability over time and across countries is a priority. Scrutinizing the fine details of consumption aggregation while disregarding the treatment of data issues can be like "watering the garden while the house is on fire" (Lokshin 2018).

The importance of this part of the analyst's job did not escape DZ: Chapter 3 of the *Guidelines* opens with a discussion of data cleaning. While they refrain from sharing specific guidelines, the authors do deliver a general recommendation: "It is of the greatest importance that the analyst check (...) for the presence of 'gross' outliers, typically by graphing the data, for example using the 'oneway' and 'box' options in Stata" (p. 23).

In this section, we expand on DZ's recommendation and cover two topics of major relevance: missing data and outliers. More than anywhere else in this document, our intended audience is broad and encompasses data producers as well as analysts. Statistical officers are (or should be) aware of the priorities of data users as they tackle data collection and processing, and welfare analysts are (or should be) cognizant of the minute details of the data generating process as they evaluate data cleaning strategies. In practice, the two are (or should be) in the same room when the key decisions regarding data issues are taken. Anthony Atkinson's posthumous book includes a whole chapter on *The Key Role of Data*, to emphasize that welfare analysts must not take data for granted, and instead should be encouraged to adopt an all-round (AA) approach, which requires cooperation and consultation with data producers and sampling specialists (Atkinson 2019, 144—145; Brandolini and Miklewright 2020).

Accordingly, the objective of this section is broader and more instructional than elsewhere in these guidelines. The goal is not to equip the analyst with a step-by-step protocol for handling data issues: there is no consensus on best practices in this difficult terrain, where so much depends on the context. Instead, what this section offers is a conceptual framework for understanding important issues such as unit nonresponse (section 7.1), item nonresponse (section 7.2), and outliers (section 7.3), as well as some references to solutions that are commonly applied in practice.

7.1 Unit nonresponse

A nonrespondent unit (an individual or a household) is any unit for which survey data are not obtained because of refusal (persons who adamantly refuse to be interviewed), noncontact (as for the case of persons who reside at home but are temporarily away), and a number of other big and small reasons, from interviewer fraud all the way to bad weather (Platek 1977; Brick and Montaquila 2009, 164). In practice, unit nonresponse results in a missing record, such as a household was supposed to be in the database, but is missing instead. To deal with this problem, one must first understand why it is a problem in the first place, and its implications for final estimates. The goal of this section is to help the reader do just that.[73]

Even with state-of-the-art data collection practices, unit nonresponse nearly always occurs. One example is paradigmatic. Miller and Aharoni (2015) report survey nonresponse in US military populations: this is a world where the answer to everything is "Sir, yes Sir!," so one would expect response rates to be close to perfection, that is, 100 percent. Results show response rates as low as 9 percent. If this is so, then the rest of us really should expect to have to deal with low response rates. As Särndal and Lundström (2005, 1) put it, "nonresponse is a normal but undesirable feature of a survey." How normal is it, and why is it undesirable, exactly? Figure 7.1 addresses the first question, showing how nonresponse rates for income and expenditure surveys vary (1) across countries (panel a), and (2) over time (panel b). Panel a gives a general idea of the international heterogeneity in nonresponse rates: they are close to zero for some surveys (the case of Jordan stands out—see Palaniswamy and Vishwanath 2019), and as high as 94 percent for others.[74] Hlasny (2020) investigates nonresponse rates

---

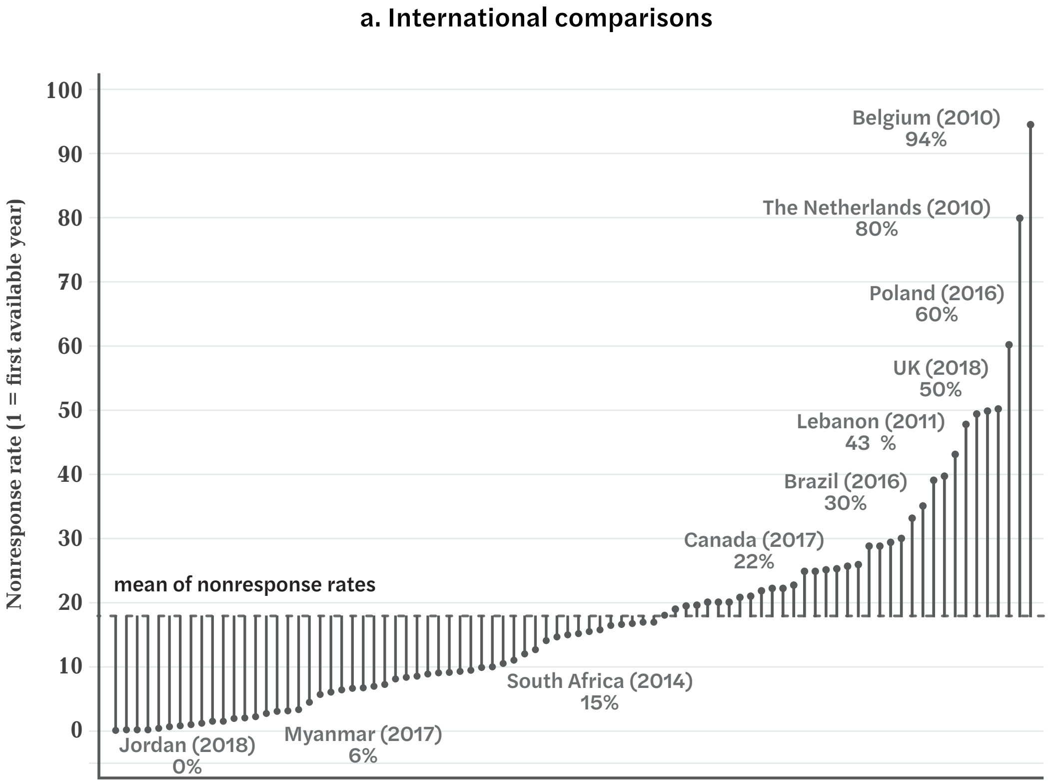
[73] This section has benefitted from suggestions by Carlos Rodríguez-Castelán, Kristen Himelein, and Juan Muñoz.

[74] The graph combines income and expenditure surveys spanning almost a decade (2010 to 2018), and it reports nonresponse rates as available from official publications. Lohr (2009, 355) points out that response rates can be easily manipulated to show high values, which would convey the idea of high quality survey data. Incentives exist to unethical reporting of inflated response rates, or equivalently, deflated nonresponse rates (Groves et al. 2009, 184). The American Association of Public Opinion Research (2016, 60) provides widely accepted guidelines for defining response rates.

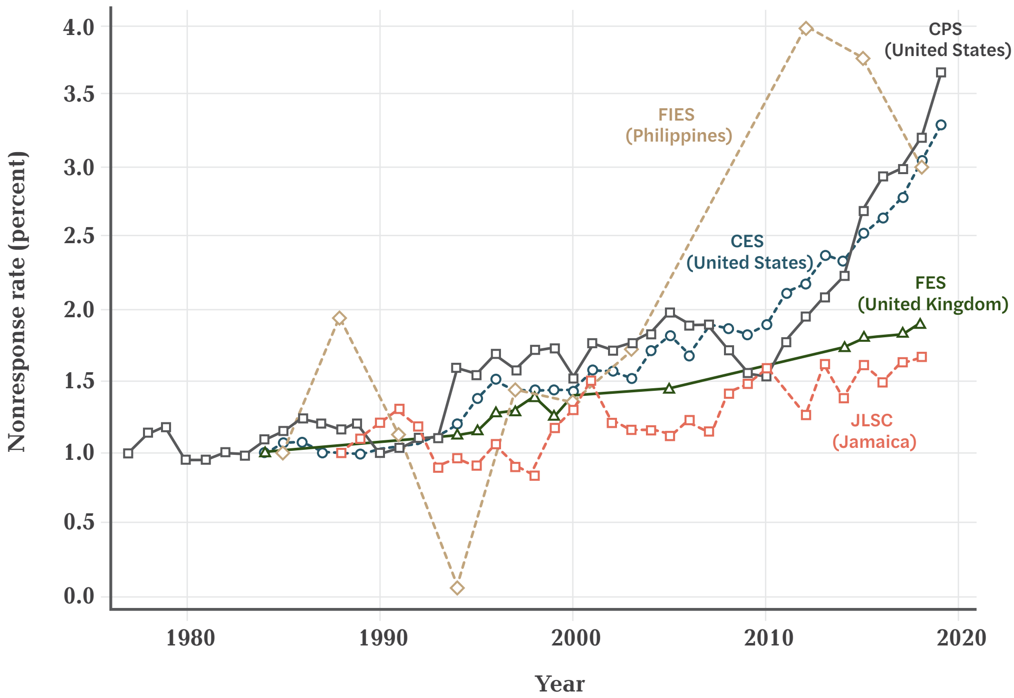
at the level of subnational regions for 38 countries and reports a similarly high international heterogeneity. Overall, the evidence suggests that nonresponse rates around 20 — 30 percent are "normal" in modern household budget surveys.

Panel b of figure 7.1 shows time trends of nonresponse rates in selected countries. Since the time DZ was written, greater challenges have emerged in assuring compliance with surveys. Meyer, Mok, and Sullivan (2015; 4) have recently argued that household surveys are in crisis, in part due to rising rates of unit (and item) nonresponse. For each time series shown in the figure, the initial nonresponse rate is set equal to 1: no attention should be paid to levels, therefore, but only to changes. The evidence is clear-cut: the US Current Population Survey (CPS), used for the official US poverty rate, shows an increase by a factor of 3.7, with a marked acceleration during the last decade. The US Consumer Expenditure Survey (CES), which provides the weights for the calculation of official US inflation, follows a similar dynamic. In general, all series considered in figure 7.1 are trending upward (although at different paces); at a minimum, nonresponse rates have doubled over the course of the last 30 years or so (see also de Leeuw and de Heer 2002; Luiten, Hox, and de Leeuw 2020).

FIGURE 7.1. Nonresponse rates (percent) around the world and over time



b. Time trends



SOURCES: Panel a: our estimates based on the database described in appendix A, and LIS data (https://www.lisdatacenter.org/frontend#/home), extracted on March 2021. Panel b: for US, the CPS series is from Morelli and Munoz (2020); CES is from Meyer, Mok, and Sullivan (2015), updated using U.S. Department of Labor (various years); for Philippines, the Family Income and Expenditure Survey (FIES), as available from the Philippine Statistics Authority's website (March 2021); for Jamaica, the Jamaica Survey of Living Conditions (JSLC), as reported in the annual reports.

To explain why the widespread and growing phenomenon of nonresponse in household surveys is undesirable and must be addressed, we take two perspectives: theory and practice.

Let us begin with theory. It is easy to show that unit nonresponse can induce *nonresponse bias* in most distributive statistics, including poverty and inequality estimates. To see this clearly, we set up some notation which allows us to express nonresponse bias in a simple and instructive form. Let N denote the number of households in the population (or the *universe*). Let N_{RU} be the number of households that, if sampled, would respond (R is for respondent, U is for universe) and N_{MU} the number of households that would not respond if interviewed (M is for missing). Of course, $N = N_{RU} + N_{MU}$. Using similar notation, let $\bar{X}_U$ be the welfare aggregate of the population, and $\bar{X}_{RU}$ and $\bar{X}_{MU}$ the welfare aggregate for respondents and nonrespondents, respectively. Based on these definitions, we can write the population welfare aggregate as a weighted average of the two groups, respondents and nonrespondents, where the weights are equal to the corresponding population shares, N_R/N and N_M/N:

$$\bar{X}_U = \frac{N_R}{N}\bar{X}_{RU} + \frac{N_M}{N}\bar{X}_{MU} \tag{7.1}$$

Equation (7.1) refers to the population from which the survey sample is drawn. Once fieldwork is completed, the data collected in the sample will only refer to respondents: this allows the estimation of $\bar{X}_{RU}$ in equation (7.1), but not of $\bar{X}_{MU}$. What are the consequences of failing to estimate $\bar{X}_{MU}$? Can the analysts still hope to get an accurate estimate for $\bar{X}_{U}$ from respondents alone? The answer can be obtained directly from equation (7.1). If we denote with $\bar{X}_{R}$ an unbiased estimator of the consumption aggregate for respondents, that is $E\left[\bar{X}_{R}\right] = \bar{X}_{RU}$, then the bias that affects $\bar{X}_{R}$ is given by the difference between the expected value of the estimator ($E[\bar{X}_{R}]$) and the true population parameter ($\bar{X}_{U}$):

$$\underbrace{E\left[\bar{X}_{R}\right] - \bar{X}_{U}}_{\text{nonresponse bias}} = \underbrace{\frac{N_{M}}{N}}_{\text{nonreponse rate}} \times \underbrace{\left(\bar{X}_{RU} - \bar{X}_{MU}\right)}_{\text{selective compliance}} \tag{7.2}$$

Equation (7.2) is the key to understanding the problem caused by unit nonresponse. Our hope is for the bias to be small, which happens if at least one of two things is true: either the ratio N_{M}/N is small, or the difference $(\bar{X}_{RU} - \bar{X}_{MU})$ is small. The first expression is the *nonresponse rate*. The second expression has to do with how similar respondents and nonrespondents are, in terms of the statistic of interest. If respondents are not systematically different from nonrespondents, then it is reasonable to expect that $\bar{X}_{RU} \approx \bar{X}_{MU}$; under this assumption, the analyst can ignore nonresponse, and use respondents as a representative sample of the population. Technically, we say that when survey compliance is random (or survey compliance is not selective), nonresponse is not a big concern. If, on the other hand, nonrespondents tend to differ from respondents ($\bar{X}_{RU} \neq \bar{X}_{MU}$), then the bias arising from using only respondents to estimate population parameters may make the entire survey worthless (Lohr 2009, 354). Note that the product on the right-hand side of equation (7.2) has no bounds: unlike N_{M}/N, which is necessarily less than 1, the difference in parentheses is unbounded, which translates into a clear and alarming message: *ignoring nonresponse can produce an arbitrarily large bias for the estimates of the parameters of interest*.[75] This is as far as elementary statistics can take us.

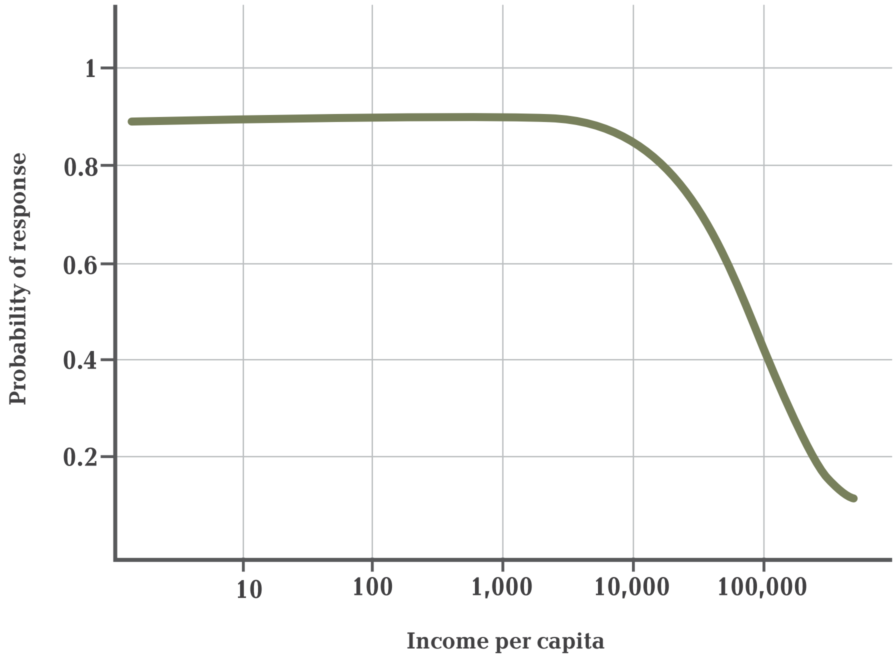
Let us now turn to nonresponse bias from an empirical standpoint. Though potentially catastrophic in theory, can one expect (or hope) nonresponse bias to be reasonably small, or at least acceptable, in practice? Figure 7.1 demonstrates that it is not uncommon for the nonresponse rate N_{M}/N to be large, typically above 20 percent. Equation (7.2), however, makes it clear that the impact of nonresponse critically depends on the interplay with the second component of the bias, $(\bar{X}_{RU} - \bar{X}_{MU})$, which in turn depends on *who* the nonrespondents are. A mountain of evidence suggests that nonresponse is in fact selective, and depends on a number of socio-demographic factors (Groves, Cialdini, and Couper 1992; Groves and Couper 1998; Groves 2006), including age, race, gender, educational level, health status, and income. In general, the higher the opportunity cost of the time required to comply, the lower the response rate, and nonrespondents will typically be better off than respondents.

Figure 7.2 illustrates that the probability of responding (vertical axis) has been found to decrease monotonically with income (horizontal axis), so that the richer a household is, the less likely it is to be in the sample (Korineck, Mistiaen, and Ravillion 2006). The expected practical consequences of selective compliance are (likely downward) biased estimates for the level of the welfare aggregate, an underestimation of inequality, and an overestimation

[75] What holds true to for the estimation of means, the case considered in the text, requires additional qualifications when estimating proportions or other nonlinear statistics, but the overall message applies.

of poverty rates.[76] It is worth noting that in low-income countries, survey compliance is likely to be selective at the bottom of the income distribution, too. The reason, in that case, tends to be noncontact rather than refusal (those struggling to make ends meet may be away from home for most of the day, especially in urban settings).[77] Ultimately, evidence suggests that analysts *cannot* ignore nonresponse, and *cannot* use respondents as a sample that is representative of the whole population.

FIGURE 7.2. Selective compliance: Better-off households are less likely to participate in surveys



SOURCE: The curve is a stylized version of an empirical finding originally proposed by Korinek, Mistiaen, and Ravallion (2006, 2007). See also Morelli and Muñoz (2020, 2021).

The single best strategy to minimize the negative effects of nonresponse is to prevent it from happening, that is, to address the problem ex ante. The research on nonresponse has highlighted several avenues to do that in practice (Brick 2013; Plewes and Tourangeau 2013; Tourangeau et al. 2010). One is to study the psychological and sociological mechanisms that cause nonresponse (Goyder 2019; Groves and Couper 1998; Toureangeau, Rips, and Rasinski 2000), with the aim of reducing respondent burden. Another strand of research focuses on data collection modes: the choice of mode (face-to-face, telephone, mail, internet, etc.) can influence cooperation significantly, as can the methods used to follow up with nonrespondents (see, for instance, Bethlehem, Cobben, and Schouten 2011, ch. 4). Other salient topics in this vast literature include interviewer effects (Schaeffer, Dykema, and Maynard 2010), and the role of incentives (Singer and Ye 2013).

---

[76] On the strong and negative income effect on survey compliance, see Ravallion (2021) and Hlasny and Verme (2018a, 2018b).

[77] Though, to our knowledge, empirical evidence on this phenomenon is still lacking, this is a widespread perception among survey experts and practitioners, and worth a passing remark.

The ex ante approach, though preferred, exists alongside a number of ex post strategies, namely statistical adjustments aimed at minimizing nonresponse bias. Section 7.1.1 does not go into the vast literature on nonresponse adjustments, beyond the citation of some key references. However, it does offer a compelling reason to engage with these, by discussing the threat that nonresponse poses to the final estimates through its impact on survey weights (expansion factors).

7.1.1 Survey weights

Setting weights correctly is integral to the computation of unbiased estimates for the distribution of household consumption, as well as for any statistics of interest derived from that distribution. In fact, one could argue that the appropriate weighting of survey data is no less important to poverty and inequality measures than other issues discussed in this document (Ravallion 2021). Also, many analysts can relate to the feeling of being in treacherous waters when using survey weights: there are often doubts on which weight variable should be used (households versus individuals), or which weight type (Stata users have four different choices—`aweight`, `fweight`, `iweight` and `pweight`—and therefore a potential quadrilemma).[78]

In the presence of a complex survey design, each sampled unit (say, each household) receives a sampling weight, that is a value that accounts for the fact that different units are selected with different probabilities. A more compact way to say this is that each of the N households in the population is assigned a sampling probability π_i ($i = 1, ..., N$). This is not the probability that the i-th household is in the sample, but, if we imagine sampling to happen as a series of consecutive draws from the population, the probability that the household is selected at each draw (Deaton 1997, 44). If the sample has size n, then the probability of inclusion in the sample is $p_i = n\pi_i$.[79] The sampling weight is defined as the reciprocal of that probability of inclusion:

$$w_i = \frac{1}{p_i} \tag{7.3}$$

The weights w_i in equation (7.3) have a straightforward interpretation: they give the number of population households represented by the sample household i. For example, if a household is included in the sample with probability 1/50, that household represents 1 out of 50 households in the population from which the sample was drawn. In fact, it is worth noting two additional facts about weights. First, consider the sum of the weights in equation (7.3) across all the n sampled households:

$$\widehat{N} = \sum_{i=1}^{n} w_i \tag{7.4}$$

[78] See Valliant, Dever and Kreuter (2013), and Valliant and Dever (2018).

[79] This formulation holds for the case of simple random sampling, which is more of a theoretical than a realistic occurrence in current survey practice. Virtually all large-scale surveys use complex designs that involve multiple sampling stages, so that the formulation of weights in eq. (7.3) should be seen as either a simplification, in the interest of clarity of the text, or as a condition that holds within stratum. This does not in any way change the substance of the discussion on weighting: we left the topic of multistage sampling aside for the sake of simplicity.

Equation (7.4) says that the sum of sampling weights provides an estimate ($\widehat{N}$) of the size of the population (N). Second, suppose that x_i is the variable of interest reported by the i-th household, say its income. The *total* income in the population can be estimated as a weighted sum of household incomes:

$$\hat{x} = \sum_{i=1}^{n} w_i x_i \qquad (7.5)$$

This is known as and commonly referred to as the *Horvitz-Thompson (HT) estimator* of the population total. Similarly, poverty rates are estimated using HT estimators, as are inequality measures and any other statistic of interest. In general, HT estimators are *the* estimators used, as a rule, in the presence of survey data. The merit of equation (7.5) is that it gives visibility to the role played by the weights: good estimates require both good data (x_i) *and* good weights (w_i).

The weights in equations (7.3) to (7.5) are referred to as *base weights* (or *design weights* or *selection weights*). They are typically constructed by the specialist responsible for sampling design before data collection commences, so that, by their own nature, base weights do not account for unit nonresponse. In practice, in the presence of (selective) nonresponse, equation (7.5) is no longer true, and the unbiasedness of weighted estimates is at stake.

Several strategies are routinely employed by survey specialists to deal with this situation. Adjusting base weights is a common approach, one that encompasses a wide array of techniques. The essence of all weighting adjustment procedures is to increase the weights of respondents so that they represent the nonrespondents (Kalton and Kasprzyk 1986, 2; Kalton and Flores-Cervantes 2003); ultimately, the way survey specialists develop these procedures is by using some *response propensity* model, that is, a model that estimates the probability that each unit will respond. The multiplication of the original sample selection weight (w_i in equation 7.5) for each sample unit by the reciprocal of its modelled response propensity creates a new weight, which, *if* the model is correct, enables an unbiased or nearly unbiased estimation of population statistics from the survey data (Heeringa, West, and Berglund 2017, ch. 2).[80]

7.2 Item nonresponse

Item nonresponse refers to missing *values* of particular items in the questionnaire (when a respondent has completed the questionnaire, but some of her answers are missing), as opposed to unit nonresponse, which indicates missing *records* (when a subset of sampled households does not complete the questionnaire, as described in the previous section). In this section, we refrain from recommending a specific protocol for dealing with missing values, given that the optimal course of action is highly dependent on the context, namely the variable under consideration, and a host of country- and survey-specific circumstances. However, understanding the nonresponse mechanism underlying the observed patterns of missingness is a precondition for drawing up any strategy to deal with the problem, and this is the focus of the section.

[80] Other popular methods are discussed in Lohr (2009, ch. 8), Bethlehem, Cobben, and Schouten (2011, ch. 8), and Kolenikov (2016).

We start from a stylized representation of the consumption aggregate for a given household, which is obtained as the sum of individual expenditure components:

$$x = \sum_{j=1}^{J} x_j = x_1 + x_2 + \ldots + x_J \tag{7.9}$$

where x_j denotes a household's consumption expenditure for item *j*, with *j* running from the first to the last expenditure-related question that is put to respondents. Equation (7.9) can similarly represent the way an income-based welfare indicator is constructed (appendix C), where x_j would denote elementary income sources (*e.g.* wages, income from self-employment, pensions, and so forth).

Chances are that one or more of the items on the right-hand side of equation (7.9) are missing: in other words, the value of one or more of the x_j is unknown. Missing values arise when, for instance, a section of the questionnaire, such as the food diary, is lost, or deemed invalid. As a result, food expenditures are missing, while nonfood expenditures collected in other sections of the questionnaire are known. Loss of information need not be this extensive to be concerning: in the case of households failing to report the amount of rent paid, for instance, nonresponse to a single question is responsible for a fundamental component of the consumption aggregate missing.

From a theoretical standpoint, there is good reason for a welfare analyst to be concerned with missing values. First, the presence of missing data implies a reduction of the information available, and consequently, a loss of *precision* (efficiency) of estimates: smaller sample sizes imply larger standard errors. Second, and perhaps more important, missing values may cause *bias* to estimates. Whether or not estimates of, say, inequality or poverty are affected by the presence of missing values crucially depends on the reason *why* data are missing, which we call the *nonresponse mechanism* (Rubin 1976).

The best-case scenario is when data are missing by pure accident: a respondent forgets to answer a question, a random part of the data is lost during processing, or other such circumstances. When this is the case, we say that data are *missing completely at random* (*MCAR*). What we mean, more precisely, is that the probability that a value is missing does not depend on the value of the variable, or on any other characteristic of the respondent. Under MCAR, the available sample, though incomplete, can be regarded as a random subset of the ideal data (what we would observe if there were no missing values). The implication is that there is a loss of information, and therefore precision, but no risk of bias in the statistics of interest. If the analyst can support the MCAR hypothesis, for instance by showing that there are no systematic patterns of missing data, then there is no need to bother any further with missing values, and the analysis may be carried out on the complete cases.

Unfortunately, MCAR rarely occurs in practice. Typically, missing values depend on the values of auxiliary variables—for instance, the burden of filling out a long food diary may be felt more strongly by less literate households in rural areas than by urban households. If that were the case, area of residence (urban or rural) would be the auxiliary variable correlating with missingness, and expenditures would not be missing completely at random. If missing values *within* each group, that is, among rural households only, can be assumed to arise by pure chance, then data are said to be *missing at random* (*MAR*). The implication is that we can assume MCAR *within* appropriately defined groups. A number of simple solutions are available to deal with this situation, such as using *hot-deck* imputation or any of its variants to impute missing values (Andridge and Little 2010; Little and Rubin 2019).

Finally, the worst-case scenario occurs when data are *missing not at random* (*MNAR*): for instance, persons who are better off (that is, with a high total income and/or expenditure) may be more likely to refuse to report their expenditures (because of reticence or lack of interest in responding). When data are MNAR, the probability that a value is missing depends on the value of the variable itself (and possibly also on the values of auxiliary variables). This situation is the most difficult to deal with analytically. In general, one has to make assumptions in order to model the dependence of the nonresponse mechanism on the values of the target variable (Nicoletti 2010; de Waal, Pannekoek, and Scholtus 2011, ch. 1).

From this general discussion, we can draw several conclusions that are relevant to the work of the welfare analyst. First, in most practical applications, restricting the analysis to complete case records (that is dropping observations with missing values from the analysis) will produce biased (and inefficient) estimates, as missing values rarely occur completely at random (they are *not* MCAR).

Second, while taking some action to deal with nonresponse bias is considered appropriate in many practical applications, there is no such thing as a universal method that applies to all variables and all contexts. What is clear is that, before embarking on imputing missing values, the most common approach when item nonresponse is concerned, it is essential to investigate (1) the *cause* of missingness, and (2) the *pattern* of missingness in the sample. Sometimes the cause can be easily determined: for instance, when the Computer Assisted Personal Interviewing (CAPI) system has an incorrect skip, or the data processing routines incorrectly replace zeros with missing values.[81] In these circumstances, mistakes can and should be fixed by going back to the raw data. When the cause is not clearly identified, it is advisable to investigate the pattern of missingness. Even simple two-way tables where the distribution of missing values is examined by region, urban-rural areas, per capita consumption deciles, or other dimensions are often insightful enough, despite their simplicity, to explore the MCAR (or MAR) hypothesis.

In the most common situation, when data are MAR, the analyst may take several routes, depending on how much is known about the data generating process, how relevant the likely impact of the missing expenditure component over the aggregate is, and so on. No "best imputation method" exists. A review of the current practice suggests that regression-based imputation methods may be preferred in cases where a model can be built by tapping into an existing theoretical and empirical apparatus, such as for housing expenditure: hedonic models are discussed in section 4.5, or food expenditure: econometric models for consumption analysis abound from Stone's (1954) Linear Expenditure System, to Deaton and Muellbauer's (1980) Almost Ideal Demand System, and Banks, Blundell, and Lewbel's (1997) Quadratic Almost Ideal Demand System. Simple *hot-deck* techniques are sometimes used when little is known about the process generating the data, and when an investment in complex analytical machinery is excessive compared to the likely size of the missing expenditure component.

If there is evidence that data are MNAR, then the problem is more serious and requires developing ad hoc imputation models, a topic that, by its own nature, does not lend itself to general recommendations.

---

[81] Because of the way some questionnaires are constructed, respondents may skip questions that do not apply to their situation (for instance, no household members are currently in school, therefore the interviewer does not ask any of the questions on education expenditures). These responses, which are known and equal to zero, are sometimes coded as empty records, but they are certainly not missing values.

A final remark on the consequences of imputing missing values is in order. Dempster and Rubin's (1983, 3—10) admonition is worth revisiting:

> *"The idea of imputation is both seductive and dangerous. It is seductive because it can lull the user into the pleasurable state of believing that the data are complete after all, and it is dangerous because it lumps together situations where the problem is sufficiently minor that it can legitimately be handled in this way and situations where standard estimators applied to real and imputed data have substantial bias."*

Whatever the imputation technique, *imputed* values cannot be treated as *genuine* values. A dataset containing a large share of imputed values will imply smaller estimated standard errors than those that would have been obtained in the absence of missing data. This implies that welfare comparisons, hypothesis tests, and ultimately any sort of statistical inference will be invalid.[82] The presence of missing values, in other words, ultimately reveals that something has not worked properly with the survey, and while it is desirable to "fix" as much as is required, the general recommendation is threefold: (1) bear in mind that "data" and "imputed data," despite appearances, are not synonyms; (2) document the presence of missing values and any consequent imputation procedures (a recent report on living standards in Afghanistan 2016/17 provides a useful example of how to do so (Central Statistics Organization 2018, 344); and (3) carry out sensitivity analysis to assess the impact of changes in the distribution of the consumption (or income) aggregate when imputations for missing values have been implemented (Coudouel, Hentschel, and Wodon 2002, 46).

7.3 Outliers

Outliers—values "too small' or "too large" compared to the bulk of the data—are everywhere: they crop up in any sample, any dataset, any empirical application. The definition of outlier provided half a century ago by Grubbs (1969, 1) underlies most of the many definitions that are in use today:

> *"An outlying observation, or 'outlier,' is one that appears to deviate markedly from other members of the sample in which it occurs. An outlying observation may be merely an extreme manifestation of the random variability inherent in the data. If this is true, the values should be retained and processed in the same manner as the other observations in the sample. On the other hand, an outlying observation may be (...) an error in calculating or recording the numerical value. In such cases, it may be desirable to institute an investigation to ascertain the reason for the aberrant value. The observation may even eventually be rejected as a result of the investigation, though not necessarily so."*

The definition makes two main points. First, it highlights the inherently *relative* nature of an outlier: an outlier is an observation that appears to be abnormal, where the norm is set

[82] A number of methods are available to mitigate the problem, most notably multiple imputation (MI), where each missing value is replaced by two or more imputed values in order to represent the uncertainty about which value to impute (Rubin 1987, vi). A proper discussion of the method is beyond the scope of this section—Ardington et al. (2006) illustrates the method for the case of South Africa; see also the 2018 South Sudan poverty report (Pape and Parisotto 2019), Myanmar, Lebanon and a few others.

by the remainder of the data.[83] Second, it debunks a common misconception: that outliers are a problem that *must* be fixed. Equating "outlier" with "error" plays into the practice of taking excessive liberties with the data, indiscriminately dropping observations from the sample, or even over-editing. One should always keep in mind that a few observations can be genuinely abnormal, and can represent, at least in principle, novel and important information—scientific discovery depends, at least in part, on outliers. For example, Rutherford discovered the atomic nucleus because of the outlying behavior of a few particles.

A third point, that instead does not appear in the definition, is that an outlier, be it an error or not, may not be *influential:* whether outliers matter at all depends on the context, and more precisely on the statistic of interest. Inequality estimates, for instance, tend to be extremely sensitive to the presence of extreme values (Cowell and Victoria-Feser 1996a). On the other hand, poverty estimates are generally insensitive to what happens above the poverty line, regardless of how extreme the top values are (Cowell and Victoria-Feser 1996b).[84] Practical comparisons of income distributions (for example, Lorenz comparisons, or rankings based on stochastic dominance) are also highly sensitive to outliers (Cowell and Victoria Feser 2002, 2006, and 2007).

Overall, these considerations lead to a first important conclusion: in no way can we devise a single "best" strategy to deal with outliers in all situations. Different courses of action, including taking no action at all, are likely to be appropriate in different contexts. However, it is hard to deny that, in the specific settings frequented by welfare analysts, for the specific distributions that are routinely analyzed (consumption expenditure, calorie intakes, unit values, and so on), extreme values are typically seen as *potentially* inaccurate, and the need for examining the data and detecting outliers is not questioned; rather, the debate is focused on methodology.

At a very fundamental level, the analyst is faced with the choice between two alternative approaches to tackle the issue of outliers. The first approach is to use *robust* estimation procedures, that is, estimators that are not influenced by the presence of outliers in the sample. The second approach is to retain the use of *standard* estimation and testing procedures, detect outliers in the distributions of interest, and apply whatever adjustment is deemed appropriate, if any, to the data itself (Huber 1981, 4; 1996). Almost invariably, in the context of welfare analysis, the preferred strategy is to embrace the latter approach, given the need to produce standard poverty and inequality indicators that are comparable across countries and over time, as well as "clean" micro-data for general and public use (Filzmoser, Gussenbauer, and Templ 2016).

In this scenario, do welfare analysts share a common conceptual framework when detecting and treating outliers? The answer is negative (Agunis, Gottfredson and Joo, 2013). Our

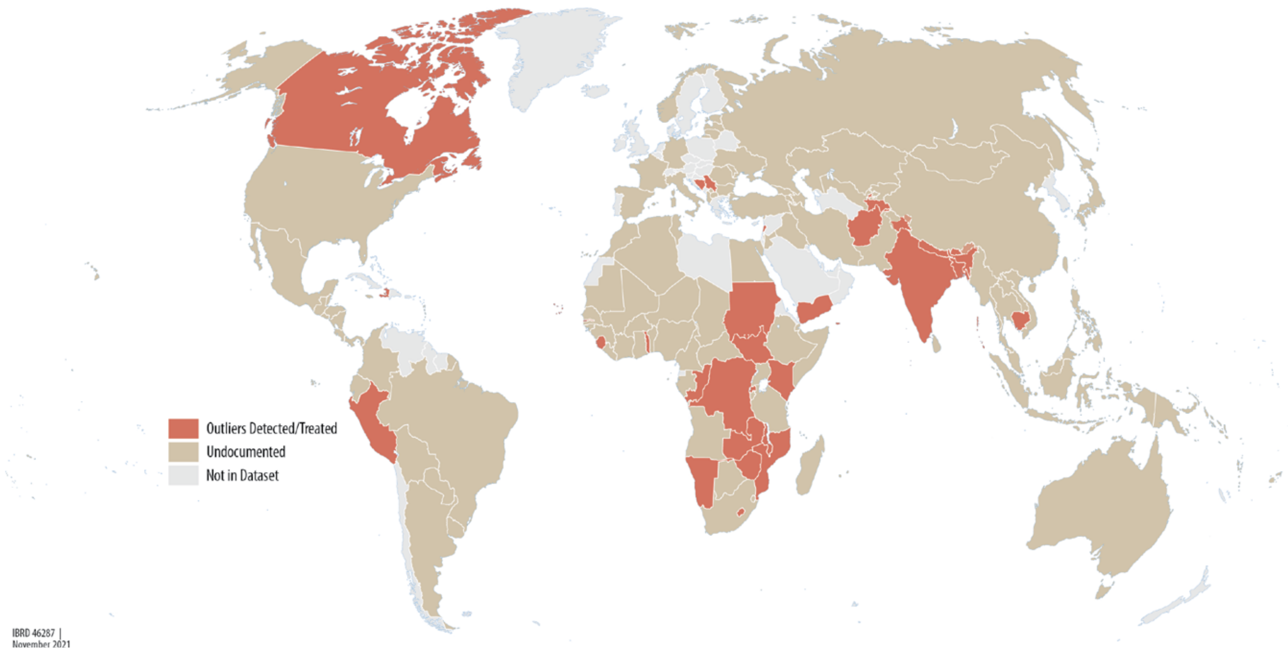
---

83 See Barnett and Lewis (1994) for an extensive yet accessible discussion of the relative nature of an outlier.

84 This holds true if the poverty line is set exogenously, which is often the case in practice, even if not necessarily so in theory. An important exception is the case of *relative* poverty lines. For example, Eurostat takes as a poverty line for any member country 60 percent of the median income. Here the poverty line is explicitly dependent on the data; "this (...) means that the poverty line itself is being estimated from the data, and it thus implies that there is an additional channel by means of which data contamination may bias the estimates of poverty." (Cowell and Victoria-Feser 1996b, p. 1768).

review of the current practice shows that the most popular strategy is to let the issue go *undocumented*. Map 7.1 shows that, for an overwhelming majority of countries, the documentation accompanying the release of official poverty and inequality estimates does not even mention whether outliers were dealt with, and how. Where outliers are treated and documentation is available, methods vary: countries may detect outliers by graphical inspection of the distribution of interest (e.g., Zambia 2016), by flagging the top and bottom 1st to 5th percentiles (e.g., Nepal 2011, 41; Mozambique 2018, 71), by setting detection thresholds based on the mean and standard deviation of the normalized target variable (e.g., Namibia n.d., 28; Maldives 2018, 11), or by applying other "miscellaneous" rules (ECASTD 2016). Potentially, this is a serious threat to the comparability of results, at least for what concerns inequality, both over time (the effect of domestic redistribution policies will be unclear) and across space (international and intra-country geographic comparisons will be at risk). A number of theoretical results justify this concern. Cowell and Flachaire (2007, 2015) demonstrate the high sensitivity of the most popular inequality indices to the presence of extreme values in *both tails* of the distributions. In the presence of outliers, the expenditure distribution becomes skewed and heavy-tailed, a feature that causes problems not only to the point estimates of most inequality estimators (including Lorenz curves) but also to their standard errors, and consequently to any statistical inference exercise, for example, tests on the difference between Gini indices from two regions or two surveys in different years—see Davidson (2012), Schluter and van Garderen (2009), and Schluter (2012).

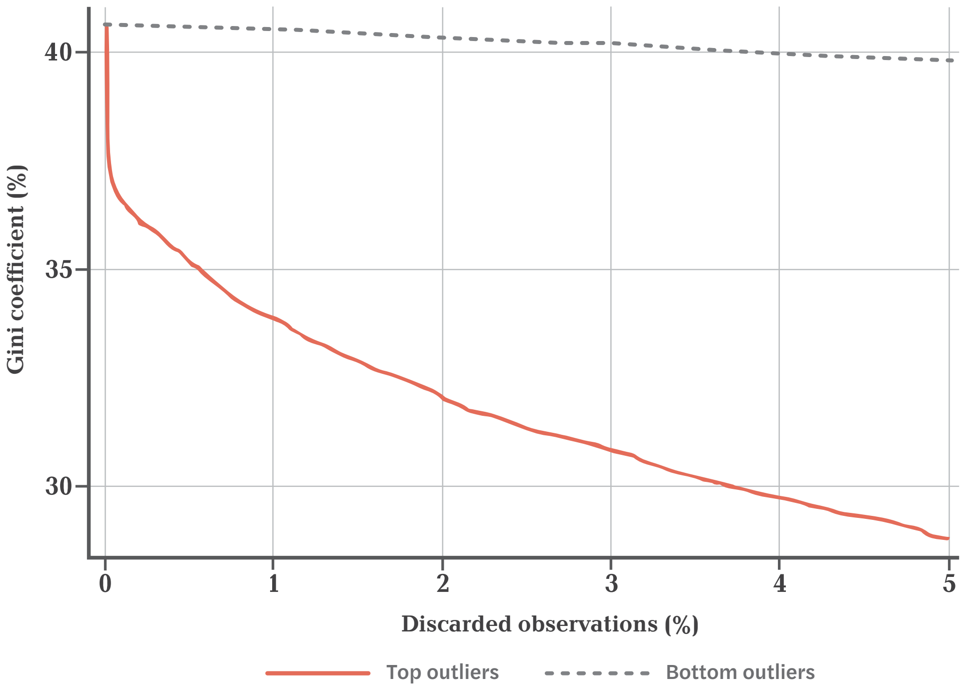
MAP 7.1. Countries that detect and treat outliers of the welfare aggregate



SOURCE: Authors' elaboration of the dataset described in appendix A.

While the findings emerging from this literature are unequivocal, a practical experiment can help convey the message. Figure 7.3 shows the impact of extreme values on the Gini index.[85] First, the Gini index is computed using the raw data (that is, on the per capita consumption variable "as is" in the dataset released by the National Statistical Office [NSO]). This value, 40.6 percent, can be read on the vertical axis in correspondence of zero in the horizontal axis. The solid line shows the value of the Gini index as the *largest* observations are excluded from calculations, one at a time (each time the weights of the remaining observations are recalibrated so as to sum up to the entire population); the dashed line shows Gini when the *smallest* observations are removed, one at a time. The figure shows that, by excluding the five largest observations, corresponding to just 0.04 percent of the sample size, the Gini index decreases to 36.9 percent. Despite the relatively large size of the sample (almost 12,500 observations) each of the five discarded observations is worth almost 1 Gini point. The Gini index is not as sensitive to small outliers, a finding explained by Cowell and Flachaire (2007), and Ceriani and Verme (2019), but other indices are.[86]

FIGURE 7.3. Sensitivity of Gini estimates to the presence of extreme values



SOURCE: Data are from Malawi's 2016 Fourth Integrated Household Survey (IHS4), as available from RuLIS (2020)—data extracted in December 2020. The figure is purely illustrative and does not reflect official statistics.

[85] This is not a special case (van Kerm 2007); see the Incremental Trimming Curve (ITC) in Belotti, Mancini, and Vecchi (2021), which borrows from Hampel (1974), Hampel, Ronchetti, Rousseeuw and Stahel (1986, chapter 2), and Cowell and Flachaire (2007, section 6).

[86] See also Hlasny and Verme (2018a) for a similar application in Egypt. For measures other than the Gini index, such as the Atkinson index or the Generalized Entropy Indices introduced by Shorrocks (1980), with parameter less than zero, sensitivity is greater for small outliers.

The overall conclusion from both the theoretical literature and empirical applications is that the detection and treatment of outliers cannot be an afterthought. The application of a consistent methodology to detect extreme values, paired with careful documentation of their treatment, would be a step forward in the direction of comparability and transparency of final estimates.

7.3.1 Detection and diagnostics

While it is not possible to point to a universally accepted strategy for dealing with outliers, it is useful for the analyst to be aware of the main options at her disposal. In general, all approaches can be seen as consisting of two steps: detection and treatment. Outlier detection entails deciding what makes a value "extreme" in the context at hand: going back to Grubbs' definition in the opening paragraph of section 7.3, what does it mean, exactly, for an observation to be far away from the bulk of the distribution? Outlier treatment is deciding what to do about it: replacing or otherwise rejecting the extreme value, versus leaving it as is. In practice, the two decisions are often intertwined, but thinking about them as separate steps makes for a more transparent discussion of the underlying reasoning.

It bears repeating here that checking for "gross mistakes" and other context-specific sources of measurement error is a given when handling extreme values: these checks are usually performed by NSOs as part of a set of routine edits. At some point, though, all planned checks will be completed, and the data will be deemed final. This is exactly the stage this section focuses on, when upon receiving the NSO-cleared datasets, welfare analysts consider taking further steps to shield the statistics of interest from any residual "noise."

Regarding outlier detection, analysts regularly resort to both "subjective" approaches and "objective" rules. The former are often based on manual or visual inspection of the data: checking the largest and smallest values of a given variable, graphing its distribution, and so on, and determining whether or not anything "looks off." Naturally, this can be difficult to decide, and even more difficult to document. In many cases, analysts find it useful to apply "objective" outlier detection rules, that is, pre-determined statistical criteria to flag extreme values.[87] Typically, such rules rely on some definition of distance from the bulk of the distribution, and on the identification of a threshold beyond which this distance is considered "too large," so that observations falling past it get flagged. This basic logic gives rise to the abundance of criteria found in the literature, which range from simplest (such as the widely known boxplot rule) to most technical (such as the multivariate methods summarized in Filzmoser, Gussenbauer, and Templ 2016).

A version of the latter approach is worthy of illustration, as it provides analysts with a practical diagnostic tool, particularly useful in the context of sensitivity analysis (section 8). Let X denote the variable of interest, the target for outlier detection, and let $f(x)$ denote its

[87] The word "objective" in this particular context is merely used to indicate that the decision to flag an observation as an outlier is based on some algorithm or ex ante rule, rather than on a case-by-case evaluation of the plausibility of observations. However, "objective" rules still require to make a judgment call on what it takes for an observation to be considered "extreme"—the judgment is just made explicit.

probability density function (pdf). Detecting outliers in the distribution of X requires one to determine how high or low a value should be to qualify as extreme. If the distribution of interest is known—for example, if $f(x)$ is Normal—then one can consider an observation to be an outlier if it falls into a range of values that occurs with arbitrarily low probability (say 5 percent or 1 percent). An observation x falling into an *outlier region* defined in this way could conceivably be produced by the theoretical distribution of income, but that would be a rare occurrence, making x an extreme value, or an outlier (Davies and Gather 1993; Gather and Becker 1997). A conventional application of this criterion, sometimes called the "three sigma rule," identifies the bounds of the outlier region for a Normal distribution as the mean ($\bar{x}$) plus or minus three times the standard deviation(σ_x) (each tail region defined in this way has a probability of about 1 percent: in formulas, all values of x falling outside the range $[\bar{x} - 3 \times \sigma_x, \bar{x} + 3 \times \sigma_x]$ would be flagged as outliers). Equivalently, the outlier detection rule can be formulated by means of the z-score:

$$\left| \frac{x - \bar{x}}{\sigma_x} \right| > 3 \qquad (7.10)$$

The rule in equation (7.10) would flag as outliers all values of x whose z-score exceeds 3 in absolute value.

The application of this simple criterion runs into two problems in practice. First, the empirical pdf of income, and of most other variables of interest to welfare analysts, is certainly *not* Normal. Rather, it is typically *unimodal*, *asymmetric*, and *heavy-tailed* compared to a normal distribution. If the analyst can transform the raw distribution into something that is approximately Normal, the algorithm can still be applied: observations that are flagged in the transformed distribution are also outliers of the untransformed distribution. One must find a normalizing transformation that works well enough for the distribution at hand. Fortunately, candidates abound. During the early 2000s, in a contribution to the "great Indian poverty debate," Deaton and Tarozzi (2005) used the natural logarithm as a normalizing transformation for unit values of commodities consumed by households. Dupriez (2007) explored the use of the Box-Cox transformation (Box and Cox 1964), which includes the log transformation as a special case. Other useful transformations include Yeo and Johnson (2000), Friedline, Masa, and Chowa (2014), and many others.[88]

A second, subtler problem is related to the definition of the outlier detection region in terms of mean and standard deviation of the distribution: the empirical mean and standard deviation are vulnerable precisely to the outliers one is concerned about. The presence of a few extremely large observations in the income distribution, for instance, is likely to "pull" the sample mean in equation (7.10) of the transformed variable with them, increase it, and inflate its standard deviation. As a consequence, the outlier detection region would be smaller than when such extreme observations are absent, and the rule for outlier detection defined in equation (7.10) would be more forgiving to all other observations elsewhere in

[88] Goodness-of-fit criteria such as the Pearson chi-squared test (Snedecor and Cochran 1989) can be used to determine which transformation is best in any given context. The Pearson statistics divided by its degrees of freedom converges to 1 when the data approaches a Gaussian distribution: it can be interpreted as a measure of how close a distribution is to normality and used to rank transformations according to how successful they are in normalizing the data.

the distribution, regardless of whether they, too, are extreme compared to the bulk of the sample. Once again, the remedy is simple enough, as robust measures of location and scale are not difficult to come by: the median can replace the mean, and as for the standard deviation, alternatives include the interquartile range, the mean absolute deviation (Hampel 1974), and many others.

In light of these considerations, the outlier detection strategy resulting from the discussion above can be summarized with the following two steps: (1) transform the variable of interest to induce *normality* in its empirical distribution, and (2) use *robust* statistics to set the thresholds of the outlier region. The z-score in equation (7.10) is therefore replaced by its robust counterpart, and the resulting (two-tailed) outlier detection rule is as follows:

$$\left| \frac{t - med(t)}{Q_t} \right| > 3 \tag{7.11}$$

where t denotes the transformed (normalized) variable, $med(t)$ is its median, and Q_t is a robust estimator for the dispersion, or scale, of t (Rousseeuw and Croux 1993). According to equation (7.11), we would flag as outliers all values of x whose transformation t falls outside the region $[med(t) - 3 \times Q_t, med(t) + 3 \times Q_t]$. This approach is explained in more detail in Belotti, Mancini, and Vecchi (2021), and paired with an extensive empirical application.

A number of practical questions remain: Should one detect outliers at the national, regional, or subregional level? Are we to worry only about outliers of the consumption aggregate itself, or should we care about the distribution of expenditure components, too? Further still, once outliers are flagged, what should we do with them? If suggesting a specific outlier detection criterion is controversial, then recommending a specific treatment procedure is utopian. Too many factors that cannot be generalized come into play: the variable of interest, the number of outliers detected, the relationship with other variables in the dataset. It is not coincidental that, while the consensus on the relevance and impact of data cleaning is nearly universal, best practices to deal with data issues continue to be elusive. This should not discourage analysts from investing as much effort as necessary at this stage of the analysis. The usual recommendations apply: irrespective of where the choice may fall in terms of outlier detection and treatment, it is crucial to document what was done, and to present results based on "raw" as well as "clean" variables (sensitivity analysis). Methods for sensitivity analysis are discussed more generally in section 8.

8. Sensitivity Analysis

The process of constructing a consumption aggregate and estimating inequality and poverty is riddled with methodological dilemmas. Even when sticking by the recommendations from the theory and literature, rarely does the analyst have one clear path toward the end result. More often, she will find herself having to choose between several equally valid options: for instance, there is typically more than one viable imputation strategy for missing or extreme values (section 7.3). Even more often, she will be faced with the need to pick one of a few "bad" alternatives: is it better to include an unreliable self-reported rental value in the aggregate, or one that is modelled, even though we may know the model may not be very good, either? (section 4.5) Should we construct a spatial price index based on "noisy" survey-based unit values, or should we skip spatial deflation altogether? (section 5.2). Finally, sometimes there simply is no consensus in the literature on the best course of action: the choice of the equivalence scale that is most appropriate for adjusting the consumption aggregate is a notable example (section 6).

Ultimately, at each of these turns, a choice must be made. Arbitrariness is unavoidable, but it can and should be managed. Good practice requires that the analyst provide an evaluation, possibly a quantification, of the impact of arbitrary choices on final estimates. This can be accomplished through *sensitivity analysis*, which may be defined loosely as the study of how changes in the inputs of a process affect its output.[89] In our case, the process is the computation of a welfare indicator, the inputs are the methodological choices that shape it, and the output can be one or more statistics of interest, such as inequality and poverty estimates. In general, the goal of sensitivity analysis is testing whether results are robust to the assumptions made by the analyst. Which of the welfare analyst's decisions should be investigated? And how?

The first question poses the problem of "picking one's battles." One cannot investigate every single controversial choice made when constructing the aggregate: "There are so many points where judgment calls have to be made, and they combine with one another to produce an impossibly large number of alternatives. Decisions have to be made for better or worse" (DZ, p. 63). The *Guidelines* go on to argue that not every controversial decision is also influential for the statistics of interest: the sensible approach is to focus efforts on the choices that have the potential to be both. DZ shortlist two of them: the choice of an equivalence scale and the inclusion of an expenditure component that is "atypical" or measured with error, which they indicate as top candidates for sensitivity analysis. In fact, depending on the context and the statistics of interest, the list could be easily expanded. This section will not attempt to propose an inventory of all the situations where sensitivity analysis should be performed; this would be an impossible feat. Instead, it focuses on how to investigate the

[89] The terms "sensitivity" and "robustness" may assume specific technical meanings in different contexts. In this section, "sensitivity analysis" and "robustness checks" are used interchangeably, and the word "robust" should be taken to mean "not sensitive."

methodological choices that the analyst will deem critical. We discuss the main tools for the job—tables (section 8.1) and curves (section 8.2)—and do so by means of practical examples. Section 8.3 focuses on testing the sensitivity to the choice of a poverty line and a poverty index. Section 8.4 summarizes the main recommendations.

8.1 Tables: Side-by-side comparisons

The most straightforward test of the impact of measurement assumptions on results consists in a side-by-side comparison of the statistics of interest (e.g., inequality and poverty indicators) computed under different scenarios: alternative definitions of the welfare aggregate, alternative methods for dealing with outliers, alternative spatial deflators, and so forth.

Table 8.1 showcases a few examples of sensitivity analysis done this way, taken from recent poverty assessment reports. To illustrate, let us take the case of row 1, which reproduces the results of a test for the sensitivity of poverty rates to the imputation of missing data, for the case of Bangladesh in 2016. In that year's Household Income and Expenditure Survey, an abnormally high percentage of households were found to report zero expenditures for education, despite having household members currently enrolled in school. Analysts decided to impute missing and zero values of education expenditures so as to reconcile the information provided in different modules of the questionnaire. Because the decision of imputing missing values is a controversial one (see section 7), the impact of this choice on final results was duly investigated. This was done by comparing the headcount poverty rate under the "no imputation" scenario (25.1 percent, corresponding to Method A in table 8.1) with the headcount poverty ratio under the "imputed data" scenario (24.8 percent, corresponding to Method B in table 8.1). The comparison leads to the conclusion that the impact of the imputation on the headcount poverty rate is not significantly different from zero. The remaining rows of table 8.1 reproduce the results of similar checks.

TABLE 8.1. Sensitivity analysis in practice: Side-by-side comparisons of poverty headcount rates (percent)

Source	Decision	Method A	Method B	Diff A–B
1	Data imputation: no imputation (A) vs. imputing zeros and missing values in education expenditures (B)	25.1	24.8	0.3
1	Outlier detection: extreme values identified by strata (A) vs. by divisions (B)	24.3	24.0	0.3
2	Food expenditures: annualization using actual diary-keeping days (A) vs. assumed 30-day period (B)	23.2	22.5	0.7
2	Data source for spatial deflation: price survey for nonfood and fish, unit values for food (A) vs. unit values for all items (B)	23.2	27.8	−4.6***
3	Health expenditures: included in the consumption aggregate (A) vs. excluded (B)	17.0	27.0	−10.0***
3	Durable goods: included in the consumption aggregate (B) vs. excluded (A)	26.6	27.4	−0.8

(Continued)

TABLE 8.1. (Continued)

Source	Decision	Method A	Method B	Diff A–B
4	Durable goods: acquisition approach (A) vs. consumption flow with geometric depreciation (B)	31.0	30	1.0
5	Spatial deflation: use Laspeyres price index (A) vs. nominal unadjusted consumption aggregate (B)	27.0	25.0	2.0**
5	Inflation: adjustment for within-year inflation (A) vs. use of nominal consumption aggregate (B)	27.1	27.6	−0.5
5	Equivalence scale: use per capita expenditure (A) vs. per adult equivalent (B)	32.0	21.4	10.6***

NOTE: Standard errors are omitted to avoid cluttering the table. The final column, reporting differences between point estimates, is a back-of-the-envelope calculation by the authors, based on published information. Stars indicate the significance of differences: * p<0.05, ** p<0.01, *** p<0.001.
SOURCES: [1] Bangladesh (2017, 19); [2] Lao PDR (2014, 43); [3] Lebanon (2015, 27); [4] Myanmar 2017, 44; and [5] Palestine (2018, 29).

The approach illustrated by the collection of examples in table 8.1 is easy enough to implement in most situations and can be very insightful, despite its simplicity. For these reasons, it is a useful first step for any sensitivity analysis.[90] However, in most real-life situations the number of findings to be checked can quickly get out of hand. A typical case where this occurs is when the analyst is interested in testing the robustness of a poverty profile. A poverty profile examines "the pattern of poverty to see how it varies by geography (by region, urban or rural, mountain or plain, and so on), by community characteristics (for example, in communities with and without a school), and by household characteristics (for example, by education of household head or by household size). Hence, a poverty profile is a comprehensive poverty comparison, showing how poverty varies across subgroups of society." (Haughton and Khandker 2009, 122; Ravallion and Bidani 1994, 75). When sensitivity comes into play, questions like "does the poverty ranking between urban and rural areas change when the consumption aggregate is computed differently? What about the ranking between large and small households? Between households headed by men and women?" quickly pile up.

An effective way of summarizing the robustness of key findings in a poverty profile is shown in table 8.2, which will be referred to as a *robustness matrix*.[91] Each row in the table represents a finding or a statement whose robustness is under investigation; each column represents a methodological aspect whose impact on findings is being checked. Cell (1,1), where row 1 and column 1 meet, provides the answer to the question: is the finding "rural poverty is higher than urban poverty" robust to the choice of an outlier treatment procedure that is different from what is chosen as the "baseline"? The checkmark in cell (1,1) indicates that the answer is positive, that is, the poverty ranking of urban and rural areas is not affected by how outliers are treated. As a counter example, cell (1,2) of

[90] While table 8.1 only refers to headcount poverty rates, similar tables can be produced for testing the sensitivity of any statistic of interest. For example, one may wish to check the sensitivity of other poverty measures (the poverty gap index, the poverty gap squared, etc.), but also of inequality measures, mean expenditures, and so on.

[91] See Bosnia and Herzegovina (2003, 65), Table 6.2, which inspired the robustness matrix in table 8.2. See also D'Alessio (2020).

the robustness matrix indicates that rural areas are no longer found to be poorer than urban areas if we base our estimates on a nominal aggregate, that is, if we forego spatial deflation. In this instance, the conclusion is that the urban-rural poverty ranking is not robust to the choice of adjusting for cost-of-living differences. By extension, when a row of the matrix contains all checkmarks, the corresponding finding is robust to each of the methodological tweaks considered in the columns—or, more pragmatically, qualifies as a "story" that can be safely included in the executive summary, given that it is not vulnerable to discretionary assumptions. This reassures the analyst, as well as the reader. This is the case of row 4 in table 8.2, for example. On the other hand, when one or more x marks are present in a row, caution is needed before highlighting the corresponding result as a solid finding—in the example of table 8.2, this is the case for row 1, for example, which reports that the finding "rural poverty is higher than urban poverty" critically depends on the spatial cost-of-living adjustment.

TABLE 8.2. The robustness matrix

	Baseline findings	Data	Price adjustment	Adjustment for household composition	
		Imputation of outliers *1*	No spatial deflation *2*	OECD-II scale *3*	DZ $\alpha = 0.25$, $\theta = 0.9$ *4*
	Geography				
1	Poverty incidence is higher in rural than in urban areas	✓	×	✓	✓
	Demography				
2	Poverty incidence increases with the age of the household head	✓	✓	×	×
3	Elderly face a poverty risk lower than children	✓	✓	×	×
4	Poverty incidence is the same for men and women	✓	✓	✓	✓
	Education and labor				
5	Poverty incidence decreases as educational status increases	✓	✓	✓	✓
6	Poverty is higher for the employed than for those not in employment	×	✓	✓	✓
	Housing				
7	Poverty is higher among tenants than among homeowners	✓	✓	✓	✓
	Miscellaneous				
8	Poverty is higher among refugees	✓	✓	✓	✓

LEGEND: ✓ indicates that a finding (row) is confirmed under a certain methodological modification (column); × indicates that it is not.
NOTE: The table depicts a hypothetical example of sensitivity analysis. By 'baseline findings', we indicate results obtained using a baseline consumption aggregate (in this example, one with no outlier imputation, with spatial deflation, and expressed in per capita terms). The columns of the table indicate changes to the baseline methodology whose impact on the poverty profile is being checked.
SOURCE: Authors' elaboration.

In some cases, the use of more sophisticated methods of sensitivity analysis can offer further insight. DZ's *Guidelines* demonstrated how the concept of *stochastic dominance* (Shorrocks 1983; Atkinson 1987; Foster and Shorrocks 1988) could be useful for the purposes of sensitivity analysis, but the last 20 years show that this set of tools failed to become as widespread as it perhaps should have (this point is discussed further in section 8.4). In section 8.2, we give it another chance. The computational burden of stochastic dominance analysis is now close to zero, thanks to the availability of standard statistical packages: there really is no good reason for it not to be a fixture of applied welfare analysis (Garciá-Gómez, Pérez, and Prieto-Alaiz 2019).

8.2 Curves: Stochastic dominance analysis

The starting point for understanding the concept of stochastic dominance is the cumulative distribution function (CDF), arguably the single most useful weapon in the welfare analyst's arsenal (see, for instance, Duclos and Araar 2006, and Chakravarty 2019). A formal definition of the CDF can be found in any statistics textbook, and goes along the following lines: if X is any random variable, then its CDF, indicated by $F(x)$, is defined for any real number x as the probability that X is less than or equal to x :

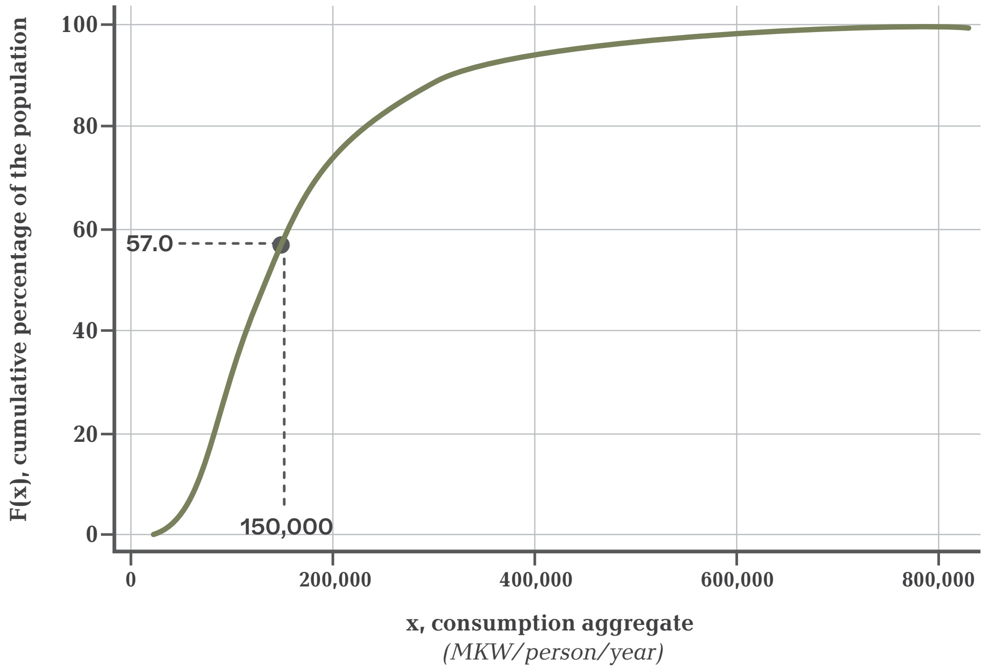
$$F(x) = P(X \leq x) \tag{8.1}$$

In practice, if X is expenditure, then equation (8.1) says that the CDF measures the proportion of individuals with expenditures of at most x. Figure 8.1 provides a graphical representation of the CDF.[92] To fix ideas, imagine that X is real consumption per capita, and x is any monetary amount measured in, say, Malawian kwacha. If, for instance, we take x to be 150,000 kwacha, then $F(150{,}000) = 0.57$ is the proportion of the population for which annual real consumption per capita is less than or equal to 150,000 kwacha. This is typically reported as a percentage (57 percent). By construction, the CDF ranges between 0 (nobody has less than the minimum expenditure observed in the data) and 1 (everybody has less than the maximum), it is monotonically increasing, and typically sigmoid-shaped (initially convex and then concave), as in figure 8.1. Familiarity with these concepts is important for the discussion that follows.

The significance of the CDF for welfare analysts becomes apparent when one considers that when the value chosen for x (horizontal axis) is equal to the poverty line—we typically indicate this value with the letter z—then $F(z)$ on the vertical axis corresponds to the poverty headcount rate, H. If 150,000 kwacha were the value of the national poverty line for the country depicted in figure 8.1, then 57 percent would be the estimated share of the population living in poverty for that year. Because of this interpretation of the CDF as a curve that plots the poverty headcount rate for any given value chosen for the poverty line, it is sometimes referred to as the *poverty incidence curve* (Ravallion 1994, 67).

[92] Figures 8.1 to 8.4 shown in this section are based on Malawi's 2016 Fourth Integrated Household Survey (IHS4), as available from RuLIS (2020)—data extracted in December 2020. In some cases, the data have been adapted to facilitate the illustration of the points covered in the discussion; the graphs do not reflect official results and statistics.

FIGURE 8.1. The empirical cumulative distribution function (CDF)



NOTE: MWK stands for Malawian kwacha. The figure is purely illustrative and does not reflect official statistics.
SOURCE: Authors' elaboration based on Malawi's 2016 Fourth Integrated Household Survey (IHS4).

In what way is the CDF useful for sensitivity analysis? Suppose we construct a consumption aggregate, which we denote by X_1. Let X_2 denote a second consumption aggregate, born out of different methodological decisions. Take the following case as an example: unlike X_1, which we may think of as a nominal aggregate, X_2 may be adjusted for cost-of-living differences by means of a spatial price index. The goal is to assess the impact of such a decision (to deflate or not to deflate?) on the estimates of poverty incidence. Let $F_1(x)$ and $F_2(x)$ denote the CDFs corresponding to each aggregate. We say that $F_1(x)$ has *first-order stochastic dominance* (FOD) over $F_2(x)$ if we observe that:

$$F_2(x) \geq F_1(x) \quad \text{for any } x \tag{8.2}$$

Equation (8.2) contains no mistakes: the curve that dominates is the one *below*, not above, as most of us might assume, going by the common meaning of the word "dominance." This is illustrated in figure 8.2 (panel a), where the two curves are plotted on the same graph. It all makes sense once we realize that, irrespective of where one draws the poverty line, the proportion of individuals below the line itself (the headcount poverty ratio) is always lower according to F_1 than F_2. In this sense, we would all prefer a society described by F_1 over one described by F_2—hence, F_1 dominates F_2. Panel b of figure 8.2 will come into play shortly.

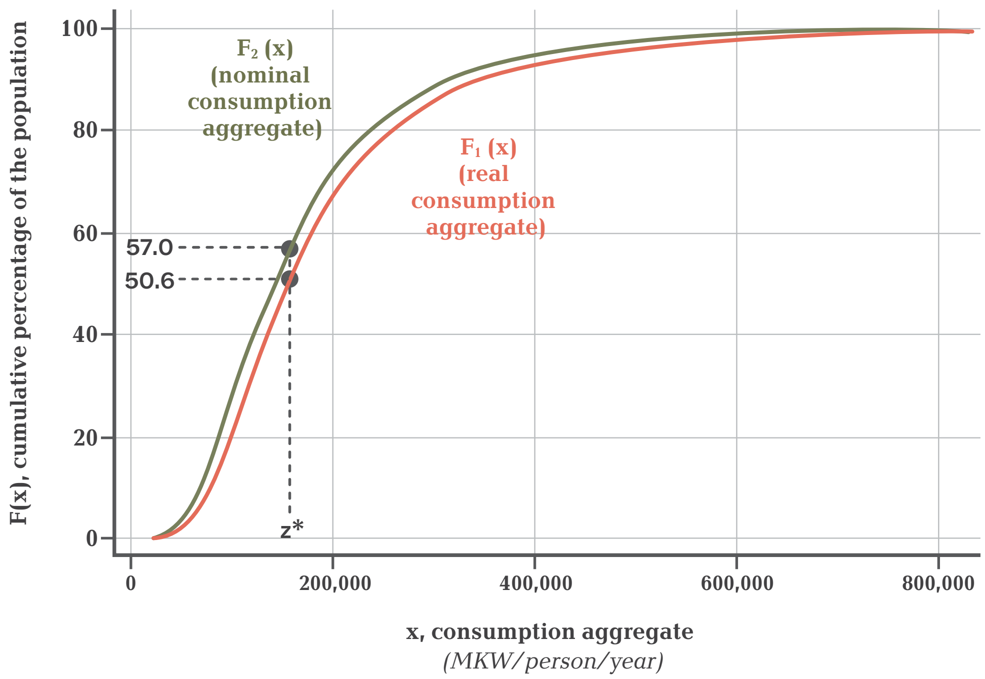
How can FOD help determine whether estimated poverty is sensitive to the choice between X_1 and X_2? If we take the poverty line as given and fixed—say at the value

$z^\star$ =150,000 kwacha—then the answer is straightforward.[93] In that case, the vertical distance between the two curves evaluated at $z^\star$, that is $F_1(z^\star) - F_2(z^\star)$ or, equivalently, $H_1 - H_2$, tells us precisely the extent to which the choice between the two consumption aggregates affects the headcount poverty rate. In our example, $H_1 - H_2 = 50.6 - 57.0 = -6.4$: spatial deflation is responsible for a decrease of 6.4 percentage points in poverty incidence, and in this sense the vertical distance between the two curves *at the poverty line* $z^\star$ shows how sensitive the estimated poverty rate is to the choice of adjusting the consumption aggregate for price differences.

Admittedly, this does not add much to simply tabulating poverty headcount rates, as in table 8.1 (section 8.1). The power of FOD is in giving us the answer to the question "is poverty higher with X_1 or X_2?," for *any* value of the poverty line. If $F_1(z)$ lies below $F_2(z)$ *for all values of z*, as in the panel a of figure 8.2, then $H_1 \leq H_2$ for any choice of poverty line. This way, a *poverty ordering* is established according to which poverty is lower when welfare is measured using X_1 than when using X_2, irrespective of where the line is drawn. This is quite useful because the poverty line itself is a discretional methodological choice, as we shall see in section 8.3.

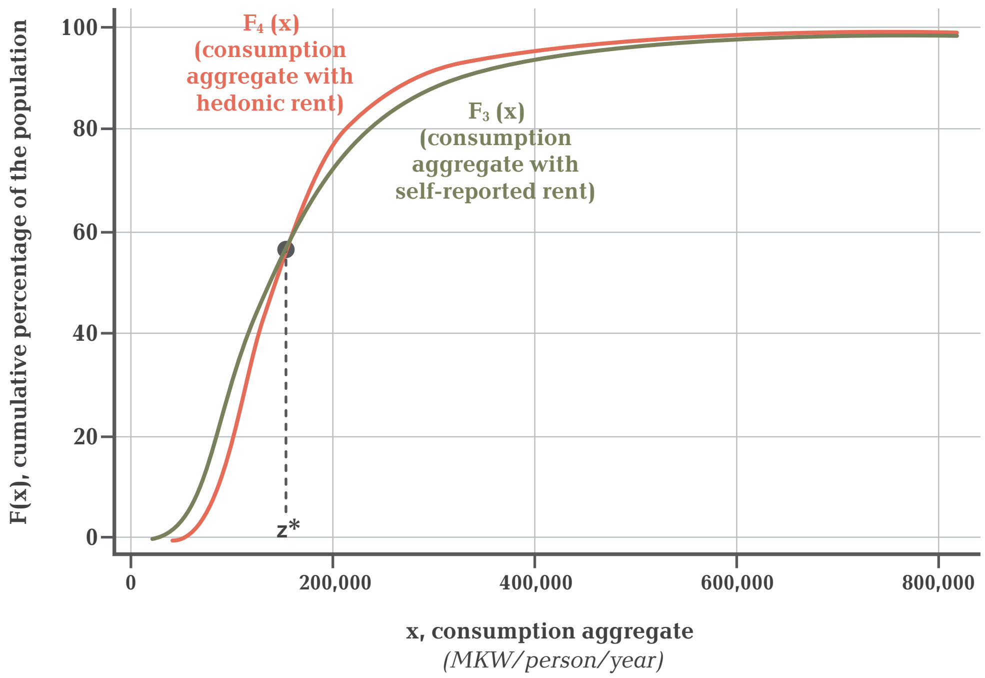
FIGURE 8.2. An illustration of first-order stochastic dominance

a. *$F_1(x)$ first-order stochastically dominates $F_2(x)$*



93 The analysis in the text follows DZ's setup, and assumes an exogenous poverty line.

b. $F_3(x)$ does not first-order stochastically dominate $F_4(x)$



NOTE: MWK stands for Malawian kwacha. The figure is purely illustrative and does not reflect official statistics.
SOURCE: Authors' elaboration based on Malawi's 2016 Fourth Integrated Household Survey (IHS4).

Unfortunately, the scenario depicted in panel a of figure 8.2, where the CDFs do not cross, is not guaranteed to occur in practice. Panel b of figure 8.2 shows another comparison between consumption aggregates, X_3 and X_4. This time we imagine the analyst evaluating a different methodological aspect: the estimation of imputed rent. One can think of X_3 as the consumption aggregate where imputed rent is the owners' self-reported value, while for X_4 imputed rent has been predicted using a hedonic regression model. FOD cannot be established in this instance: the CDFs *intersect*, and do so almost exactly where the poverty line is set. This gives rise to a most unfortunate situation, one where even a small variation of the poverty line would imply a ranking reversal between alternative consumption aggregates. Any candidate value for the poverty line to the left of z^* would lead to the conclusion that poverty decreases when switching from X_3 to X_4 (F_4 dominates F_3), while any value to the right of z^* would, on the contrary, imply an increase in poverty (F_3 dominates F_4).[94]

Dealing with one curve is simpler than dealing with two curves: it is often convenient to plot the vertical distance between two CDFs, rather than the CDFs themselves, to focus more closely on the difference between the poverty rate estimated in one scenario versus another.

[94] See, for instance Amendola and Vecchi (2008), where the stochastic dominance analysis of second and higher orders (Davidson and Duclos 2000) is used to solve a situation like the one depicted in panel b of figure 8.2.

Figure 8.3 is the exact counterpart of figure 8.2, in that the underlying data are the same, but what appears in the graph is the following:

$$\Delta H_{12} = F_1(Z) - F_2(Z) = H_1 - H_2 \text{ (panel a)}$$

$$\Delta H_{34} = F_3(Z) - F_4(Z) = H_3 - H_4 \text{ (panel b)}$$

For brevity, we may call these *headcount difference curves*.[95]

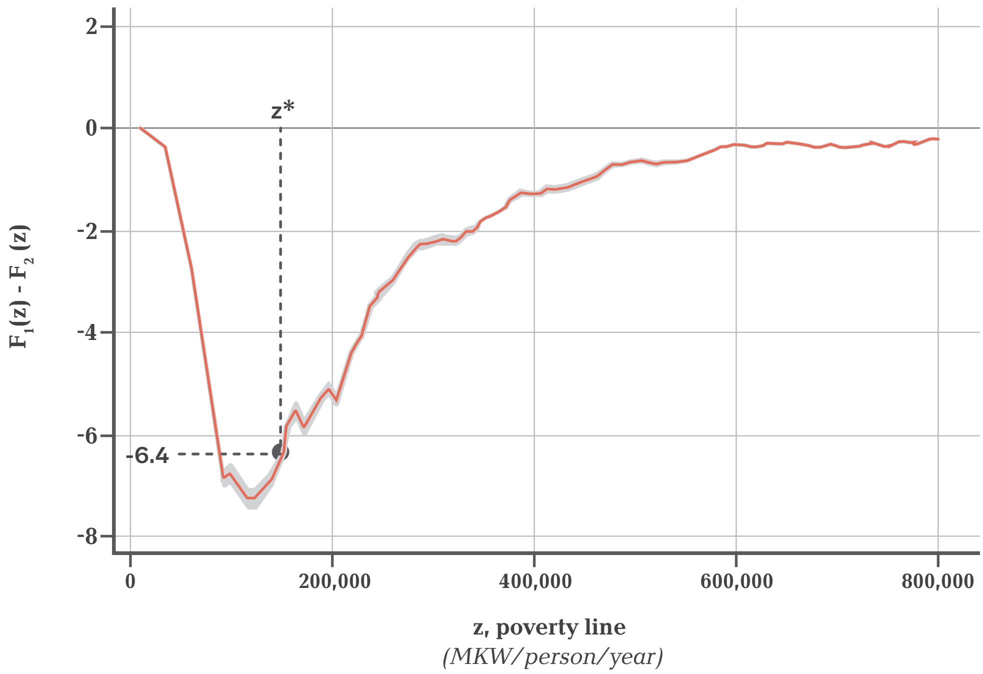
Let us see how the story that emerges from panel a of figure 8.2 is retold by panel a of figure 8.3. Our eye should focus on three main aspects. First, the *position* of the curve relative to the horizontal axis reflects the *poverty ordering* between the two alternative consumption aggregates. In figure 8.2, this could be inferred by the relative position of the two CDFs (one curve lies above the other), while in figure 8.3, it can be inferred from the fact that the curve ΔH_{12} always lies below zero: $\Delta H_{12} = H_1 - H_2 \leq 0$ implies that $H_1 \leq H_2$, for any choice of the poverty line, that is, the poverty rate measured on the basis of (in our example) the nominal aggregate X_1, is always lower than the poverty rate based on the real (price-adjusted) aggregate. Second, the *distance* of the curve from the horizontal axis makes it easy to gauge the *magnitude* of the impact of spatial deflation: at the poverty line z^*, the difference between alternative poverty rates amounts to 6.4 percentage points (=57.0 − 50.6, as from the labels added to figure 8.2); now the headcount difference curve in figure 8.3 makes it easier to see that this difference varies, according to where the poverty line is drawn. If the line were slightly lower than z^*, the impact of spatial deflation on final estimates would be even greater, while it would slowly wane if the threshold were moved up instead. Third, the *confidence bands* that surround the curve offer a simple and immediate way of testing the *significance* of differences between headcount rates (Chen and Duclos 2011): in the case at hand, the band does not intersect the horizontal axis, indicating that observed differences are significantly different from zero.

The same reasoning can be applied to panel b of figure 8.3. This is the difference $\Delta H_{34} = H_3 - H_4$, corresponding to aggregates X_3 (which in our example includes self-reported rent), and X_3 (which includes hedonic rent). The *position* of the curve is not everywhere above or below the x-axis: instead, there is a crossing, showing that for some value of the poverty line (which, coincidentally, is almost exactly equal to z^*), there is a *reranking* of poverty rates obtained from the two alternative welfare aggregates. In general, in the absence of (first-order) stochastic dominance, that is, when the curve crosses from positive to negative territory (or vice versa), we cannot say for sure which method of estimation of rent yields higher estimated poverty. The *distance* of the curve from the x-axis tells us that the *magnitude* of the impact of rent estimation is close to zero when the poverty line is z^*, but that it could be 9 percentage points if the line were set around 100,000 kwacha, or −4 percentage points for a line set at 200,000 kwacha. *Confidence bands* signal that most of the observed differences are significant.

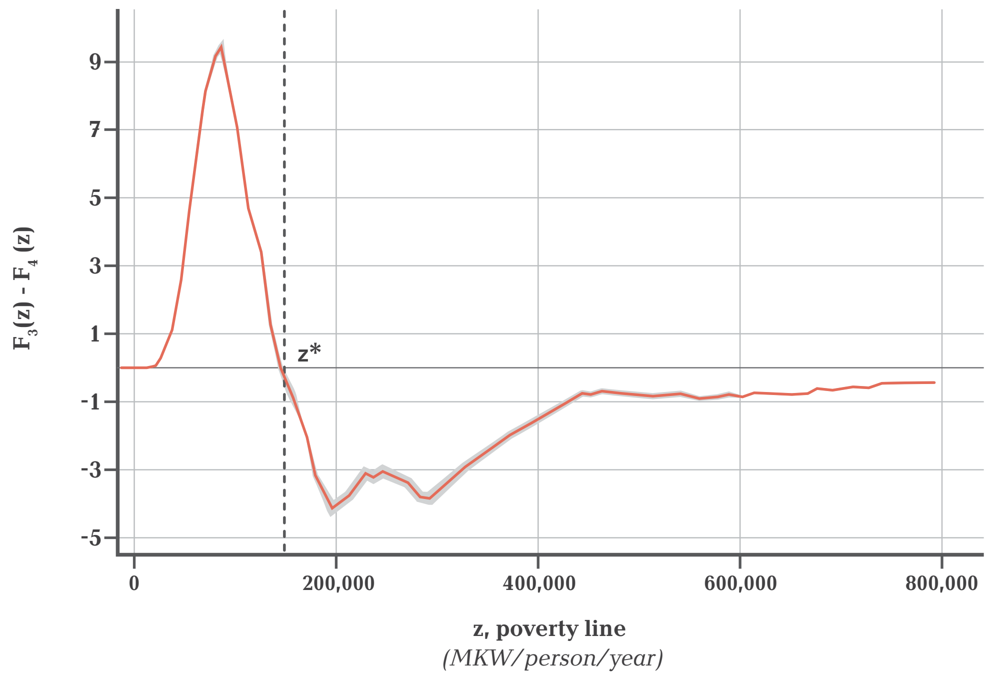
[95] See Duclos and Araar (2006, ch. 10) for an excellent, even if mathematically more demanding, discussion which extends the stochastic dominance analysis beyond headcount ratios.

FIGURE 8.3. Headcount difference curves

a. $\Delta H_{12} = F_1(z) - F_2(z) = H_1 - H_2$



b. $\Delta H_{34} = F_3(z) - F_4(z) = H_3 - H_4$



NOTE: MWK stands for Malawian kwacha. The figure is purely illustrative and does not reflect official statistics. Gray bands represent the 95 percent confidence interval of Δ*H*. The graphs were produced using the DASP Stata package (command `cfgts2d`).
SOURCE: Authors' elaboration based on Malawi's 2016 Fourth Integrated Household Survey (IHS4).

To recap, panel a of figure 8.3 is telling us that the poverty headcount rate is sensitive to spatial deflation, but that we can be absolutely sure of the impact of this choice: when the consumption aggregate is adjusted for price differences, estimated poverty is sure to decrease, and the size of the change can be expected to be between 4 to 7 percentage points (under the assumption of using values in a reasonable range around the "official" poverty line, z^*). In contrast, panel b of figure 8.3 tells us that we cannot really say anything definitive about the impact of how imputed rent is calculated on the incidence of poverty. The choice between self-reported rent over hedonic rent could significantly increase the headcount, leave it unchanged, or decrease it.

The headcount difference curve, ΔH, is even more useful when the goal is to test the sensitivity of the *poverty profile* to the use of different methods. This is, in fact, the application originally suggested by DZ (DZ 2002, 57). In figure 8.4 we simulate a common scenario in some detail, one where the focus is on the adjustment for household size and composition. The situation could be described as one where the analyst asks the following questions: "What is the impact of the choice of an equivalence scale on poverty *comparisons*? For instance, is the finding that rural poverty is higher than urban poverty robust to this choice? What about the fact that children are found to be poorer than older people: does it depend on the chosen equivalence scale?" The headcount difference curve in panel a of figure 8.4 shows the difference between headcount poverty rates in rural and urban areas, calculated using a per capita versus a per adult equivalent welfare measure. More precisely, the two curves in panel a are defined as follows:

$$\Delta H_{ru} = H_r - H_u = F(x|rural) - \mathrm{F}(x|urban)$$

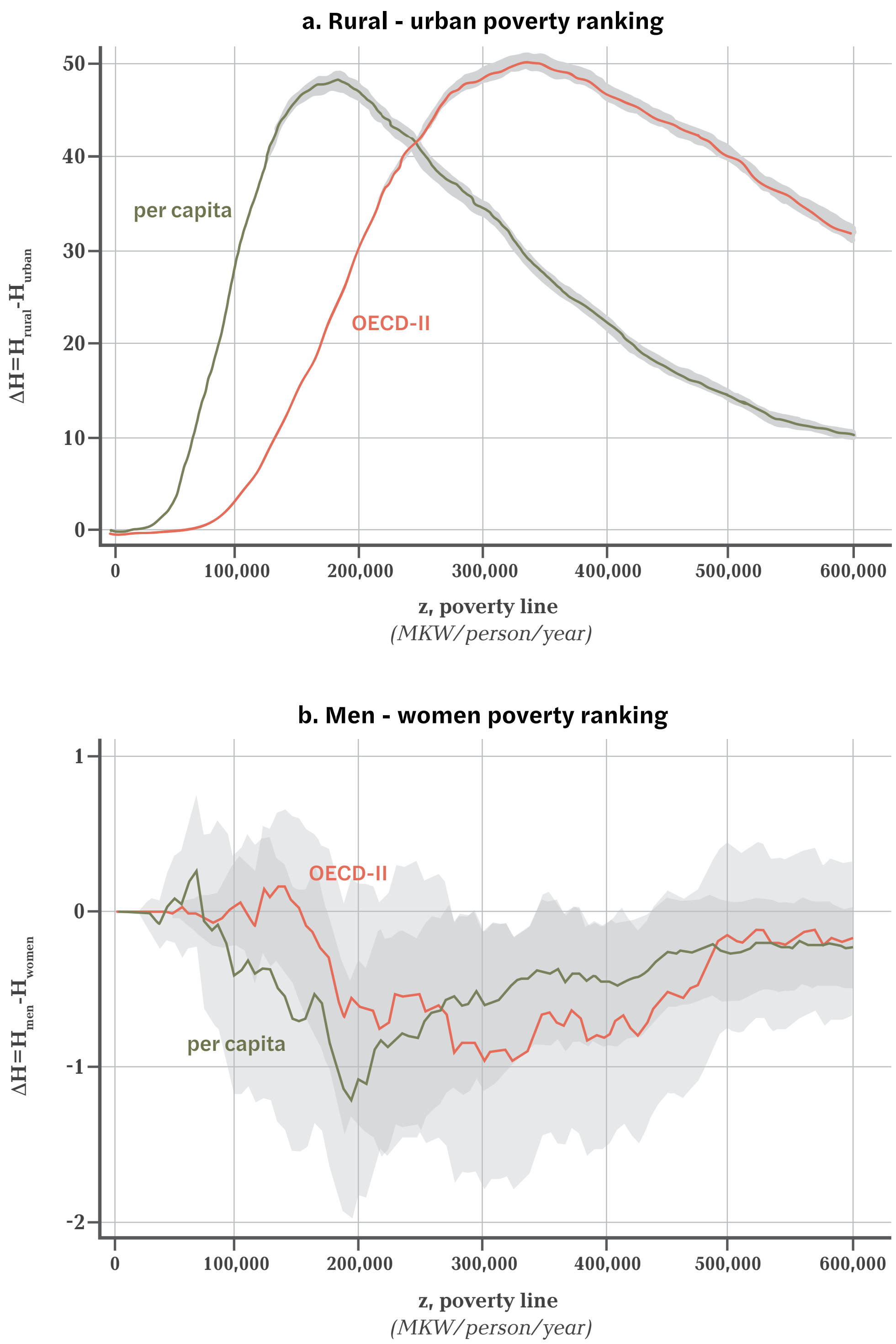
For each curve, the underlying welfare aggregate is defined in a different way: in per capita terms (blue solid line) and in per adult equivalent terms (red solid line), using the OECD-II equivalence scale. Both curves are positive for all values of x on the horizontal axis—this indicates that it is always the case that $H_r > H_u$, that is, the urban-rural poverty ranking is robust to the method of adjustment of the welfare indicator. On the other hand, the vertical distance between two curves is substantive, which suggests that the impact of the adjustment on poverty levels is also substantive.[96]

On the other hand, the "elderly vs. children" comparison in panel c of figure 8.4 is not as robust: in this case, the poverty ranking between individuals over 75 (the elderly) and under 15 (children) is reversed by the choice of an equivalence scale. When using a per capita aggregate, children are estimated to be poorer (the line is always in the positive), while the reverse happens with a per capita indicator.

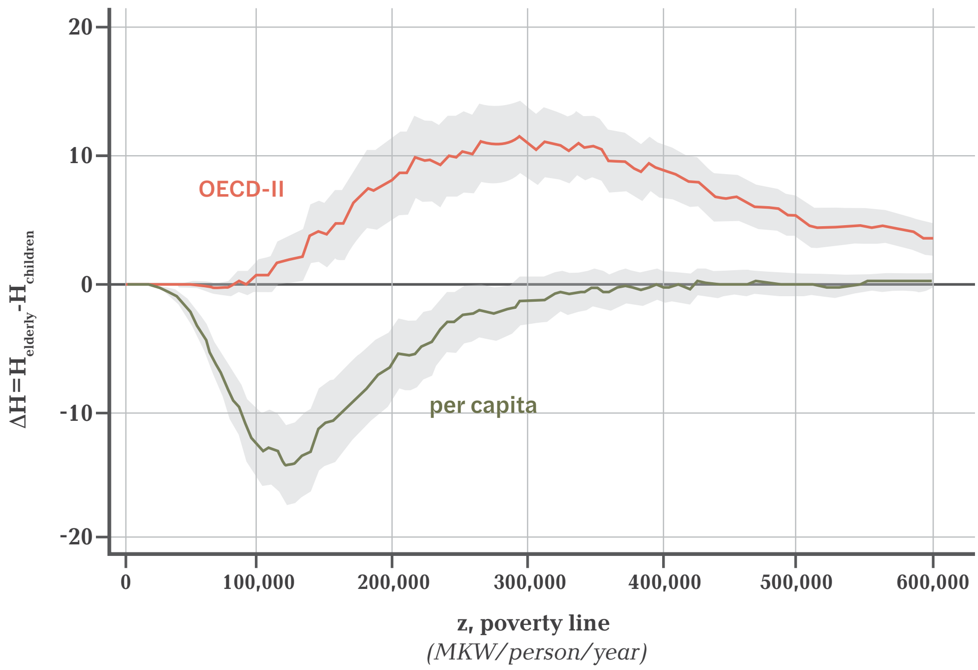
From panel b, we learn that men are not found to be significantly poorer than women, regardless of the chosen scale (both confidence bands overlap with zero); while panel d, comparing employed and not employed working-age individuals, tells a similar story as panel a—the fact that poverty is higher among the employed than among those out of the labor force can be thought of as robust (both curves and their confidence bands lie above the zero line).

[96] The comparison per capita versus per adult equivalent poverty rates requires the adjustment of the poverty line—see DZ (2002, 61) for details.

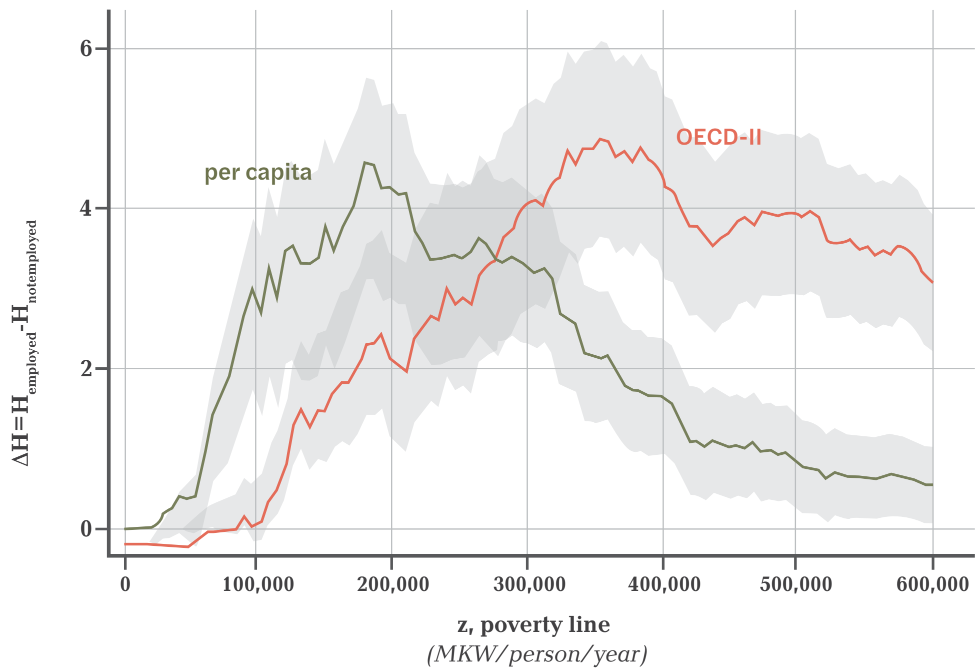
FIGURE 8.4. Headcount difference curves and the robustness of the poverty profile to the choice of the equivalence scale



c. Elderly - children poverty ranking



d. Employed - not employed poverty ranking



NOTE: MWK stands for Malawian kwacha. The figure is purely illustrative and does not reflect official statistics. Gray bands represent the 95 percent confidence interval of ΔH. The graphs were produced using the DASP Stata package (command `cfgts2d`).
SOURCE: Authors' elaboration based on Malawi's 2016 Fourth Integrated Household Survey (IHS4).

The results of the analysis summarized in figure 8.4 could find their place in the robustness matrix, (table 8.2 in the previous section): the results in the four graphs would translate into three checkmarks (panels a, b, and d), meaning that the poverty ranking is robust, and one x mark (panel c), meaning that the ranking is sensitive to the choice of the equivalence scale, in column 3 of the table.

We left the discussion of how to generalize the headcount poverty curve introduced in this section for last, because they have been illustrated with great clarity in Ravallion (1994), Deaton (1997), Lambert (2001), and Duclos and Araar (2006, ch. 10). Analysts interested in going beyond the headcount poverty curves can easily produce graphs similar to those in figure 8.4 for the poverty gap index, the poverty gap squared index, and other higher-order poverty measures.

8.3 Sensitivity to the choice of the poverty line and poverty measure

When a new poverty line is proposed, no matter how technically sound and politically participated its estimation, it is almost invariably met with criticism and skepticism: things would be different if a different poverty line were used, the argument goes. But how different?

The sensitivity of poverty estimates to the choice of the poverty line can be gauged by means of a simple method, illustrated in table 8.3.[97] Using selected examples from published poverty reports for each country, the table shows (1) the "official" poverty headcount ratio (row labelled "Actual"), and (2) the poverty rates that would be observed by moving this line up or down by 5 percent, 10 percent, and 20 percent. Considering the case of Armenia, for example, the table shows that a 10 percent increase in the poverty line would imply an increase in the incidence of poverty from 29.8 percent to 30.9 percent, that is, +0.37 percentage points, corresponding to an elasticity of 0.4. This value is relatively low compared to the one resulting from the same calculation for Botswana (1.4), or Mongolia (2.3), or Lao People's Democratic Republic (2.6)—poverty incidence rates in Armenia therefore qualify as "robust to the choice of the poverty line."

The results of sensitivity checks like those shown in table 8.3 critically depend on the shape of the CDF of the welfare aggregate, in particular on how steep the curve is in the neighborhood of the poverty line: the steeper the curve, the more the incidence of poverty will react to small variations in the poverty line (Ravallion and Huppi 1991, 66−67). Some analysts interpret this measure—the elasticity of poverty headcount to changes in the poverty line—as a proxy of the *vulnerability* to poverty. The reasoning is that the fraction of the population situated just above the poverty line is "vulnerable" to becoming poor after a small shock to the line itself (or, equivalently, after a small shock to incomes/expenditures).[98] As noted by Ravallion (2016, 258), this is a deceptive label (measuring vulnerability implies estimating

[97] As shown in the previous section, stochastic dominance analysis is the chief instrument for analyzing the sensitivity of poverty *comparisons* to changes in the poverty line—see for instance the comment to figure 8.2.

[98] See Dercon (2005b) and Gallardo (2018).

or modelling the risk element that is inherent to it), but if the approach shown in table 8.3 is used with this caveat in mind, then it can provide useful insights regarding the robustness of poverty estimates to the choice of a poverty line (Foster et al. 2013, ch. 3).

TABLE 8.3. Sensitivity of the headcount poverty rate (percent) to the choice of the poverty line

Poverty line	**Armenia 2015 [1]**	Change from actual (%)	**Botswana 2009/10 [2]**	Change from actual (%)	**Mongolia 2016 [3]**	Change from actual (%)
+20%	36.1	*21.1*	25.4	*31.3*	42.2	*42.6*
+10%	30.9	*3.7*	22.2	*14.4*	36.3	*22.6*
+ 5%	30.0	*0.7*	20.8	*7.3*		
Actual	**29.8**	***0.0***	**19.4**	***0.0***	**29.6**	***0.0***
−5%	22.4	*−24.8*	17.7	*−8.8*		
−10%	16.5	*−44.6*	16.2	*−16.2*	23.1	*−22.0*
−20%	9.2	*−69.1*	12.8	*−34.1*	17.1	*−42.2*
Poverty line	**Mozambique 2014/15 [4]**	Change from actual (%)	**Lao PDR 2013 [5]**	Change from actual (%)	**Madagascar 2012 [6]**	Change from actual (%)
+20%	57.2	*24.1*	34.9	*50.2*	78.4	*10.9*
+10%	51.7	*12.1*	29.2	*25.7*	74.9	*5.8*
+ 5%	49.0	*6.3*	26.4	*13.4*	73.0	*3.2*
Actual	**46.1**	***0.0***	**23.2**	***0.0***	**70.7**	***0.0***
−5%	42.9	*−6.9*	20.4	*−12.3*	68.3	*−3.4*
−10%	39.7	*−13.9*	17.6	*−24.3*	65.5	*−7.4*
−20%	33.1	*−28.2*	12.4	*−46.7*	59.9	*−16.3*

NOTE: [1] Armenia (2016, 44); [2] Botswana (2015, 31); [3] Mongolia (2017, 11); [4] Mozambique (2016, 14); [5] Lao PDR (2014, 46); and [6] Madagascar (2016, 34). See Appendix A for full references.

SOURCE: Authors' calculations.

A last remark is on the sensitivity of poverty rates to the choice of the poverty *index*, a point that has received less attention than the choice of the poverty *line*. Choosing a poverty index is an arbitrary choice: many indices exist, and most of them have been shown to be sensible choices (Zheng 1997). The Foster, Greer, and Thorbecke (FGT) (1984) class of poverty measures, which includes the headcount ratio (capturing the *incidence* of poverty), the poverty gap index (capturing the *depth* of poverty), and the poverty gap squared index (capturing the *severity* of poverty) has become the standard for international evaluations of poverty. These measures are reported by the World Bank PovcalNet, a host of UN agencies, and countless individual countries (Foster, Greer, and Thorbecke 2010, 492). Given the different weighting schemes (value judgments) underlying the FGT indices (Ravallion 1998), reporting and comparing the three estimates constitute a simple way of testing the sensitivity of results to the choice of poverty index, and should become routine in poverty analysis.

In fact, other poverty measures could be usefully added to the list: one is the Watts index (Watts 1968), long neglected by poverty analysts (Ravallion 2016, 233—236), but later brought back to the fore due to its desirable properties (Zheng 1993) and interesting economic interpretation (Morduch 1998). Others are the Sen-Shorrocks-Thon (SST) index (Shorrocks 1995; Osberg and Xu 2001) and the Clark-Hemming-Ulph (CHU) family of indices (Clark, Hemming, and Ulph 1981). It is good practice to estimate Watts, SST, and CHU indices, and follow the recommendation in Foster et al. (2013, 204): "if these [alternative] measures, capturing different aspects of poverty and inequality among the poor, agree with the results from the measures in the FGT class, then the poverty analysis is robust. In contrast, if these measures do not agree with each other, the policy conclusion should be drawn with more care."

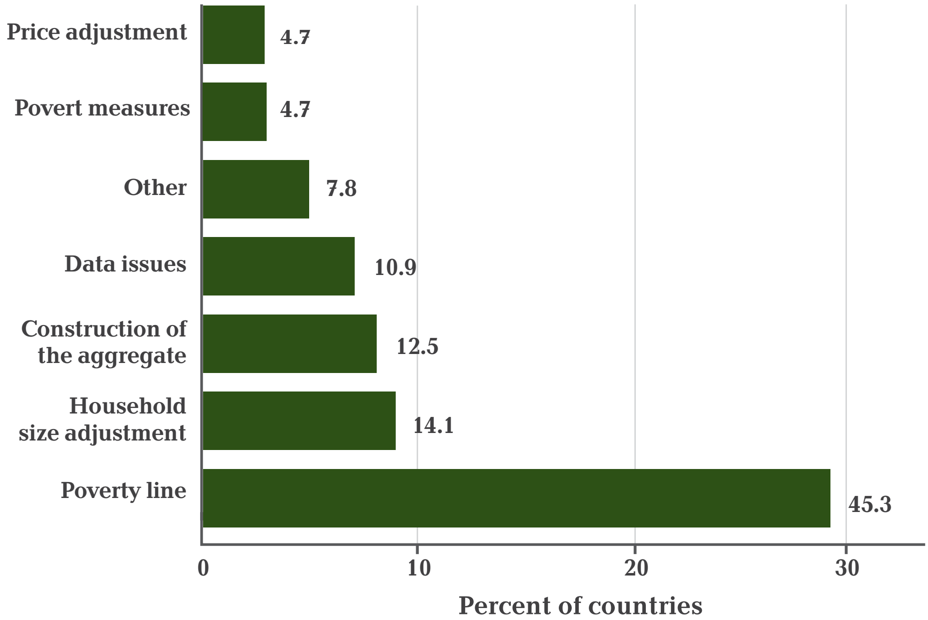
8.4 Discussion

Welfare measurement entails value judgments: a number of parameters—including equivalence scales, price indices, imputation methods, poverty thresholds, and so forth—must be selected by the analyst, who in doing so, exercises at least some degree of discretion. This tends to attract criticism. Responding to criticism requires an assessment of the impact that discretionary choices have on final estimates. The natural instrument for this is sensitivity analysis, but we find that, despite a general consensus on its desirability, it has not received due attention in applied work.[99] The inspection of some 220 plus official publications reporting inequality and poverty estimates shows that only one out of five includes a chapter or an appendix devoted to sensitivity analysis. A first recommendation is that *systematic* sensitivity analysis should be integral to the process of constructing a welfare aggregate and producing final estimates. A section or appendix dedicated to sensitivity testing should become the norm for any technical report presenting inequality and poverty estimates.

A second recommendation concerns the object of sensitivity analysis. Figure 8.5 shows that, among the reports that do include some sensitivity checks, almost half focus on the sensitivity of poverty estimates to changes in the poverty line, as discussed in section 8.3. We argue that this is good practice, and should also become a routine task. However, other less "eye-catching" analytical choices can have a potentially large impact on bottom-line findings. Only a minority of official reports engage with a *systematic* investigation of the extent to which results are sensitive to choices made when constructing the consumption aggregate, and, in particular, to the options discussed in sections 4 (aggregation), 5 (prices), and 6 (household size). Subjecting these choices (which ones depends on the data and the context) to closer scrutiny would only strengthen the credibility of results. To this, we shall add that it is also worth checking the sensitivity of poverty estimates to the choice of different poverty *indices*, not only by varying the "poverty aversion" parameter within the class of the FGT measures, but also by checking whether the overall FGT-based estimates are consistent with other poverty measures. Following Foster et al. (2013), we suggest the Watts index, and the Sen-Shorrocks-Thon and Clark-Hemming-Ulph poverty measures.

[99] The recent literature on multidimensional poverty indicators is a notable exception. Alkire et al. (2015), for example, devote an entire chapter to robustness analysis based on stochastic dominance.

FIGURE 8.5. Sensitivity analysis in official poverty reports



NOTE: Categories on the vertical axis indicate the methodological choices tested via sensitivity analysis. For instance, 45.3 percent of all checks focus on the sensitivity of poverty estimates to the poverty line; 14.1 percent are on the sensitivity of poverty estimates to household size adjustment procedures, and so on.
SOURCE: Our elaboration based on the database in appendix A.

A third recommendation concerns the methods available for performing sensitivity analysis. When the focus is on poverty estimates, stochastic dominance analysis stands out as the main tool for performing extensive sensitivity checks. That way the analyst can identify those findings that can be upgraded to "facts", and brought to the attention of policy makers. Operationally, side-by-side comparison tables, as well as the *robustness matrix* introduced in section 8.1 and the *headcount difference curve* in section 8.2 are simple yet insightful tools, that are useful when organizing the findings needed to craft an empirically robust poverty profile.

Inequality estimates and other statistics of interest (including summary statistics of the welfare aggregate itself) have not been mentioned as often as poverty estimates in this section, which belies their importance. The tools discussed in section 8.1 (tables 8.1 and 8.2) are, in fact, easily adapted to any statistic whose robustness is under examination.

A final note on the interpretation of sensitivity exercises. The analyst typically hopes to get a verdict of "robustness" out of her checks (DZ 2002, 65). Findings can then be declared sound, no matter the assumptions, and thus iron clad against criticism. As noted by DZ's *Guidelines*, however, even when results cannot be deemed robust—for instance, the adjustment for household composition usually matters a great deal—a sensitivity analysis can make for a better understanding of exactly how our choices impact final estimates, which goes a long way toward informing the decision of what should be done.

9. Reproducibility of Results

9.1 What is it?

If final results generated by the welfare analyst are to be trusted, the process leading to their production must be rigorous, transparent, participated, and open for debate. These are crucial features of any research endeavor and are usually described with the term *reproducibility*: in our context, the word is taken to indicate a situation where some external researcher with no prior knowledge of the analysis "can re-create the final reported results of the project, including key quantitative findings, tables, and figures, given only a set of files and written instructions." (Kitzes, Turek, and Deniz 2017).[100] This section focuses on how to make sure that the work of the welfare analyst is reproducible.

Reproducibility of the process leading from the raw data, to the construction of a welfare aggregate, to the production of final estimates, is important for at least three reasons. First, and perhaps most importantly, reproducibility is a tenet of the scientific method, and not only for the "hard" sciences. Pleas to devote more efforts to the reproducibility of applied economic research have been made by the highest echelons of the discipline (though not always heard). In the inaugural issue of *Econometrica*, one of the most important academic journals in the field, Ragnar Frisch—its first editor and future recipient of Noble prize in economics—went so far as to advocate for raw data to be shared publicly: "the original raw data will, as a rule, be published (...) This is important in order to stimulate criticism, control, and further studies" (Frisch 1933; 3, cited in Dewald, Thursby, and Anderson 1986, 588). The findings of welfare analysis are central to public debate and instrumental for policy makers: all the more reason for them to follow the rules of scientific rigor.

Second, reproducibility is a necessary condition for the *consistency* of welfare comparisons, intertemporal or otherwise. A regular occurrence is the release of data from a new survey wave, which almost invariably brings about the need to estimate time trends for poverty, inequality, and other statistics. The identification of time trends boils down to a replication exercise: the new methodology must be consistent with the one used to produce previous estimates, or no comparisons can be drawn. More often than not, the minutiae of what was done (whether this or that expenditure component was included, or how exactly some

[100] The term "reproducibility" is used interchangeably with "replicability" in some contexts, but the two words can also have distinct meanings. In this section, we stick to the definitions offered by the National Academies of Sciences, Engineering, and Medicine (2019): "*Reproducibility* means computational reproducibility—obtaining consistent computational results using the same input data, computational steps, methods, code, and conditions of analysis. *Replicability* means obtaining consistent results across studies aimed at answering the same scientific question, each of which has obtained its own data."

adjustment or imputation was made) are lost to time, or entrusted to the memory—perhaps prodigious, but never perfect—of some experienced colleague. Instead, building the analysis in a way that is reproducible at a later time, down to the smallest detail and without close guidance, is essential.

Third, analysis that is built to be reproducible tends to have certain characteristics—it is well-organized, easy to understand, and efficient—that are extremely valuable when working in a team. A welfare analyst rarely operates in isolation: more often, the work of "crunching the numbers" is shared in a group that includes analysts as well as data producers. In such a context, it is important that the analysis be easy to share, comment upon, and review collaboratively. Incidentally, this greatly increases the chance to spot and correct mistakes.

So, how to ensure that welfare analysis is reproducible, in practice? It might be useful to run a thought experiment where we imagine "handing over" our work to another researcher, or even to ourselves a few months from now, and have them replicate our results exactly. In principle, it should be enough to pass on two elements: the complete raw data, and a set of detailed instructions describing each and every step of the analysis. In practice, however, the way that data and instructions are organized and documented is at least as important. Are the data files easy to locate and clearly labelled? Do instructions take the form of computer scripts (Stata do-files, for instance) that run smoothly and are easy to interpret? There are countless examples of how failing to meet these apparently simple, but actually rather demanding conditions creates delays, mistakes, even complete failures in the day-to-day practice of research. Economists Matthew Gentzkow and Jesse Shapiro paint some evocative pictures of this kind of impasse: "In trying to replicate the estimates from an early draft of a paper, we discover that the code that produced the estimates no longer works because it calls files that have since been moved. When we finally track down the files and get the code running, the results are different from the earlier ones." Another example: "We are keen to build on work a research assistant did over the summer. We open her directory and discover hundreds of code and data files. Despite the fact that the code is full of long, detailed comments, just figuring out which files to run in which order to reproduce the data and results takes days of work. Updating the code to extend the analysis proves all but impossible. In the end, we give up and rewrite all of the code from scratch." (Gentzkow and Shapiro 2014, 4).

9.2 How to achieve it?

The combination of data, code, organization, and documentation can be described as the *workflow* of data analysis (Long 2009). "Everything from file and variable names to folder organization to data storage to efficient and readable programming is part of workflow" (Christensen, Freese, and Miguel 2019). The choices made by the analyst when setting up the workflow are central to determining whether or not her work is reproducible: a "bad" workflow can frustrate any attempt to piece together what exactly was done to obtain results. So, how does one build a "good" workflow? Although this question is not addressed as often as it should be in the typical educational curriculum of a social scientist, the literature is rife with good practices, down to the most minute details. For the purposes of this section, the discussion can be usefully limited to three "golden rules" suggested by Kitzes, Turek, and Deniz (2017), which we adapt slightly.

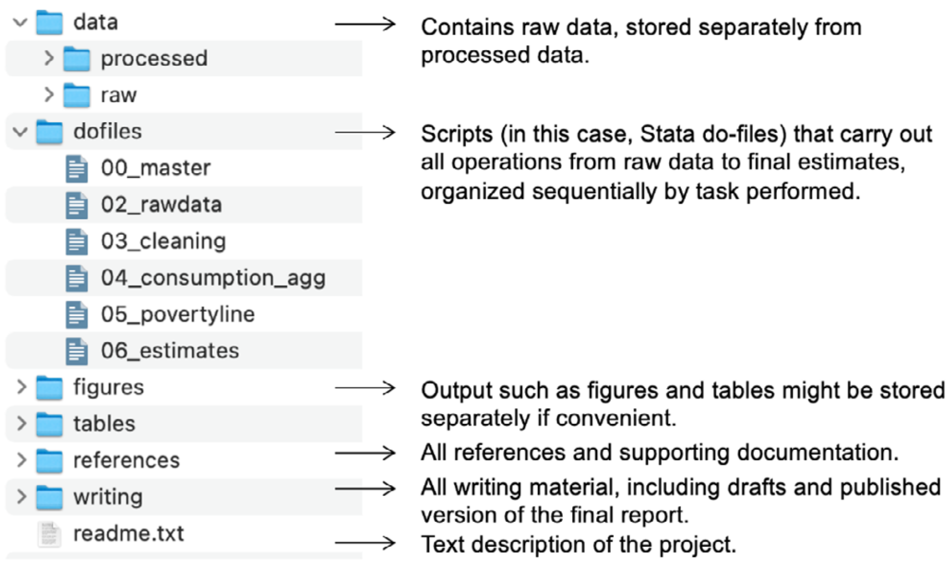
1. *Metadata*. Clearly describe, label, and document all data, files, and operations that occur on data and files. In practice, this principle involves archiving all materials related to the research project and linking them to appropriate supporting information that describes their content and purpose in detail.
2. *Automation*. Automate operations as much as possible, avoiding manual intervention in the workflow. Work that is done manually, by pointing and clicking (for instance, by entering or deleting data in Excel, or by using drop-down menus in Stata) is, by its own nature, undocumented. Programming scripts (files containing code in some programming language), on the other hand, are self-documenting. Given that most research work in applied economics is computational, scripts are the natural and best vehicle for keeping track of all operations that generate final estimates.
3. *Structure*. Design the workflow as a sequence of small steps, with intermediate outputs from one step feeding into the next step as inputs. A typical step-wise sequence will involve data acquisition, data preparation and "cleaning", data analysis, and the outputting of results for presentation. Even if, in practice, the analyst will often move up and down the steps, rather than completing them sequentially, it is useful to think of the workflow in terms of distinct stages, an follow this structure when organizing data, scripts and documentation.

These principles are simple, but powerful, if applied systematically. To be truly useful, however, they must be seen in action. What does a reproducible workflow look like, exactly, in the context of welfare analysis?

9.3 Guiding principles for workflow organization

Everything starts with organizing all files, scripts, and documents related to a project into a dedicated, self-contained directory structure. The fact that the folder includes *all* and *only* the material related to the project is important: this allows to share the whole workflow easily with others. For the same purpose, it is also useful for the structure of the folder to be standardized, predictable, ideally conventional within the team of collaborators, or, even better, the community of researchers in the field. Figure 9..1 illustrates a template for the project directory, and the rest of this section clarifies the use of its components.

FIGURE 9.1. Template for the project directory structure



SOURCE: Authors' elaboration.

Let us describe figure 9.1 in detail. The project directory contains four relevant sub-directories, separated by function:

1. data,
2. scripts (do-files in this example),
3. references, and
4. writing.

This basic partition may be made more complex if needed (for instance, figure 9.1 contains subdirectories for figures and tables), although experience suggests to keep things simple. However, there is one rule that should not be broken when organizing and archiving data files: *raw data should be kept raw*. For one, this makes the job of the analyst easier and safer: "It is tempting to overwrite raw data files with cleaned-up versions, but faithful retention is essential for rerunning analyses from start to finish, for recovery from analytical mishaps, and for experimenting without fear." (Wilson et al. 2017). It also facilitates future re-analysis by other researchers (Hart et al. 2016). In practice, following this rule means that raw data must be stored separately from processed data and never modified. Scripts use raw files to generate processed data, which are then saved elsewhere.

As for scripts, the names of the Stata do-files in figure 9.1 indicate that they follow the step-wise sequence advocated by the three "golden rules." This puts anyone in a position to understand what each script does at a glance, and makes it easy to track down any specific portion of the analysis that may be up for review.

FIGURE 9.2. The master.do file

```
1   /* ------------------------------------------------------------
2       New Project - Master file
3
4       Author
5       author@mail.com
6
7       March 1, 2022
8    ------------------------------------------------------------ */
9
10  //  Set working environment
11
12      version 17
13
14
15  //  Set paths and working directory
16
17      global path "/Users/`c(username)'/Documents/projects"
18
19      global wd        "$path/data"
20      global raw       "$path/data/raw"                    // raw data
21      global processed "$path/data/processed"              // processed data
22
23      cd "$wd"
24
25
26  //  Run analysis
27
28      do 00_master.do
29      do 01_rawdata.do
30      do 02_cleaning.do
31      do 03_consumption_agg.do
32      do 04_povertyline.do
33      do 05_estimates.do
34
35  exit
```

SOURCE: Authors' elaboration.

The way of organizing scripts shown in figure 9.1 enables an extremely convenient "one-click" approach to reproducing results. The "master" do-file, pictured in figure 9.2, is instrumental for this feature of the workflow. Here, one line of code (line 17) guarantees that, by customizing one's "path," the whole workflow can run smoothly from top to bottom. Line 17 should be the only user-specific line in the entire set of scripts; anywhere else in the code, the analyst should only use relative paths (those that appear in lines 19 — 22, and are defined on the basis of line 17). The rest of the code in the master do-file (lines 29 — 34) executes all other scripts, in sequence. This way, any new user can reproduce the entire analysis in two simple moves: (1) changing line 17 to the path of the "project" directory on her own computer, and (2) executing the master do-file (quite literally in one click).

A lot could be said regarding the content of all other scripts, that is, those that actually carry out data analysis. Writing "good" code—accurate, dependable, well organized, easy to read—is certainly important for reproducibility. Luckily there is no shortage of coding guidelines, both general and software-specific (checking out the Stata Press catalogue is always a good starting point): we see no advantage to duplicating them here.

The final element in figure 9.1 that is worthy of a comment is the "readme" file. Its name is truly meant as an invite: its goal is to attract the attention of anyone approaching the project for the first time (or after some time has passed), and to provide an introduction to it, in plain English. It is a simple text file that contains, at a minimum, a description of how the project directory is organized, and what each of its subdirectories contains. A brief description that walks the reader through the main steps of the analysis can also be included here, as well as any "warnings" that she should be aware of.

The structure of the project directory in figure 9.1 can be endlessly customized according to need and context.[101] More complex projects with a lot of moving parts typically require more complex directory structures, data analysis software comes in many flavors, and so on. What matters is that the discipline underlying figure 9.1 and figure 9.2 be followed "in spirit": it pays enormous dividends.

We close with a consideration on *open data* and the openness of the research process more generally. Reproducibility is closely linked with the notion of sharing: with colleagues, peers, the general public. If the ultimate goal of reproducibility as a research practice is to generate knowledge in a way that is transparent and participated, then it certainly matters whether the analyst's work—not just results, but data, scripts, and documentation—is freely accessible. There is little disagreement that openness and transparency are a good thing, but in practice, the incentives of analysts, data producers, and research institutions often get in the way. National statistic offices are typically constrained by confidentiality concerns, but in the last few decades there has been a progressive move toward public data availability (Dupriez and Boyko 2010). In this context, subscribing to the "golden rules" of reproducibility may seem like a small choice. Instead, it is a prerequisite for larger changes toward openness in the practice of welfare measurement.

While the discussion offered in this section is by no means comprehensive, it does highlight some basic principles that have direct practical implications for the hands-on work of the welfare analyst. Ultimately, regardless of the individual variations to the template proposed in this section, working in a reproducible manner is what makes the difference between producing new scientific knowledge and simply coming up with some (unsubstantiated) numbers.

[101] See, in particular, the Development Impact Evaluation (DIME) Project (World Bank 2020, ch. 1), and Christensen, Freese, and Miguel (2019, ch. 11).

10. Summary of Recommendations

Originally written for a small group of analysts (some 50 people or so) doing highly specialized work, Deaton and Zaidi's *Guidelines* (2002) have reached a much broader audience than expected. With the development of a massive data infrastructure for monitoring global poverty—the availability of household surveys suitable for poverty measurement worldwide has increased by more than two-and-a-half times over the course of 20 years, from the 1990s to the 2010s (Roser and Ortiz-Ospina 2017)—the need for an accessible "how-to" guide for analyzing this data has also grown. In this context, the *Guidelines* have become the key reference for the construction of a consumption-based welfare measure.

In this document we have addressed the question of whether DZ's recommendations hold up to 20 years of applied work and academic debates on the measurement of living standards. In light of our assessment, the 30 or so recommendations set out in the original paper (we counted as many, between explicit prescriptions and less visible suggestions peppered through the text) stand the test of time remarkably well. The overall longevity of DZ notwithstanding, we shall mention five areas where the current research and practice of welfare measurement has caught up with the *Guidelines.* One is that of the preference to be accorded to consumption over income as a monetary welfare indicator. While this report, like its predecessor, focuses on the construction of a consumption aggregate, both scholarship and international institutions have coalesced around a more impartial view. Openness toward the use of income will continue to grow in the future: the inclusion of a brief discussion on the construction of an income aggregate (appendix C) does not do justice to the issue, but it is a tangible step in that direction. Second, an area where the literature has made leaps in recent years is that of price deflation. A lively debate, addressing both theoretical and empirical questions, now looms over DZ's original discussion of adjustments for cost-of-living differences. Third, techniques for the estimation of housing expenditures and of the consumption flow from durable goods have been dissected by recent contributions: we are now in a position to offer more detailed guidance on these topics for those in the trenches of welfare analysis. A fourth aspect that this report emphasizes is the interconnectedness of data production (survey design, data collection, data processing) and data analysis. Arguably, the choices made before the data is "user-ready" have at least as much of an impact on final estimates as those made after. Analysts should be aware of, or better involved in, the decisions made since the inception of the survey and the trade-offs they imply. Fifth and finally, the importance of sensitivity analysis and reproducibility of the analyst's work cannot be overstated. Poverty and inequality estimates remain as politically sensitive and potentially controversial as ever, and analysts should spare no effort to make the analysis as sound, transparent, and participated as possible.

This document was not conceived as a purely academic exercise: its ambition is to complement the assessment of DZ's legacy with a constructive effort, and to formulate "updated" recommendations to complement DZ's, whenever justified by the available research. We feel that this task would not be fully accomplished without the pragmatic resolve that has contributed to making the *Guidelines* so successful. The boxes included in the original paper are a most useful contribution, particularly in the eyes of practitioners: the authors managed to distill complex debates and nuanced arguments into a set of actionable directions, each merely a few lines long. In the rest of this concluding section, we follow their lead and borrow the box format (see Boxes 10.1 to 10.7) to provide a concise summary of the original recommendations, and of the discussion contributed by these guidelines, by topic.

Box 10.1.
Theoretical Definition of the Welfare Indicator

	Original recommendation	Discussion
1. Money Metric Utility (MMU) vs. Welfare Ratio (WR)	Attempts should be made to use MMU and to calculate the Paasche price indices with individual household weights. (page 21)	Among monetary measures of welfare, MMU is still to be preferred to the WR. Its theoretical foundations, that is, standard consumer theory, have remained substantially unchanged. When the computation of a Paasche price index is empirically arduous (because of a lack of suitable information or because of low-quality data), then the use of other price indices, such as Laspeyres, should be considered. The MMU vs. WR debate, as it was originally framed in the *Guidelines*, has lost relevance in comparison to other, broader discussions on the nature of welfare measurement. Alternative approaches (social exclusion approach, multidimensional and subjective poverty) have gained ground and represent a useful complement, rather than replacement, of a monetary approach to welfare measurement. (section 2)
2. Income vs. consumption	In most developing countries where the Living Standards Measurement Study (LSMS) and/or household expenditure surveys are available, consumption is the appropriate measure to use. (page 21)	The debate on income vs. consumption is far from settled. On the one hand, income has gained legitimacy, especially in contexts were policy targets "minimum rights" to resources, and/or where inequality is a serious concern. On the other hand, consumption is a good representation of material deprivation, and is more reliably measured in low- and middle-income countries. Ultimately, it seems unlikely for a single measure to come out on top: consumption continues to play a fundamental role, but income-based measures should be considered as a complement or even an alternative, particularly as a country reaches more advanced stages of economic development. A challenge for the coming years is the implementation of a joint analysis of household consumption, income, and *wealth*: assets and lack thereof are key for measuring material well-being. OECD (2013) provides a useful and operational framework to get started. (section 3)

Box 10.2.
The food component of the nominal consumption aggregate

	Original recommendation	Discussion
3. Comprehensiveness of the food aggregate	Include food from all possible sources. In particular, the aggregate should include not just (i) food purchased in the market place, including meals purchased away from home for consumption at or away from home, but also (ii) food that is home-produced, (iii) food items received as gifts or remittances from other households, as well as (iv) food received from employers as payment in-kind for services rendered. (page 25)	This recommendation stands. (section 4.2)
4. Computation of the value of food purchases	Include food purchased from the market as the amount spent in the typical month × 12 (or number of months typically consumed). (page 38)	The recommendation is unchanged. (section 4.2)
5. Food consumption vs. acquisition	If food consumed can be distinguished from food purchased, include the value of the former. (page 27)	The recommendation remains valid, although it is common in practice to only have information on food purchased. In such cases, it is good practice to check for extreme values in the distribution of both food expenditures and calorie availability, and, if necessary, exclude large bulk purchases. (section 4.2.1)
6. Recall period for the food aggregate	When information on food purchases has been collected for more than one recall period, (...), the choice is limited between a "last two weeks" (or shorter period) measure, and a "usual month" measure. The literature reviewed in Deaton and Grosh (2000) leads to a recommendation in favor of the latter over the former. (page 28)	The "usual month" approach has not found much support in the literature since the early 2000s. Where the analyst has a choice, estimates obtained via simple recall (one week to a month) are to be preferred to "usual month" estimates. (section 4.2.2)
7. Food away from home	Include the value of meals consumed outside the home as the total of: amount spent in restaurants, amount spent on prepared foods, amount spent on meals at work, amount spent on meals at school, amount spent on meals on vacation. (page 40)	New methodological research has pointed to the inadequacy of standard questionnaire designs in capturing food away from home (FAFH). The analyst should still follow the original recommendation. (section 4.2.3)

Box 10.2. (Cont)

The food component of the nominal consumption aggregate

	Original recommendation	Discussion
8. Home-produced food	Food that is home produced: quantity in typical month × farmgate price × number of months typically consumed. (page 38)	The recommendation is unchanged, with the exception of the focus on a "typical month" measure. (section 4.2.4)
9. Farm-gate vs. market prices	Treat the farm household as a business selling to the household. Attempt to value produce at "farmgate" rather than "market" prices. (page 22)	The recommendation stands. It is acknowledged that actual farm-gate prices are virtually never available, and that in practice, welfare analysts use either self-reported valuations (if available from the survey) or unit values from food purchases. The former option is, in principle, to be preferred.(section 4.2.3)
10. Food received as gift or in-kind payment	Include total for a year. (page 38)	The recommendation stands. (section 4.2.4)
11. Missing prices or unit values	first choice is price (unit value) reported by the household; if not available, use as a proxy the median—not mean—price paid by "similar" households in the neighborhood, subject to checks that such prices are plausible. Check data for outliers; miscoding or misunderstanding of units for quantities causes errors in unit values. (page 38)	The recommendation stands. (section 4.2.4)
12. Food rations	*[DZ do not discuss food rations explicitly.]*	Food rations are quotas of food provided for free or at below-market price. Their "official" value cannot be included in the consumption aggregate: rations must be re-priced using an estimated market price, one that adequately represents the marginal utility of the ration for the consumer. Among the several options to do so, we recommend to exploit the information generated by a secondary market for rations, if it exists, and compute primary sampling unit (PSU) median ration unit values. (section 4.2.5)

Box 10.3.

The nonfood nondurable component of the nominal consumption aggregate

	Original recommendation	Discussion
13. Comprehensiveness of the nonfood aggregate	Daily use items, annualize the value. Clothing and housewares, annualize the value. Exclude taxes paid, purchase of assets, repayment of loans, expenditure on durable goods and housing, as well as other lumpy expenditures such as marriages and dowries. To the extent that local property taxes bear a relation to services rendered, we recommend their inclusion. (page 38)	These recommendations stand. They follow from selecting consumption expenditure as the welfare measure. Transfers to other households (such as gifts and remittances given) and charitable contributions are also to be excluded. Section 4.3 of this document provides a point-by-point summary of recommended inclusion and exclusion choices for each item category, based on the Classification of Individual Consumption According to Purpose (COICOP) scheme.
14. Subsidies	Expenditures on utilities, water, gas, electricity, or telephone can be problematic if some households are subsidized and some are not. (...) In some cases, making accurate regional (and certainly international) welfare comparisons will make it necessary to make corrections to (by repricing) the reported expenditures. (page 32)	The recommendation stands. Hentschel and Lanjouw (2000) provide a useful conceptual framework and practical recommendations for repricing subsidized goods. (section 4.3)
15. Regrettable necessities	Include expenditure on items that may or may not be regrettable necessities. (page 22)	The recommendation stands. (section 4.3)
16. Work-related expenditures	To the extent possible, purely work-related expenditures should be excluded. This recommendation does not include transport to work or work clothing. (page 38)	The recommendation stands. (section 4.3)
17. Health expenditures	Health expenses should only be included if they have high income elasticity in relation to their transitory variance or measurement error. (page 38)	The inclusion of health expenditure in the consumption aggregate continues to be a contentious issue. In contrast to the original recommendation, we argue that health expenditures should be included in the aggregate. The main argument for exclusion (health expenditures are associated with a loss in welfare in the health dimension) has little merit within the framework of monetary welfare measurement. Exclusion of selected health expenditures is still justified when they are considered to be atypical and "lumpy", but this need not be the case. (section 4.3.1)
18. Education expenditures	Education expenses: Typically measured quite accurately in most surveys—our recommendation is to include them. (page 38)	The recommendation stands. (section 4.3.1)
19. Leisure	Omit time and leisure in the calculation of consumption. (page 21)	The recommendation stands. (section 4.3.2)
20. Public goods	Do not include any valuation of public goods in the calculation of the household consumption aggregate. (page 22)	The recommendation stands. (section 4.3.2)

Box 10.4.
Durable goods and housing

	Original recommendation	Discussion
21. Purchase price of durables	A measure of use-value, not purchase, of durable goods is the right measure to include in the consumption aggregate from a welfare point of view. Exclude expenditures—instead, calculate a rental equivalent/user cost for housing and durable goods owned by the household. (page 21)	This recommendation is unquestioned. (section 4.4)
22. Consumption flow from durables	Calculate an annual rental equivalent using an appropriate real rate of interest and median depreciation values for each item calculated across all households owning that item. (page 38) *[Note that DZ are referring here to what the present paper calls the user-cost method]*	This recommendation remains valid. Estimate the consumption flow of consumer durables based on the user-cost method, and estimate the depreciation parameter using the geometric model. If the information required by the geometric model is not available, use the economic life depreciation model. If all else fails, consider excluding durable goods from the aggregate. (section 4.4)
23. Rent expenditures	If a household pays rent, annualize the amount of rent paid. Even if the dwelling is owned by the household or received free of charge, an estimate of the annual rental equivalent must be included in the consumption aggregate. In countries where few households pay rent, rental equivalents are potentially inaccurate, and the benefits of completeness need to be weighed against the costs of error. (page 38)	Including both actual rent and a measure of imputed rent for owners and nonmarket tenants remains the goal. Self-reported imputed rents may be used, if deemed accurate; hedonic regression models offer a viable alternative (Duan's retransformation should be used for predicted values from log-linear models); the user-cost and rent-to-value approaches may be useful if the first two methods fail. If no reliable estimate of rental expenditures can be produced, consider excluding rent (both actual and imputed) from the aggregate. (section 4.5)

Box 10.5.
Adjusting for price variation, household size and composition

	Original recommendation	Discussion
24. Choice of the price index	Use price indexes to adjust nominal consumption. Use within-survey prices supplemented by prices from the price questionnaire, if available. The Paasche index is our preferred price index to use to adjust for cost-of-living differences faced by different households. (page 52)	The core recommendation, rooted in consumer theory, stands the test of time. We document that the challenge is not much on the choice of the index, but rather on the empirical challenges to the calculation of a Paasche index. We register an increasing dissatisfaction with the use of unit values, and even more so with indirect methods (most notably, Engel curves–based methods). In the absence of other reliable price data sources, the recommended strategy is to explore the extent to which quality bias is a concern, for instance by summarizing median unit values by per capita expenditure decile (the presence of a steep gradient would be taken as evidence of a non-negligible quality bias), and consider empirical adjustments, as appropriate in the context at hand. We advise to also adjust for within-survey inflation, being aware that the sequence of the adjustments (temporal first, spatial after, or vice versa) leads to different results. (section 5.2)
25. Plutocratic vs. democratic aggregation of price indices	Plutocratic vs. democratic aggregation of price indices: the latter is preferable. (page 43)	This recommendation remains valid.
26. Per capita vs. per adult equivalent	Adjust household expenditure to reflect household size. Use an appropriate measure of size/composition. Continue using PCE supplemented with measures based on the arbitrary approach. Use low alpha and high theta in poor countries, and the reverse in richer countries. (pages 22, 52)	The recommendation to use an appropriate measure of size and composition to adjust total household expenditure remains valid. The use of an equivalence scale based on the "arbitrary" approach also remains the recommended route. The definition of coefficients for the scale is based on convention: computing coefficients on the basis of caloric energy requirements (World Health Organization/Food and Agriculture Organization [WHO/FAO] scales) has become a popular approach in many countries, but it is not superior to alternatives (if anything, it is more disputable, given that it is hinges solely on food consumption). DZ's recommended specification, the OECD-II scale, or the LIS (square-root) scale, would all be better choices. Regardless of the choice of equivalence scale, it is recommended to keep computing per capita expenditure as a supplementary measure, to facilitate comparisons both over time, especially in the long run, and internationally, as part of the monitoring of global poverty. (section 6)

Box 10.6.
Data issues

	Original recommendation	Discussion
27. Unit nonresponse	*[DZ do not include a stand-alone discussion on unit nonresponse.]*	Unit nonresponse poses a growing threat to the reliability of survey weights. The best way to mitigate this issue is to prevent it by maximizing compliance ex ante, that is at the survey implementation stage. Although welfare analysts are not directly tasked with treating unit nonresponse, in the case where ex post adjustments become necessary, the involvement of a sampling specialist is advised. It is recommended that the documentation accompanying final estimates explicitly address unit nonresponse and discuss how expansion factors (weights) were handled. (section 7.1)
28. Item nonresponse	*[DZ do not include a stand-alone discussion on item nonresponse.]*	The analyst should assess the extent to which item nonresponse affects the consumption aggregate through its elementary components. If the incidence of missing data is a concern, the nature of missingness should be investigated. If data are missing at random (MCAR and MAR), a number of approaches are available to mitigate the impact of missing values on the statistics of interest, though the variety of situations faced in practice is too wide to recommend any one-size-fits-all method. If evidence suggests that data are missing not at random (MNAR), this can threaten the representativeness of certain survey estimates. Ad hoc adjustments can be evaluated in consultation with sampling specialists. In both cases, random and nonrandom item nonresponse, the recommendation is to report how any corrections were handled in the documentation accompanying the final estimates (section 7.2)
29. Outliers	It is of the greatest importance that the analyst check each item for the presence of "gross" outliers, typically by graphing the data, for example using the "oneway" and "box" options in Stata. For inherently positive quantities, it is often useful to do this in logs as well as in levels. (page 25)	Extreme values represent a potential threat to the unbiasedness of consumption statistics, poverty, and inequality estimates. Therefore, DZ's original recommendation stands: it is essential to check the variable(s) of interest and assess the incidence of outliers before producing final estimates. Three aspects should take priority when it comes to outlier detection and treatment: (1) Conduct sensitivity analysis, e.g., by comparing results obtained for key indicators with and without the inclusion of outliers. (2) Document how any outlier corrections were handled, to ensure the replicability of the final aggregate and allow comparisons with the original data. (3) When estimating trends, implement the same outlier detection and treatment routines across surveys, if possible. (section 7.3)

Box 10.6.
Sensitivity analysis and reproducibility

	Original recommendation	Discussion
30. Sensitivity analysis	(Sensitivity) techniques should not be used to check out the results of every controversial decision in constructing the consumption aggregates. (...) But there are often critical decisions, of which that about equivalence scales is one, and the inclusion of a noisy item of expenditure is often another, where we know in advance that the decision is going to matter for the poverty analysis, and where it is important to have more information on exactly how it matters. For this, stochastic dominance analysis is ideally suited. (page 63)	The recommendation stands, and is even reinforced by a need to communicate results effectively and transparently to an audience that is often critical. A section or appendix dedicated to systematic sensitivity testing should become the norm for any technical report presenting inequality and poverty estimates. Tables with *side-by-side comparisons* of results obtained under different scenarios, as well as the *robustness matrix*, help perform sensitivity analysis in a systematic and organized way (section 8.1). When the focus is on poverty analysis, stochastic dominance stands out as the main framework for performing sensitivity checks. In particular, the *headcount difference curve* is a simple yet powerful tool. (section 8.2)
31. Reproducibility of the analysis	*[DZ do not include a stand-alone discussion on documentation and reproducibility of welfare analysis.]*	Build the entire process of welfare analysis (from data preparation, to the construction of the welfare aggregate, up to the computation of final estimates) in a way that ensures *reproducibility* by external researchers. This is a requirement of rigorous, scientifically sound analysis. It also facilitates the construction of comparable estimates once new survey data are released over time and spurs greater transparency and accountability. Three practical rules: (1) impose a clear and logical structure to the workflow of data analysis, (2) automate operations as much as possible using scripts (such as Stata do-files), and (3) fully document all data and scripts in a standard structured working directory. (section 9)

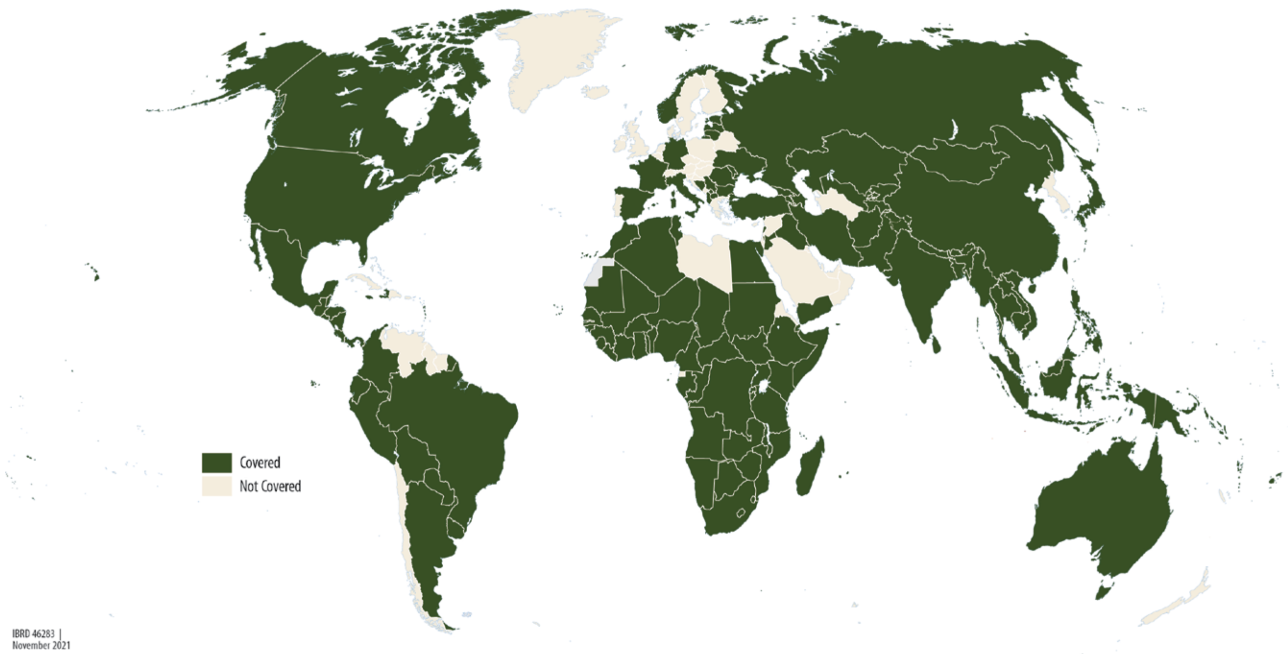
Appendix A.

Welfare Measurement Methodology Database

The database describes the main methodological choices made when constructing the welfare measure underlying official national poverty estimates. The source of the information is the official documentation released by the institutions producing the estimates (National Statistical Offices, the World Bank, etc.).

Map A.1 and table A.1 provide an overview of the coverage and sources of the database.

MAP A.1. Geographic coverage of the database



SOURCE: Authors' elaboration.

TABLE A.1. References underlying the database

Country	Year	Survey	Reference
Afghanistan	2016	Living Conditions Survey 2016–17	Central Statistics Organization (2018), Afghanistan Living Conditions Survey 2016–17. Kabul, CSO
Albania	2012	Living Standards Measurement Survey 2012	World Bank, Albania Institute of Statistics (INSTAT) (2013) Albania: trends in poverty 2002–2005—2008–2012 Albania Institute of Statistics (INSTAT) website. Living Standard Measurement Survey (LSMS) 2012, methodology section
Algeria	2011	l'Enquête Nationale sur les Dépenses de Consommation et le Niveau de Vie des Ménages Algériens (ENCNVM)	Office National des Statistiques (2011). Premiers résultats de l'Enquête Nationale sur les Dépenses de Consommation et le Niveau de Vie des Ménages 2011Gouvernement Algérien (2016). Objectifs du Millénaire pour le Développement. Rapport National 2000–2015
American Samoa	2013	2013/14 Household Income and Expenditure Survey	Samoa National Statistics Office/UNDP Pacific Centre (2016) Samoa, Hardship and Poverty Report, Analysis of the 2013/14 Household Income and Expenditure Survey
Angola	2008	Inquérito Integrado Sobre o Bem Estar da Populaçao (IDR II e MICS III) 2008–09	Instituto Nacional de Estatística (2013) Inquérito Integrado sobre o Bem-Estar da População, IBEP 2013, Relatório AnalíticoVol. III Perfil da Pobreza
Argentina	2016	Encuesta Permanente de Hogares Continua 2016, Segundo Trimestre/semestre	Instituto Nacional de Estadística y CensosI.N.D.E.C. (2016) La medición de la pobreza y la indigencia en la Argentina, Metodología INDEC Nº 22 Instituto Nacional de Estadística y CensosI.N.D.E.C. (2017) Informes Técnicos vol. 1 nº 53, Condiciones de vida vol. 1 nº 4, Incidencia de la pobreza y la indigencia en 31 aglomerados urbanos. Segundo semestre de 2016
Armenia	2015	Integrated Living Conditions Survey 2015	National Statistical Service of the Republic of Armenia, World Bank (2016) Social Snapshot and Poverty in the Republic of Armenia, Statistical Analytical Report. Based on the Results of the 2015 Integrated Living Conditions Survey of Households
Australia	2015	Household Expenditure Survey (HES) 2015–2016	Australian bureau of statistics (2017). Household Expenditure Survey, Australia: Summary of Results, 2015-16
Azerbaijan	2001	Azerbaijan-HBS 2001	World Bank (2003). Azerbaijan Republic Poverty Assessment (In Two Volumes) Volume I: Summary and ConclusionsWorld Bank (2003). Azerbaijan Republic Poverty Assessment (In Two Volumes) Volume II: The Main Report
Bangladesh	2016	Household Income and Expenditure Survey 2016–2017	Bangladesh Bureau of Statistics, Statistics and Informatics Division, Ministry of Planning (2017). Preliminary Report on Household Income and Expenditure Survey 2016 World Bank, Bangladesh Bureau of Statistics (2017). Description of the Official Methodology Used for Poverty Estimation in Bangladesh for 2016/17
Belize	2008	Household Expenditure Survey (HES) 2008–2009	National Human Development Advisory Committee. Ministry of Economic Development, Commerce and Industry,and Consumer Protection (2010). Belize Country Poverty Assessment, Volume 1. Main Report
Benin	2011	Enquête Modulaire Intégrée sur les Conditions de Vie des ménages 2011	Republique du Benin, Ministere du developpement, de l'analyse economique et de la prospective (MDAEP) Institut National de la Statistique et de l'Analyse Economique (INSAE) (2012) Enquete modulaire integree sur les conditions de vie des menages, 2ème édition (EMICoV 2011) Institut National de la Statistique et de l'Analyse Economique (INSAE) (2012) Evaluation de la pauvreté au Benin
Bhutan	2017	Living Standards Survey 2017	World Bank, National Statistics Bureau of Bhutan (2017) Bhutan Poverty Analysis report 2017 World Bank, National Statistics Bureau of Bhutan (2017) Bhutan living standards survey report 2017
Bolivia	2015	Encuesta de Hogares 2015	Instituto Nacional de Estadística (2018) Encuesta de hogares 2011–2015

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Bosnia and Herzegovina	2004	Living Standards Measurement Survey (LSMS) 2004 (Wave 4 Panel)	Department for International Development. Bosnia and Herzegovina, labour and social policy in Bosnia and Herzegovina: the development of policies and measures for social mitigation (2005) Living in BiH, Panel Study WAVE 4 Report. Draft for discussion World Bank (2003). Bosnia and Herzegovina Poverty Assessment. Volume II: Data on Poverty. Washington, DC: World Bank
Botswana	2009	Core Welfare Indicators Survey 2009–2010, Poverty Survey	World Bank (2015). Botswana Poverty Assessment. Report No. 88473-BW Statistics Botswana (2013). Botswana Core Welfare Indicators survey 2009/10. Main Report, volume 1
Brazil	2017	Pesquisa Nacional por Amostra de Domicílios ContínuaPNAD Contínua 2017	Instituto Brasileiro de Geografia e Estatística IBGE (2017). Síntese de indicadores sociais: uma análise das condições de vida da população brasileira: 2017
Bulgaria	2016	EU-SILC 2016	National Statistical Institute (n.d.) Statistics on Living Conditions Department, Demographic and Social Statistics Directorate. National Reference Metadata in ESS Standard for Quality Reports Structure
Burkina Faso	2014	Enquête Multisectorielle Continue 2013–2014	Institut national de la statistique et de la démographie (INSD) (2015). Rapport Enquête multisectorielle continue (EMC) 2014. Profil de pauvreté et d'inégalités
Burundi	2013	Enquête sur les conditions de vie des ménages 2013	Institut de Statistiques et d'Études Économiques du Burundi, Groupe de la Banque africaine de développement (2015). Burundi : profil et déterminants de la pauvreté. Rapport de l'enquête modulaire sur les conditions de vie des ménages 2013/2014
Cabo Verde	2015	inquérito às despesas e receitas familiares 2015	Instituto Nacional de Estatística (2018). Documento metodológico IDRF 2015; Medodologia da medição da pobreza monetária absoluta em Cabo Verde; Instituto Nacional de Estatística (2018). Documento metodológico IDRF 2015; Evolução da Pobreza Monetária Absoluta 2001/02, 2007 e 2015 Instituto Nacional de Estatística (2018). Perfil da Pobreza Absoluta em Cabo Verde; Evolução da Pobreza Monetária Absoluta 2001/02, 2007 e 2015
Cambodia	2011	Cambodia Socio-Economic Survey 2011	World Bank (2014). Where Have All The Poor Gone? Cambodia Poverty Assessment 2013 Kingdom of Cambodia (n.d.) National Strategic Development Plan, 2014–2018
Cameroon	2014	Quatrième Enquête Camerounaise auprès des Ménages (ECAM4) 2014	Republique du Cameroun, Institut National de la Statistique (n.d.) Troisieme Enquete Camerounaise Aupres des Menages (ECAM4)
Canada	2017	Canadian Income Survey 2017	Statistics Canada website. Detailed description of the Canada Income Survey 2017 (CIS) Employment and Social Development Canada (2018) Opportunity for All—Canada's First Poverty Reduction Strategy
Central African Republic	2008	Enquete Centrafricaine Pour Le Suivi-evaluation Du Bien-etre (ECASEB)	Backiny-Yetna, P., and Wodon, Q. (2010). Profil et corrélats de la pauvreté en République Centrafricaine en 2008. Perspective Afrique, Association Africaine pour les Sciences sociales, vol. 5(1–3), pages 1–28. World Bank (2019). Central African Republic: Priorities for Ending Poverty and Boosting Shared Prosperity. Systematic Country Diagnostic. World Bank, Washington, DC
Chad	2011	Enquête sur la Consommation des Ménages et le Secteur Informel au Tchad 2011, Troisième	République du Tchad, institut national de la statistique, des etudes economiques et demographiques (INSEED) (2013). Troisième Enquête sur la Consommation et le Secteur Informel au Tchad (ECOSIT3) Rapport FinalWorld Bank (2013) ChadPoverty note: dynamics of poverty and inequality following the rise of the oil sector
China	2016	National Household Survey on Income, Consumption and Living Conditions; National Rural Poverty Monitoring Survey	National Bureau of Statistics of China (2016). Poverty Monitoring Report of Rural China Lugo, M. A., Niu, C., and Yemtsov, R. (2021). Rural Poverty Reduction and Economic Transformation in China: A Decomposition Approach. Policy Research Working Paper no. 9849. Washington, DC: World Bank

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Colombia	2017	Gran Encuesta Integrada de HogaresGEIH—2017	Departamento Administrativo Nacional de Estadística (DANE) (2016) Metodología General Gran Encuesta Integrada de Hogares GEIHDepartamento Administrativo Nacional de Estadística (DANE) (2018) Boletín técnico. Pobreza Monetaria y Multidimensional en Colombia, Año 2017
Comoros	2014	Enquête sur les Dépenses de Consommation des Ménages aux Comores 2014	World Bank (2017). Comoros Poverty Assessment. World Bank, Washington, DC Union des Comores, Institut National de la Statistique, des Etudes Economiques et Demographiques, Fonds Africain de Développement (2015) Pauvrete et Consommation des menages en Union des Comores
Congo, Dem. Rep.	2012	Enquête 1-2-3 2012	Adoho, Franck M.. 2016. Congo, Democratic Republic of Poverty assessment (French). Washington, D.C. : World Bank Group
Congo, Rep.	2011	Enquête Congolaise Auprès des Ménages pour le Suivi et l'Evaluation de la Pauvreté 2011	World Bank (2017). Republic of Congo Poverty Assessment Report. Education, Jobs and Social Protection for a Sustainable Reduction of Poverty. Washington, DC: World Bank
Costa Rica	2017	Encuesta Nacional de Hogares 2017	Instituto Nacional de Estadística y Censos (INEC) (2017). Encuesta Nacional de Hogares Julio 2017, Resultados generales
Côte d'Ivoire	2015	Enquête sur le Niveau de Vie des Ménages (ENV) 2015	République de Côte d'Ivoire, Institut National de la Statistique (2015). Enquête sur le niveau de vies des ménages (ENV 2015) Profil de pauvreté
Croatia	1998	Household Budget Survey	World Bank (2001). Croatia, Economic Vulnerability and Welfare Study. Washington, DC: World Bank
Djibouti	2017	Enquête Djiboutienne Auprès des Ménages pour les Indicateurs Sociaux 2017	Republique de Djibouti, Direction de la Statistique et des Etudes Demographiques (DISED) (2018). Resultats de la quatrieme enquête djiboutienne aupres des menages pour les indicateurs sociaux (EDAM4-IS)
Dominica	2007	Survey of Living Conditions and Household Expenditure and Income 2007–2008	Kairi Consultants Limited, Caribbean Development Bank, National assessment team of Dominica (2010). Country poverty assessment Dominica. Volume Main report; Kairi Consultants Limited, Caribbean Development Bank, National assessment team of Dominica (2010). Country poverty assessment Dominica. Volume Participatory Poverty Assessment
Ecuador	2013	Encuesta Condiciones de Vida—Sexta Ronda (ECV) 2013–2014	Instituto Nacional de Estadística y Censos (INEC) (n.d.) Informe de Resultados, ECV 2013–2014 Instituto Nacional de Estadística y Censos (INEC) (2015) Nota Técnica-INEC-001. Metodología de construcción del agregado del consumo y estimación de línea de pobreza en el Ecuador
Egypt, Arab Rep.	2015	Income,Expenditure and Consumption Survey	World Bank (2021). Unlocking Egypt's Potential for Poverty Reduction and Inclusive Growth : Egypt Systematic Country Diagnostic Update. World Bank, Washington, DCWorld Bank Poverty and Equity Brief, Arab Republic of Egypt. Retrieved from the Web in April 2021
El Salvador	2017	Encuesta de Hogares de Propósitos Múltiples 2017	Dirección General de Estadística y Censos DIGESTYC (2018). Encuesta de Hogares de Propósitos Múltiples 2017
Estonia	2000	Household Budget Survey 2000	Statistical Office of the Republic of Estonia (n.d.) Estonia 2000: Survey Information Ministry of Social Affairs (2009) Health, Labour and Social Life in Estonia 2000–2008
Eswatini	2010	2009/10 Swaziland Household Income And Expenditure Survey	Central Statistical Office (2010). 2009/10 Swaziland Household Income And Expenditure Survey. Poverty in a decade of slow economic growth: Swaziland in the 2000's
Ethiopia	2015	2015/16 Ethiopian household consumption expenditure (HCE) survey	Federal Democratic Republic of Ethiopia (2017). Ethiopia's Progress Towards Eradicating Poverty: An Interim Report on 2015/16 Poverty Analysis Study Federal Democratic Republic of Ethiopia, Central Statistical Agency (2018) The 2015/16 Ethiopian Household Consumption—Expenditure (HCE) Survey. Statistical Report
Fiji	2013	Household Income and Expenditure Survey 2013	Fiji Bureau of Statistics (2015). Report on the 2013–14 Household income and expenditure survey

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
France	2015	Enquete SRCV-SILC 2015	INSEE (n.d.) National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRSSI), 2015INSEE (2018) Tableaux de l'économie française
Gabon	2005	Enquête Gabonaise pour l'Evaluation et le Suivi de la Pauvreté (EGEP) 2005	République Gabonaise (2015). Rapport National OMD (Version Provisoire) World Bank (2006). Gabon, Diagnostique de la pauvreté
Gambia, The	2015	Integrated Household Survey 2015	The Government of The Gambia, Gambia Bureau of Statistics (2017). Integrated Household Survey 2015/16. Volume III: Prevalence and Depth of Poverty Gambia Bureau of Statistics (2017). Integrated Household Survey 2015/16. Volume III: Prevalence and Depth of Poverty
Georgia	2017	Household Income and Expenditure Survey (HIES)	World Bank (2019). Poverty & Equity Brief; Europe & Central Asia; Georgia
Germany	2018	EU-SILC (Leben in Europa)	German Institute of Statistics (DESTATIS) website. Section on Living conditions, risk of poverty
Ghana	2017	Living Standards Survey (GLSS 7) 2017	Ghana Statistical Service (2018) Ghana Living Standards Survey Round 7 (GLSS 7). Poverty Trends in Ghana 2005–2017
Guatemala	2014	Encuesta Nacional de Condiciones de Vida (ENCOVI) 2014	Instituto Nacional de Estadística (2015) República de Guatemala: Encuesta Nacional de Condiciones de Vida 2014. Principales resultadosInstituto Nacional de Estadística (INE) (2016) Encuesta Nacional de Condiciones de Vida 2014. Tomo I Instituto Nacional de Estadística (INE) (2016) Resultados de la encuesta nacional de condiciones de vida -ENCOVI-2014 con énfasis en datos de vivienda
Guinea	2012	Enquête Légère pour l'Evaluation de la Pauvreté (ELEP) 2012	Republique de Guinee, Ministere du Plan, Institut National de la Statistique (2012) Enquete Legere pour l'Evaluation de la Pauvrete ELEP-2012. Rapport Final
Guinea-Bissau	2010	Inquerito Ligeiro para a Avaliação da Pobreza 2010	República da Guiné-Bissau, Instituto Nacional de Estatística (2011) Inquérito Ligeiro para Avaliação da Pobreza (ILAP2), Resultados definitivos
Haiti	2012	Enquête sur les Conditions de Vie des Ménages Après le Séisme (ECVMAS)	Backiny-Yetna, P., Marzo, F. (n.d.) Poverty in Haiti: Methodological Note on the aggregate of consumption using ECVMAS 2012 World Bank (2014) Investing in people to fight poverty in Haiti. Reflections for evidence-based policy making
Honduras	2018	Encuesta Permanente de Hogares de Propósitos Múltiples (EPHPM) 2018	Instituto Nacional de Estadística (n.d.) Encuesta Permanente de Hogares de Propósitos Múltiples 2018. Resumen Ejecutivo Instituto Nacional de Estadística (n.d.) Encuesta Permanente de Hogares de Propósitos Múltiples 2018. Metodología
India	2011	Household Consumer Expenditure in India, NSS 68th Round July 2011–June 2012	Government of India Planning Commission (2013) Press Note on Poverty Estimates, 2011–12 Government of India Planning Commission (2009) Report of the Expert Group to review the methodology for estimation of poverty Government of India, Ministry of Statistics and Programme Implementation (2016) SARVEKSHANA Journal of National Sample Survey Office, 100th issue Government of India, Ministry of Statistics and Programme Implementation (2013) Key Indicators of Household Consumer Expenditure in India Nss 68' Round (July 2011–June 2012) World Bank (2018) International Comparisons of Poverty in South Asia; Poverty and equity global practice working paper 184
Indonesia	2016	National Social Economic Survey (SUSENAS) 2016	Badan Pusat Statistik-Statistics Indonesia (2016) Welfare Statistics 2016. Education Inequality in Indonesia Badan Pusat Statistik-Statistics Indonesia (2016) Welfare Indicators 2016. Education Inequality in Indonesia Badan Pusat Statistik (2016) Penghitungan dan Analisis Kemiskinan Makro Indonesia
Iran, Islamic Rep.	2017	Household Income and Expenditure Survey	World Bank Poverty and Equity Brief, Islamic Republic of Iran. Retrieved from the Web in October 2020

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Iraq	2012	Household Socio-Economic Survey 2012, Second Round	Government of Iraq and Kurdistan Region, World Bank (n.d.) Poverty in Iraq 2012–2014 World Bank (2014a). Poverty Estimates and Trends in Iraq: 2007–2012. Methodological Note. Washington, DC: World Bank World Bank (2014b). Iraq: The Unfulfilled Promise of Oil and Growth Poverty, Inclusion and Welfare in Iraq, 2007–2012. (In Four Volumes) Volume 1: Main Report. Washington, DC: World Bank
Italy	2017	Indagine sulle spese delle famiglie	Istituto nazionale di statistica ISTAT (2016). Spese per Consumi delle Famiglie, Anno 2015 Istituto nazionale di statistica ISTAT (2016). Il Benessere Equo e Sostenibile in Italia (BES) 2016
Jamaica	2012	Jamaica Survey of Living Conditions (JSLC) 2012	Planning Institute of Jamaica (n.d.) 2012 Jamaica Survey of Living Conditions (JSLC) Executive Summary
Japan	2016	Comprehensive Survey of Living Conditions 2016	Ministry of Health, Labour and Welfare, Household Statistics Office (2017) Summary Report of Comprehensive Survey of Living Conditions 2016
Jordan	2010	Household Expenditures & Income Survey 2010	Department of Statistics, Jordan (2015). Poverty in Jordan. United Nations Economic Commission for Europe, Conference of European Statisticians, Seminar on poverty measurement May 5–6, 2015, Geneva, Switzerland The Hashemite Kingdom of Jordan, Department of Statistics (2012). Household Expenditures & Income Survey 2010
Kazakhstan	2002	Household Budget Survey (HBS) 2002	World Bank (2004). Dimensions of poverty in Kazakhstan, Volume 1. Policy Briefing World Bank (2004). Dimensions of poverty in Kazakhstan, Volume 2. Profile of Living Standards in Kazakhstan in 2002
Kenya	2015	Integrated Household Budget Survey 2015–2016	World Bank (2018). Kenya Gender and Poverty Assessment 2015/16. Reflecting on a Decade of Progress and the Road Ahead Kenya National Bureau of Statistics (2018). Basic Report on Well-Being in Kenya. Based on the 2015/16 Kenya integrated Household Budget survey (KiHBs)
Kiribati	2006	2006 Household Income and Expenditure Survey	Kiribati National Statistics Office, UNDP Pacific Centre (2010). Kiribati, Analysis of the 2006 Household Income and Expenditure Survey. A Report on the Estimation of Basic Needs Poverty Lines, and the Incidence and Characteristics of Poverty in Kiribati
Kosovo	2005	Household Budget Survey 2005	World Bank, Statistical Office of Kosovo (2007). Kosovo Poverty Assessment. Volume I: Accelerating Inclusive Growth to Reduce Widespread Poverty World Bank, Statistical Office of Kosovo (2007). Kosovo Poverty Assessment. Volume II: Estimating Trends from Non-comparable Data Kosovo Agency of Statistics (2016). Series Social Statistics. Results of Household Budget Survey 2015
Kyrgyz Republic	2013	Kyrgyz Integrated Household Survey (KIHS) 2013	World Bank (2015). Kyrgyz Republic: Poverty profile for 2013
Lao PDR	2012	Lao expenditure and consumption survey 2012/13	Lao Statistics Bureau, World Bank (2014). Poverty Profile in Lao PDR. Poverty Report for the Lao Consumption and Expenditure Survey 2012–2013
Latvia	2016	EU SILC 2016	Central Statistical Bureau of Latvia (n.d.) National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRSSI), 2016
Lebanon	2011	Household budget survey 2011/2012	Lebanon Central Administration for Statistics, World Bank (n.d.) Snapshot of Poverty and Labor Market outcomes in Lebanon Based on Household Budget Survey 2011/2021 Lebanon Central Administration for Statistics, World Bank (2015). Measuring Poverty in Lebanon using 2011 Household Budget Survey
Lesotho	2017	Continuous Multi-Purpose Household Survey (CMS/HBS) 2017/18	Sulla, Victor; Zikhali, Precious; Mahler, Daniel Gerszon (2019). Lesotho Poverty Assessment: Progress and Challenges in Reducing Poverty Bureau of Statistics (2019) Lesotho Poverty Trends and Profile Report 2002/2003 to 2017/2018

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Liberia	2016	Household Income and Expenditure Survey 2016	Liberia Institute of Statistics and Geo-Information Services (LISGIS) (2017). Household Income and Expenditure Survey 2016. Statistical Abstract
Lithuania	2016	EU SILC 2016	Statistics Lithuania (n.d.) National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRSSI), 2016
Madagascar	2010	Enquête Périodique auprès des Ménages 2010	World Bank (2016). Shifting fortunes and Enduring Poverty in Madagascar, Recent findings Institut National de la Statistique (2011). Enquête périodique auprès des ménages 2010, rapport principal
Malawi	2016	Integrated Household Survey-IV (HIS-IV)	World Bank, the National Statistics Office of Malawi (2018). methodology for poverty measurement in Malawi (2016/17) Republic of Malawi (2017). integrated household survey 2016–2017
Malaysia	2016	Household Income and Basic Amenities Survey 2016	Department of Statistics, Malaysia (2017). Household Income and Basic Amenities Survey Report 2016
Maldives	2016	Household Income & Expenditure Survey 2016	Maldives, National Bureau of Statistics (n.d.) Household Income and Expenditure Survey (HIES) Administrative Report 2016 Maldives, National Bureau of Statistics (n.d.) Household Income and Expenditure Survey (HIES). Analytical Report IV: Poverty & Inequality 2016 Maldives, National Bureau of Statistics (2018) Poverty measurement methodology in the Maldives Technical report Maldives, National Bureau of Statistics (n.d.) Household Income and Expenditure Survey (HIES). Technical Report 2016
Mali	2018	Enquete modulaire et permanente auprès des ménages (EMOP) 2018–2019	Institut national de la statistique du Mali (2019) Rapport annuel 2018 Consommation, pauvreté et bien-être des ménages
Mauritania	2014	Enquête Permanente sur les Conditions de Vie des ménages 2014	République Islamique de Mauritanie, Office National de la Statistique (2015) Profil de la Pauvreté en Mauritanie 2014 World Bank (2016) Islamic Republic of Mauritania, Poverty Dynamics and Social Mobility 2008–2014
Mauritius	2017	Household Budget Survey (HBS)	Statistics Mauritius (2018). Household Budget Survey 2017 and the Updated Price Index. Methodological report Statistics Mauritius (2018). Household Budget Survey 2017. Analytical Report
Mexico	2018	Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) 2018	Consejo Nacional de Evaluación de la Política de Desarrollo Social (2019). 10 años de medición de pobreza en méxico, avances y retos en política social Consejo Nacional de Evaluación de la Política de Desarrollo Social (n.d.) Canastas alimentarias y no alimentarias, observadas y normativas
Micronesia, Fed. Sts.	2013	2013–2014 Household Income and Expenditure Survey	Government of the Federated States of Micronesia, World Bank (2017). Poverty Profile of the Federated States of Micronesia. Based on the 2013–14 Household Income and Expenditure Household Income and Expenditure Survey; technical report
Moldova	2013	Household Budget Survey 2013	World Bank (2016). Moldova Poverty Assessment 2016. Poverty Reduction and Shared Prosperity in Moldova: Progress and Prospects National Bureau of Statistics of the Republic of Moldova (2014) Aspects of the Standard of Living of population in 2013 (Results of the Household Budget Survey)
Mongolia	2016	Household Socio-Economic Survey (HSES) 2016	National Statistics Office of Mongolia (2017). Poverty Profile 2016. Ulaanbaatar: Mongolia.

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Morocco	2013	Enquête nationale sur la Consommation et les Dépenses des Ménages (ENCDM) 2013–2014	Royaume du Maroc, Haut Commissariat au Plan (2016). Présentation des résultats de l'Enquête Nationale sur la Consommation et les Dépenses des ménages 2013/2014. Inégalités sociales et territoriales à la lumière des résultats de l'enquête nationale sur la consommation et les dépenses des ménages 2014 Royaume du Maroc, Haut Commissariat au Plan (2017). Pauvreté et Prospérité partagée au Maroc du troisième Millénaire, 2001–2014 Royaume du Maroc, Haut Commissariat au Plan (n.d.) Enquete Nationale sur la consommation et les dépenses des ménages 2013/2014. Rapport de synthèse
Mozambique	2014	Inquérito sobre Orcamento Familiar 2014–2015	World Bank (2018). Strong but Not Broadly Shared Growth. Mozambique Poverty Assessment Moçambique, Direcção de Estudos Económicos e Financeiros (2016). Pobreza e Bem-Estar em Moçambique: Quarta Avaliação Nacional (IOF 2014/15) Moçambique, Instituto Nacional de Estatística (2015). Relatório Final do Inquérito ao Orçamento Familiar—IOF 2014/15
Myanmar	2017	Myanmar Living Conditions Survey (MLCS) 2016/17	Government of Myanmar and The World Bank (2017). Technical poverty estimation report. Myanmar Poverty and Living Conditions Survey World Bank, UNDP (2019). Myanmar Living Conditions Survey 2017. Poverty report World Bank, UNDP (2019). Myanmar Living Conditions Survey 2017. Key indicators report
Namibia	2015	Namibia Household Income and Expenditure Survey, 2015/16	Namibia Statistics Agency (2017). Namibia Household Income and Expenditure Survey (NHIES) 2015/2016 Report Namibia Statistics Agency (2016). Namibia Household Income and Expenditure Survey (NHIES) 2015/2016 Key poverty Indicators (preliminary figures)
Nauru	2012	2012/13 Household Income and Expenditure Survey	Government of Nauru National Statistics Office, UNDP Pacific Centre (n.d.) Nauru Hardship and Poverty Report. Analysis of the 2012/13 Household Income and Expenditure Survey
Nepal	2010	Nepal Living Standards Survey (NLSS) 2010	Government of Nepal, Central Bureau of Statistics (2011). Nepal Living Standards Survey 2010/11. Statistical Report, Volume One Government of Nepal, Central Bureau of Statistics (2011). Nepal Living Standards Survey 2010/11. Statistical Report, Volume Two
Nicaragua	2014	Encuesta Nacional de Hogares sobre Medición de Nivel de Vida 2014	Instituto Nacional de Información de Desarrollo (INIDE) (2016). Encuesta Nacional de Hogares sobre Medición de Nivel de Vida 2014
Niger	2011	Enquête Nationale sur les Conditions de Vie des Ménages et l'Agriculture 2011	Institut National de la Statistique du Niger (2013) Profil et determinants de la Pauvrete au Niger en 2011. Premiers resultats de l'Enquete Nationale sur les Conditions de Vie des menages et l'Agriculture au niger (ECVMA); Institut National de la Statistique du Niger (2013) 2011 National Survey on Household Living Conditions and Agriculture (ECVM/A-2011). Basic Information Document;
Nigeria	2009	Harmonised Nigeria Living Standards Survey 2009/2010, First Round	Federal Republic of Nigeria, National Bureau of Statistics (2012). Nigeria Poverty Profile 2010 World Bank (2013). Nigeria Where Has All the Growth Gone? A Policy Note
North Macedonia	2013	Survey on income and living conditions, 2013	World Bank (2015). FYR of Macedonia: Measuring Welfare using the Survey of Income and Living Conditions (SILC)World Bank (2017). Poverty Reduction, Shared Prosperity and Inequality in FYR Macedonia in the Post Financial Crisis Period (2009–2013) Republic of Macedonia State Statistical Office (2015). Survey on Income and Living Conditions, 2013. Statistical review: Incomes, Expenditures and Prices
Norway	2015	Survey on Living Conditions EU-SILC 2015	Thorsen, L. R. (2018). Materielle og sosiale mangler i den norske befolkningen. Resultater fra Levekårsundersøkelsen EU-SILC

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Pakistan	2013	Household integrated economic survey (HIES) (2013–14)	Government of Pakistan, Minsitry of finance (n.d.) Pakistan Economic Survey 2015–16. Annex III: Poverty Government of Pakistan, Statistics Division (2015). Household Integrated Economic Survey (HIES) (2013–14)
Palau	2006	2006 Household Income and Expenditure Survey (HIES)	Palau Office of Planning and Statistics (2008). Palau, Analysis of the 2005/2006 Household Income and Expenditure Survey. Final report on the estimation of basic needs poverty lines, and the incidence and characteristics of poverty in Palau Republic of Palau (n.d.) 2006 Household Income and Expenditure Survey
Palestine	2016	Palestinian Expenditure and Consumption Survey (PECS)	Palestinian Central Bureau of Statistics and The World Bank (2018). Measuring Poverty in West Bank and Gaza: Methodology Review Using PECS 2016. Washington, DC: World Bank
Panama	2008	Encuesta de Niveles de Vida (ENV) 2008	Encuesta de Niveles de Vida (ENV) 2008, nota: el agregado de consumo Encuesta de Niveles de Vida (ENV) 2008, nota: medición de la pobreza World Bank (2011). Panama Poverty Assessment: Translating Growth into Opportunities and Poverty Reduction
Papua New Guinea	2009	Household Income and Expenditure Survey 2009–2010	Gibson, J. (2012). Technical Report, Papua New Guinea Poverty Profile. Based on the household income and expenditure survey Papua New Guinea, National Statistical Office (n.d.) 2009–2010 Papua New Guinea Household Income and Expenditure Survey. Summary Tables
Paraguay	2017	Encuestas Permanentes de Hogares (EPH) 2017	Paraguay Dirección General de Estadística, Encuestas y Censos (DGEEC) (2018). Principales Resultados de Pobreza y Distribución del Ingreso. EPH: Encuesta Permanente de Hogares 2017 Paraguay Dirección General de Estadística, Encuestas y Censos (DGEEC) (2018). "Condiciones de Vida," Encuestas Permanentes de Hogares (EPH)
Peru	2017	Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza 2017	Instituto Nacional de Estadística e Informática (2018). Evolución de la Pobreza Monetaria 2007–2017. Informe Técnico Instituto Nacional de Estadística e Informática (2018). Encuesta Nacional de Hogares 2017. Calidad de la Encuesta Instituto Nacional de Estadística e Informática (n.d.) Ficha Técnica. Encuesta Nacional de Hogares sobre Condiciones de Vida y Pobreza 2017. Unpublished
Philippines	2015	2015 Family Income and Expenditure Survey (FIES)	World Bank (2018). Making Growth Work for the Poor. A Poverty Assessment for the PhilippinesPhilippine Statistics Authority (2017). 2015 Family Income and Expenditure Survey (FIES)
Romania	2016	EU-SILC 2016	Romania National Institute of Statistics (2016). National Reference Metadata in ESS Standard for Quality Reports Structure (ESQRSSI)
Russian Federation	2015	ОБСЛЕДОВАНИЯ БЮДЖЕТОВ ДОМАШНИХ ХОЗЯЙСТВ 2015	RosStat (2016). СОЦИАЛЬНО-ЭКОНОМИЧЕСКИЕ ИНДИКАТОРЫ БЕДНОСТИ в 2012–2015 гг. (Socio Economic Indicators of Poverty in 2012–2015)
Rwanda	2016	Enquête Intégrale sur les Conditions de Vie des ménages (EICV-V)	National Institute of Statistics of Rwanda (NISR) (2018). The Fifth Integrated Household Living Conditions Survey, EICV5; Rwanda Poverty Profile Report 2016/17
Samoa	2013	2013/14 Household Income and Expenditure Survey	Samoa National Statistics Office, UNDP Pacific Centre (2016). Samoa Hardship and Poverty Report. Analysis of the 2013/14 Household Income and Expenditure Survey
São Tomé and Principe	2010	Inquérito Orçamento Familiar 2010	São Tomé et Príncipe, Ministère du Plan et du Développement (2011). Profil de pauvreté à São Tomé et Príncipe 2010 Instituto Nacional de Estatística (n.d.) Inquérito aos orçamentos familiares, Perfil de pobreza em São Tomé e Príncipe 2010
Senegal	2010	Enquête de suivi de la pauvreté au Sénégal (2010–2011)	République du Sénégal, Agence Nationale de la Statistique et de la Démographie (ANSD) (2013). Deuxième Enquête de Suivi de la Pauvreté au Sénegal (ESPS-II 2011). Rapport Définitif
Serbia	2007	SerbiaLiving Standards Measurement Survey 2007	Statistical Office of the Republic of Serbia (2008). Living Standards Measurements Study Serbia 2002–2007

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Seychelles	2013	Household Budget Survey 2013	National Bureau of Statistics Seychelles (2016). A Poverty Profile of the Republic of the Seychelles. Poverty Report for the Household Budget Survey 2013 Republic of the Seychelles (n.d.) Household Budget Survey 2013
Sierra Leone	2011	Integrated Household Survey 2011	World Bank (2013). A Poverty Profile for Sierra Leone
Solomon Islands	2012	2012/13 Household Income and Expenditure Survey	Solomon Islands National Statistics Office, World Bank (2015). Solomon Islands Poverty Profile based on the 2012/13 Household Income and Expenditure Survey
Somalia	2013	Somaliland Household Survey 2013	World Bank (2015). Somaliland Poverty Assessment
South Africa	2014	Living Conditions Survey 2014–2015	Statistics South Africa (2017). Poverty Trends in South Africa. An examination of Absolute Poverty between 2006 and 2015 World Bank (2018). Overcoming Poverty and Inequality in South Africa, An Assessment of Drivers, Constraints and Opportunities Statistics South Africa (2017). Living Conditions of Households in South Africa. An analysis of household expenditure and income data using the LCS 2014/2015
South Sudan	2015	High Frequency South Sudan Survey 2015	World Bank (2016). South Sudan Poverty Profile 2015. Findings from the 2015 wave of the High Frequency South Sudan Survey. World Bank.
Spain	2004	Encuesta de Condiciones de Vida 2004	Instituto Nacional de Estadística (n.d.) Estudio descriptivo de la pobreza en España Resultados basados en la Encuesta de Condiciones de Vida 2004 Instituto Nacional de Estadística (2005). Encuesta de Condiciones de Vida (ECV) Año 2004. Principales resultados
Sri Lanka	2016	Household Income and Expenditure Survey (HIES) 2016	Sri Lanka, Ministry of National Policies and Economic Affairs. Department of Census and Statistics (2018). Household Income and Expenditure Survey 2016. Final Report Department of Census and Statistics (2018). Household Income and Expenditure Survey 2016, Final ReportWorld Bank (2018). International Comparisons of Poverty in South Asia; Poverty and equity global practice working paper 184
St. Lucia	2016	Survey of Living Conditions and Household Budgets 2016	Kairi Consultants Limited (2018). Saint Lucia National Report of Living Conditions 2016 Final Report
Sudan	2009	National Baseline Household Survey (NBHS) 2009	World Bank (2011). A Poverty Profile for the Northern States of Sudan National Bureau of Statistics (NBS) (2012). National Baseline Household Survey 2009. Report for South SudanMartín Cumpa Castro (2010). Poverty in Northern Sudan, Estimates from the NBHS 2009Sudan Central Bureau of Statistics (2010). Sudan National Baseline Household Survey 2009 North SudanTabulation Report
Tajikistan	2014	HBS 2014	Tajikistan Agency of Statistics (2015). Poverty measurement in Tajikistan: A methodological note
Tanzania	2017	2017–18 Household Budget Survey	Ministry of Finance and Planning Poverty Eradication Division (MoFP-PED) [Tanzania Mainland] and National Bureau of Statistics (NBS) (2019). Tanzania Mainland Household Budget Survey 2017–18, Key Indicators Report World Bank (n.d.) Mainland Poverty Assessment, executive summary
Thailand	2013	Household Socio-Economic Survey (HSES) 2013	World Bank (n.d.) Getting Back on Track. Reviving growth and securing prosperity for all. Thailand Systematic Country Diagnostic National Statistical Office (2014). The 2013 Household Socio-Economic Survey. Whole Kingdom
Timor-Leste	2014	Timor-Leste Survey of Living Standards (TLSLS) 2014–2015	General Directorate of Statistics (2016). Poverty in Timor-Leste 2014
Togo	2015	Questionnaire des Indicateurs de Base du Bien-être (QUIBB) 2015	Institut National de la Statistique et des Etudes Economiques et Démographiques (INSEED) (2016). Togo: Profil de Pauvreté 2006-2011-2015

Continued

TABLE A.1. (Continued)

Country	Year	Survey	Reference
Tonga	2015	2015 Household Income and Expenditure Survey	Tonga Statistics Department, the Secretariat of the Pacific Community (2017). New CaledoniaHousehold Income and Expenditure Survey 2015/2016, Full report
Tunisia	2010	Enquête Nationale sur le Budget, la Consommation et le Niveau de Vie des Ménages 2010	République Tunisienne, Institut National de la Statistique (2012). Mesure de la pauvreté des inégalités et de la polarisation en Tunisie 2000–2010
Turkey	2018	Income and Living Conditions Survey, 2018	Income and Living Conditions Survey, 2018, Metadata. From the Turkish Statistical Institute website. Retrieved on November 2021
Tuvalu	2016	2016 Household Income and Expenditure Survey	Tuvalu Central Statistics Division, Pacific Community, New Caledonia (2018). Tuvalu 2015/2016 Household Income and Expenditure Survey, Full report
Uganda	2016	Uganda National Household Survey (NHS) 2016–2017	Uganda Bureau of Statistics (2018). Uganda National Household Survey 2016/2017 Report
Ukraine	2012	Household living conditions survey 2012	State Statistics Service of Ukraine (2013). Poverty Measurement in Ukraine: Criteria, Challenges and Prospects. Working paper 12 in United Nations Economic Commission for Europe
United States	2017	Current Population Survey Annual Social and Economic Supplements (CPS ASEC) 2017	Fontenot, K., Semega, J., and Kollar, M., U.S. Census Bureau (2018). Income and Poverty in the United States: 2017. In Current Population Reports P60-263
Uruguay	2017	Encuesta Continua de Hogares 2017	Instituto Nacional de Estadística Uruguay (2018). Estimación de la pobreza por el Método del Ingreso 2017 Instituto Nacional de Estadística Uruguay (2018). Ficha técnica Encuesta Continua de Hogares
Uzbekistan	2000	Household Budget Survey (HBS) 2000	World Bank (2003) Uzbekistan Living Standards Assessment. Policies to Improve Living Standards (In Two Volumes) Volume I: Summary Report; World Bank (2003) Uzbekistan Living Standards Assessment Policies To Improve Living Standards. (In Two Volumes) Volume II: Full Report; World Bank (2007) Republic of Uzbekistan, Living Standards Assessment Update
Vanuatu	2010	Vanuatu Household Income and Expenditure Survey 2010	Vanuatu National Statistics Office, UNDP Pacific Centre (2013). Vanuatu Hardship and Poverty Report. Analysis of the 2010 Household Income and Expenditure Survey Vanuatu National Statistics Office (2012). Vanuatu Household Income and Expenditure Survey 2010. Report
Vietnam	2016	Viet Nam Household Living Standards Survey 2016	World Bank (2018). Climbing the ladder. Vietnam poverty and shared prosperity update report. Poverty Reduction and shared Prosperity in Vietnam General Statistics Office (n.d.) Result of the Vietnam household living standards survey 2016
West Bank and Gaza	2017	Expenditure and Consumption Survey, PECS 2017	Palestinian Central Bureau of Statistics (n.d.) note: Poverty Profile in Palestine, 2017
Yemen, Rep.	2005	Household Budget Survey (HBS) 2005–2006	World Bank (2007a). Republic of Yemen Poverty assessment (Vol. 1): Main report. Washington, DC: World Bank World Bank (2007b). Yemen Towards Qat Demand Reduction. Washington, DC: World Bank
Zambia	2015	2015 Living Conditions Monitoring Survey	Republic of Zambia, Central Statistical Office (2016). 2015 Living Conditions Monitoring Survey Report
Zimbabwe	2017	Poverty Income Consumption and Expenditure Survey (PICES)	Zimbabwe National Statistics Agency (2018). Poverty, Income, Consumption and Expenditure Survey 2017 Report

NOTE: The table also contains survey reports cited throughout the text. Croatia and Palestine are not covered in the Poverty Measurement Methodology Database, but they appear in the table, as the related reports have been cited elsewhere in the text.
SOURCE: Authors' elaboration.

Appendix B.

Consumption vs. Income as Welfare Indicators

TABLE B.1. Advantages and disadvantages of income and consumption as welfare indicators

	Income	Consumption expenditure
ADVANTAGES	**Conceptual** 1. Reflects "potential living standards" [17] 2. Direct link with basic-income policies [11] 3. Facilitates microsimulation analysis [17]	**Conceptual** 1. Reflects "achieved living standards" [11] 2. Foundations in modern economic theory [1]
	Empirical 4. Income surveys are more regular than expenditure surveys [6] 5. Estimates with smaller standard errors due to larger sample size 6. Cost-effectiveness: in general income data are more cost effective to collect than expenditure data [11]	**Empirical** 3. Consumption smoothing allows to proxy of "typical" living standard/permanent income [17] 4. Captures the size of savings and access to credit [1] 5. More likely to capture private and public transfers [2] 6. Better reflects the insurance value of government programs [1] 7. Better measure of material well-being for those at the bottom of the distribution [3,7] 8. Collecting reliable information on consumption is much simpler than for income [4,5] 9. Disaggregation by consumption categories is informative about changes in material well-being [3]
DISADVANTAGES	**Conceptual** 1. Lack of theoretical foundations [1] 2. Fails to capture the role of savings and access to credit [1] 3. Does not reflect in-kind transfers [1] 4. Low-income not so well correlated with other non-income bad welfare outcomes [2]	**Conceptual** 1. Indirect policy levers [11] 2. Individual choice [11]
	Empirical 5. Sensitive to short-term fluctuations [11] 6. Subject to severe underreporting [2] 7. Unit and item nonresponse [21] 8. Underreporting is selective [9, 10, 13] 9. Social desirability bias [8] 10. Some components (e.g., self-employment) are difficult to measure [11, 14] 11. May overstate standards of living in the case of singular market structures (rationing, etc.) [16]	**Empirical** 3. Estimates with larger standard errors due to smaller sample size 4. Subject to some underreporting [1] 5. Unit and item nonresponse [21] 6. Certain items (e.g., alcohol and cigarettes) are underreported [12] 7. Data collection complex and expensive [11] 8. Opportunities for consumption smoothing may vary according to welfare status [17]

NOTES: Numbers in brackets denote the references listed in the sources.
SOURCES: [1] Meyer and Sullivan (2012); [2] Meyer and Sullivan (2011); [3] Meyer and Sullivan (2008); [4] Deaton (1997, 29), [5] World Bank (2015, 33); [6] Atkinson (2019, 59–60); [7] Brewer and O'Dea (2012); [8] Moore, Stinson, and Welniak (2000); [9] Carletto, Tiberti, and Zezza (2021); [10] Fisher, Reimer, and Carr (2010); [11] UNECE (2017); [12] Atkinson (2015); [13] Brewer, Etheridge, and O'Dea (2017); [14] Hurst, Li, and Pugsley (2014); [15] Johnson, Smeeding, and Torrey (2005); [16] Atkinson (1991); [17] Ravallion (2016); [18] Attanasio and Pistaferri (2016) ; [19] Beegle et al. (2016) ; [20] Browning, Crossley, and Winter (2014); [21] Meyer, Mok, and Sullivan. (2015).

Appendix C.

Construction of an Income Aggregate

This appendix will *not* discuss the construction of an income-based welfare indicator in detail. Rather, it will take advantage of an external source, the *Canberra Group Handbook on Household Income Statistics* (UNECE 2011). The *Handbook* lays out a conceptual framework that helps to understand what is meant by household (disposable) income, exactly, and provides useful practical guidelines for coming up with a comprehensive and accurate measure using the pieces of information that are typically collected by household surveys.

The *Handbook* is the culmination of decades of efforts toward the international harmonization of micro-level household income statistics, consistently with the standards that govern both the System of National Accounts and the collection of labor statistics. Early initiatives in this direction date back to the mid-1960s, and were carried out under the aegis of the United Nations Statistical Commission (OECD 2013, ch. 1). Building on this and other subsequent efforts, the Canberra Group, an international task force composed by experts from national statistics offices (NSOs), government departments, international organizations, and research agencies, was formed in 1996. The first release of the *Handbook* came in 2001 (Canberra Group 2001), and it represented a new benchmark for the harmonization of the concept of income and its measurement at the household level across countries. The 2001 guidelines were updated, with minor modifications, in 2011, establishing the current standard (UNECE 2011).[102]

The recommendations laid out in the *Handbook* greatly facilitate the task of constructing an income aggregate. To start, one must understand the income concept put forward by the *Handbook*, which is encapsulated in the definition below:

> *"Household income consists of all receipts whether monetary or in kind (goods and services) that are received by the household or by individual members of the household at annual or more frequent intervals, but excludes windfall gains and other such irregular and typically one-time receipts. Household income receipts are available for current consumption and do not reduce the net worth of the household through a reduction of its cash, the disposal of its other financial or non-financial assets or an increase in its liabilities."*
>
> (UNECE 2011, 9—10).

[102] "The only exceptions are in regard to the *Value of unpaid domestic services* and the *Value of services from household consumer durables*. These components were not included in the conceptual income definition of the first edition of the handbook, but were listed as "issues for the future." In this second edition of the handbook the two components have been included in the conceptual definition to align with the 2004 ICLS standard." (UNECE 2011, 9).

The quote contains all the elements that set household income apart from other components of the household's budget. In essence, the definition recommends that a measure of income be *comprehensive* and include, in principle, all monetary and in-kind resources flowing into the household's budget (Ellwood and Summers 1985); but the measure must also be *specific*, and take care to exclude some of these receipts.

First, changes in the value of financial and nonfinancial assets and liabilities over a reference period (holding gains or losses) are considered to be changes in net worth, and thus excluded from the conceptual and operational definitions of income. It may be helpful to mention that some common items, such as the sale of assets, loans obtained, and withdrawals from savings, are all examples of receipts that result from a reduction in net worth, and thus are not to be considered income.

The second type of exclusion concerns "atypical" incomes: windfall gains and other irregular lump sum receipts are excluded from the definition of income because they are not seen as representative of a household's usual situation and standard of living. A level of arbitrariness is embedded in the definition of what constitutes "atypical" gains, as for the case of the consumption aggregate, which also excludes infrequent expenditures (a more detailed discussion of this principle is in section 4.1).

The conceptual definition clarifies what income is and which monetary receipts should not be part of the income aggregate, but it remains silent on a number of issues of practical importance (for instance, whether or not imputed rent should be considered an in-kind capital gain, and therefore whether it should be included in the aggregate). Table C.1 lays out an operational definition of the income aggregate—sometimes called an *aggregation plan*—that is in line with the *Handbook* (UNECE 2011), one that resolves these ambiguities. Two different versions of the *nominal* income aggregate can be computed: gross income and disposable income.[103] None is necessarily superior in all contexts; each is suited to specific analytical purposes, although net or disposable income are typically considered representative of the household's command over resources.

Given their heterogeneous nature, the income components listed in table C.1 will almost always be collected across different modules within the questionnaire of a typical household budget survey. The table can function as a checklist, helping to settle the analyst's doubts as she scans the questionnaire in search of the information she needs.

The first set of income components, *employee income*, is typically collected in the employment module and usually rather unproblematic. The likeliest source of measurement error for this class of incomes is a failure to include all of its components, particularly when it comes to incomes received in kind from an employer. However, this piece of information is usually recorded in the questionnaire, together with cash incomes, and if one accepts the accuracy of respondents' valuations, then its inclusion should be easy enough (Smeeding and Weinberg 2001).

[103] Welfare analysts are often interested in *real* income, i.e., in adjusting the definition of income discussed in this section for geographic differences in the level of prices. Section 5 tackles price adjustments as they pertain to the consumption aggregate; that discussion can be applied, by and large, to the case of income.

TABLE C.1. The aggregation plan for total and disposable household income

Income component	Description
Y1 Income from employment	**Employee income =** Wages and salaries + Bonuses (commissions, tips, profit-sharing bonuses...) + Free or subsidized goods and services from an employer* + Severance and termination pay + Employers' social insurance contributions **Income from self-employment =** Revenue in cash from the sale of output + Value of goods produced for barter + Value of goods produced for own consumption* – Cost of inputs
Y2 Property income	**Property income =** Income from financial and nonfinancial assets, net of expenses + Net value of owner-occupied housing services (imputed rent)*
Y3 Current transfers received	**Public transfers =** Social security pensions or schemes + Social assistance benefits (excluding social transfers in kind) **Private transfers =** Pensions and other insurance benefits + Current transfers from nonprofit institutions + Current transfers from other households
Y = Y1 + Y2 + Y3 Total income	
Y4 Deductions	**Deductions =** Direct taxes (net of refunds) + Compulsory fees and fines + Current transfers paid to other households or nonprofit institutions + Employers' social insurance contributions
YD = Y1 + Y2 + Y3—Y4 Disposable income	

NOTE: An asterisk (*) indicates elements that are also included in the consumption aggregate.
SOURCES: Adapted from the Canberra Group Handbook (UNECE 2011) table 2.1, and McKay (2000), Table 17.1. The Canberra Group Handbook classifies imputed rent as "income from household production of services for own consumption." For simplicity, this category has been expunged here, and imputed rent has been considered part of property income.

Income from self-employment is notoriously more problematic in terms of measurement. This component includes both farm income (typically collected in the agriculture module) and nonfarm self-employment income (collected in the employment module, or in a dedicated section—see Dillon et al. 2021). The relevant concept is net income, or profit, that is, the value of output minus the cost of inputs. The main difficulty, especially in low-income contexts, is that "for a large number of households that are involved in agricultural or family business, personal and business incomings and outgoings are likely to be confused. Such households do not need the concept of income, so that respondents will not know what is required when asked about profits from farms or own enterprises." (Deaton 1997, 29).

Property income comprises returns, usually monetary, from financial assets (interest, dividends), from nonfinancial assets (rent), and from royalties (return for services of patented or copyrighted material) (UNECE 2011, 13). An estimate of imputed rent for households that occupy their own dwelling is also to be included in the income aggregate: the conceptual and empirical issues related to this component are identical to those that arise when estimating a consumption aggregate (a detailed discussion of the estimation of imputed rent is in section 4.5).

Finally, the inclusion of *public and private transfers* does not present particular challenges, provided that they are recorded by the questionnaire. This component may not represent a large share of household income overall, but it will be extremely relevant for some households, typically those at the bottom of the income distribution, so that its inclusion is important for the accurate estimation of a welfare measure (McKay 2000, 92).

Even when touching on issues related to measurement error, our discussion of income aggregation has assumed that all respondents are willing to answer all questions as truthfully as possible—as is well known, this is often not the case, given the issues of underreporting and nonresponse that are associated to income measurement through surveys (Meyer and Sullivan 2011; Meyer, Mok, and Sullivan 2015).[104] Another empirical issue is negative and zero incomes, which can be non-negligible in terms of frequency (Hlasny, Ceriani, and Verme 2020). Both negatives and zeros are, in principle, legitimate values for the income variable (for instance, think of households that are struggling with a bad business year). Unfortunately, inequality estimates are sensitive to their inclusion (or exclusion), even when the number of these values is small. Pyatt, Chen, and Fei (1980) and Chen, Tsaur, and Rhai (1982) remind us of the fact that the Gini index, for instance, can be greater than one in the presence of negative values. There is no consensus among analysts —or Stata commands that compute inequality indices— on how to deal with zero and negative values.[105] The appropriate vehicle for tackling these and other challenges to the construction of an income-based welfare indicator is, we believe, a dedicated set of guidelines. We leave this more ambitious effort to the future—a not too distant one, given the growing interest for income-based measures of living standards.

[104] Section 7.1 provides a general discussion of unit nonresponse and related adjustments.

[105] The command `ineqdeco` by Jenkins (1999) excludes zeros and negatives, while `ineqdec0` (again by Jenkins), `inequal7` by van Kerm (2001), the suite of DASP commands by Araar and Duclos (2007, 2021), and `ainequal` by Azevedo (2006) include them.

Appendix D.

Price Indices

This appendix illustrates the data underlying figure 5.3 in section 5.1.1, taken from Diewert's Chapter 19, in ILO's (2004) *Consumer price index manual*.

Table D.1 contains (artificial) prices for six commodities (p_1, ..., p_6), and corresponding quantities (q_1, ..., q_6) and expenditure shares (w_1, ..., w_6) for different situations (rows 1 — 5), that can be thought of as different years (time periods) or regions.

TABLE D.1. Example of commodity-specific price levels

Time (or region)	p_1	p_2	p_3	p_4	p_5	p_6
	Food	Energy	Traditional manufacturing	High-tech manufacturing	Traditional services	High-tech services
1	1.0	1.0	1.0	1.0	1.0	1.0
2	1.2	3.0	1.3	0.7	1.4	0.8
3	1.0	1.0	1.5	0.5	1.7	0.6
4	0.8	0.5	1.6	0.3	1.9	0.4
5	1.0	1.0	1.6	0.1	2.0	0.2
	q_1	q_2	q_3	q_4	q_5	q_6
1	1.0	1.0	2.0	1.0	4.5	0.5
2	0.8	0.9	1.9	1.3	4.7	0.6
3	1.0	1.1	1.8	3.0	5.0	0.8
4	1.2	1.2	1.9	6.0	5.6	1.3
5	0.9	1.2	2.0	12.0	6.5	2.5
	w_1	w_2	w_3	w_4	w_5	w_6
1	10.0	10.0	20.0	10.0	45.0	5.0
2	6.8	19.2	17.5	6.5	46.7	3.4
3	6.5	7.2	17.7	9.8	55.6	3.1
4	5.5	3.4	17.3	10.3	60.6	3.0
5	4.5	6.0	16.0	6.0	65.0	2.5

SOURCE: Authors' elaboration based on ILO (2004: p. 346).

As noted by Diewert (2004, p. 346), the movements of prices and quantities in table D.1 are more pronounced than the year-to-year movements that would be encountered in a typical country (and even more so the within-survey period movements). The data in table D.1 illustrate the problem facing compilers of the consumer price index (CPI): year-to-year price and quantity movements are far from being proportional across commodities. It is this phenomenon that is largely responsible for the finding in table D.2, that the choice of index number formula matters.

TABLE D.2. Estimated price indices

Time (or region)	Laspeyres	Paasche	Fisher	Törnquist
1	1.0000	1.0000	1.0000	1.0000
2	1.4200	1.3823	1.4010	1.4052
3	1.3450	1.2031	1.2721	1.2890
4	1.3550	1.0209	1.1761	1.2268
5	1.4400	0.7960	1.0706	1.2477

SOURCE: Authors' elaborations based on ILO (2004: p. 345 and p. 349).

Appendix E.

Questionnaire Design

DZ's *Guidelines* came out around the same time as *Designing Household Survey Questionnaires for Developing Countries* (Grosh and Glewwe 2000), a three-volume book laying out recommendations and lessons learned over the course of a multi-decade international program supporting the collection of survey data, the World Bank's Living Standards Measurement Study (LSMS). This is a testament to the international community's early awareness that data quality and comparability not only matter, but come first, as a *prerequisite* for sound analysis.

Time has only strengthened this conviction, as the literature on questionnaire design has continued to grow. The greatest strides have been made in amassing a wealth of empirical, often experimental evidence on exactly how much and in what ways the "details" of questionnaire design affect the measurement of living standards, and which design solutions produce the most accurate estimates. Though vast, this new evidence base remains uneven in coverage: the collection of *consumption* and expenditure data has received more attention than the collection of *income* data, and among expenditure categories, food has pulled focus. Although some gaps remain to be filled, national statistical offices (NSOs) around the world are now in a better position to make good design decisions, and analysts are ever more aware of the impact that changes in questionnaire design have on final estimates (particularly on comparisons).

Grosh and Glewwe (2000) continue to be an essential reference for their breadth of scope, commitment to practical applicability, and for the inclusion of a set of template modules to guide the task of constructing a questionnaire. Taking a similar approach to summarizing the current evidence and best practice would require far more space than what is appropriate for our scope. Instead, this section provides an annotated reference list, organized by nine (frequently asked) questions relevant to the work of both data producers and data analysts. Most of the answers are drawn (often verbatim) from the *LSMS Guidebook on Food Data Collection*, a joint effort by the Food and Agriculture Organization and The World Bank that has originated a set of recommendations endorsed by the 49th session of the United Nations Statistical Commission in New York, March 29, 2018. In what follows, page numbers refer to FAO and World Bank (2018).

1. *Should food expenditures be recorded via diary or recall? What is the most suitable reference period for the food module?*

Low-income countries are advised to adopt recall interviews and a seven-day recall period, as this method provides the best balance between accuracy and cost-effectiveness. Any survey using diary methods must be closely supervised to ensure proper completion, especially in areas where illiteracy rates are high. The reference period should not exceed 14 days. Any change in the recall period or method (recall versus diary) should be accompanied by an experimental component aimed at assessing the impact of the change on survey estimates (p. 33).[106]

2. How should food items be listed (coverage, number of items, level of detail)?

The length and level of detail of the list of food items used to collect data on food expenditures matters for final estimates. The design of the list should balance the lower costs and interview time and the smaller chance of memory lapses associated with "shorter" lists, with the better recall and more comprehensive reporting (but also the risk of respondent fatigue) associated with "longer" lists.

The adoption of a food classification system can help in meeting the criteria of comprehensiveness and specificity that the food list should be built around. Survey designers are encouraged to use the Classification of Individual Consumption According to Purpose (COICOP) system, as it is currently the basis of classification used in a wide number of datasets, in line with the requirements specified in the System of National Accounts (pp. 37—38).[107]

3. Should the questionnaire record food acquisition or food consumption? In what way?

The difference between food *acquisition* and *consumption* is discussed in section 4.2.1 of this report. The choice of collecting data on either concept depends on the purpose of the survey, and mixed approaches are not rare. Regardless of which concept is recorded, it is essential to collect data on food obtained through nonmarket sources (own production and in-kind receipts). Surveys should be designed so that it is clear to respondents, enumerators, and data users exactly what information (acquisition, consumption, or both) is to be reported. Sources of incomplete or ambiguous enumeration commonly found in current survey practices, such as filter questions that make answers conditional to having consumed a given item, followed by acquisition questions, should be avoided (pp. 34—35).[108]

4. How relevant is the seasonality of food consumption? How best to account for it when planning interviews?

There is abundant evidence that food consumption and expenditure display systematic seasonal variation on a yearly, monthly, and weekly basis. The only way to accurately capture habitual consumption for each household is to survey them multiple times over the year, but this is also the most expensive option, and in practice, it is difficult to implement. Data collection spread over the year, but with only one interview per household, results in an accurate estimate of average consumption for the population, but with excess variability around the mean; however, it is a viable option in resource-constrained contexts. Regardless of the

[106] Further readings include Backiny-Yetna, Steele, and Djima (2017); Bee, Meyer, and Sullivan (2012); Beegle et al. (2012); Bradburn (2010); Brzozowski, Crossley, and Winter (2017); De Weerdt et al. (2016); Friedman et al. (2017); Gibson, Beegle, De Weerdt and Friedman (2015); Hurd and Rohwedder (2009); Scott and Amenuvegbe (1991); Schündeln (2018); and Troubat and Grünberger (2017).

[107] Further readings include Beegle et al. (2012); De Weerdt et al. (2016); Finn and Ranchhod (2015); Jolliffe (2001); Pradhan (2009) and United Nations (2018).

[108] Further readings include Conforti, Grünberger, and Troubat (2017); Smith, Alderman, and Aduayom (2006), and Troubat and Grünberger (2017).

approach chosen, care should be exercised to ensure that enumeration is equally spread throughout the days of the week and the month, and changes in timing in holidays, festivals, and harvests must be considered (p. 34).[109]

5. What is the best way to capture food consumed away from home?

The practice of collecting information on food away from home with just one question should be discontinued. The importance of food away from home warrants the design of a separate module, based on a clear definition of food away from home. Data collection should preferably be done at the individual level. For all individuals who report having consumed meals outside the home, the minimum information attained should be on the value of the meals by meal event (breakfast, lunch, dinner, and snacks) (p. 36).[110]

6. In which cases is meal participation relevant? How to record meal partakers?

Recording the exact number of people who consumed the reported amount of food consumed by the household (meal partakers) is important for the accurate computation of food consumption per capita. Survey designers should consider adding an individual household member-based meal module. As collecting information on individuals is expensive and difficult, this can be implemented as part of the module collecting information on food away from home. If an individual member-based meal module cannot be adopted, simpler (though less preferred) alternatives are available (pp. 35—36).[111]

7. What is the best way to use nonstandard measurement units?

Qualitative feedback from field practitioners and extensive feedback from initial piloting of those methods, suggest that allowing nonstandard units of measurement can increase the accuracy of reported quantities, primarily by reducing respondent burden. To properly benefit from allowing nonstandard unit options, reporting must be paired with a framework for consistently converting nonstandard units into standard units, based on reliably documented conversion factors. Oseni, Durazo, and McGee (2017) provide detailed guidance on how to do that effectively (p. 39).

8. What do food modules actually look like in low- and middle-income countries?

Smith, Dupriez, and Troubat (2014) provide an overview of the questionnaire design choices adopted by household consumption and expenditure surveys around the world, and an assessment of their compliance with best practices.

[109] Further readings include D'Souza and Jolliffe (2012); Gilbert, Christiaensen, and Kaminski (2016); Jolliffe and Serajuddin (2015), and Troubat and Grünberger (2017).

[110] Further readings include Borlizzi. Delgrossi, and Cafiero (2017); Farfàn, Genoni, and Vakis (2017); Farfàn et al. (2019), and Smith (2015).

[111] Further readings include Bouis, Haddad, and Kennedy (1992); Bouis (1994); Conforti, Grünberger, and Troubat (2017); Gibson and Rozelle (2002); and Weisell and Dop (2012).

9. How to collect data on nonfood expenditures?

For the collection of data on nonfood expenditures, the evidence base is not yet rich enough to warrant the creation of guidelines as detailed and comprehensive as those targeting food data collection. The chapter on Consumption, as well as the chapters on Housing, Education, and Health in Grosh and Glewwe (2000), remain useful sources of applicable recommendations. The most notable recent additions to these classic references are a set of guidelines for designing household surveys, which includes recommendations on nonfood expenditure modules (Oseni et al. 2021); guidelines for the collection of education data (Oseni et al. 2018); and some new evidence on the effects of different health expenditure data collection modes on final estimates (Lu et al. 2009; Xu et al. 2009).

References

Aguinis, H., R. K. Gottfredson, and H. Joo. 2013. "Best-practice Recommendations for Defining, Identifying, and Handling Outliers." *Organizational Research Methods* 16, no. 2: 270 — 301.

Alkire, S., J. M. Roche, P. Ballon, J. Foster, M.E. Santos, and S. Seth. 2015. *Multidimensional Poverty Measurement and Analysis*. USA: Oxford University Press.

Almås, I. 2012. "International Income Inequality: Measuring PPP bias by estimating Engel curves for food." *American Economic Review* 102, no. 2: 1093 — 1117.

Almås, I., T. K. M. Beatty, and T.F. Crossley. 2018. "Lost in Translation: What Do Engel Curves Tell Us About the Cost of Living?" CESifo Working Paper Series no. 6886.

Amendola, N. and G. Vecchi. 2008. "Growth, Inequality, and Poverty in Madagascar, 2001 — 2005." World Bank, Africa Region Working Paper no. 111.

Amendola, N. and G. Vecchi. 2014. "Durable Goods and Poverty Measurement." World Bank *Policy Research Working Paper* no. 7105.

Amendola, N. and G. Vecchi. 2021. "Durable Goods and Welfare Measurement." *Journal of Economic Surveys*. Forthcoming.

Amendola, N., G. Gabbuti, and G. Vecchi. 2018. "On the Use of Composite Indices in Economic History. Lessons from Italy, 1861 — 2017." The Historical Household Budgets Project Working Papers Series no. 11.

American Association for Public Opinion Research. 2016. *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys*. 9th edition. Ann Arbor, Michigan: AAPOR.

Andridge, R. R. and R. J. Little. 2010. "A Review of Hot Deck Imputation for Survey Non-response." *International Statistical Review* 78: 40 — 64.

Araar, A., N. Choueiri, and P. Verme. 2015. "The Quest for Subsidy Reforms in Libya." World Bank Policy Research Working Paper no. 7225.

Araar, A. and J.-Y. Duclos. 2007. "DASP: Distributive Analysis Stata Package." Statistical Software Components S 456872, Boston College Department of Economics.

Araar, A. and J.-Y. Duclos. 2021. "User Manual for Stata Package DASP: Version 3.01." Partnership for Economic Policy, World Bank, UNDP and Université Laval.

Ardington, C., D. Lam, M. Leibbrandt, and M. Welch. 2006. "The Sensitivity to Key Data Imputations of Recent Estimates of Income Poverty and Inequality in South Africa." *Economic Modelling* 23, no. 5: 822 — 835.

Arrow, K. J. 1958. *The Measurement of Price Changes. The Relationship of Prices to Economic Stability and Growth.* Joint Economic Committee US Congress, US Government Printing Office.

Asfaw, A. and J. V. Braun. 2004. "Is Consumption Insured Against Illness? Evidence on Vulnerability of Households to Health Shocks in Rural Ethiopia." *Economic Development and Cultural Change* 53, no. 1: 115 — 129.

Atkinson, A. B. 1987. "On the Measurement of Poverty." *Econometrica* 1: 749 — 764.

Atkinson, A. B. 1989. *Poverty and Social Security*. Harvester Wheatsheaf.

Atkinson, A. B. 1991. "Comparing Poverty Rates Internationally: Lessons from Recent Studies in Developed Countries." *The World Bank Economic Review* 5, no. 1: 3 — 21.

Atkinson, A. B. 2015. *Inequality. What Can Be Done?* Cambridge: Harvard University Press.

Atkinson, A. B. 2019. *Measuring Poverty around the World*. Princeton and Oxford: Princeton University Press.

Atkinson, A. B., B. Cantillon, E. Marlier, and B. Nolan. 2002. *Social Indicators: The EU and Social Inclusion.* Oxford: Oxford University Press.

Atkinson, A. B., L. Rainwater, L. and T. M. Smeeding. 1995. *Income Distribution in OECD Countries. Evidence from the Luxembourg Income Study*. Paris: OECD.

Atkinson, R. and S. Da Voudi. 2000. "The Concept of Social Exclusion in the European Union: Context, Development and Possibilities." *JCMS: Journal of Common Market Studies* 38, no. 3: 427 – 448.

Attanasio, O. P. and L. Pistaferri. 2016. "Consumption Inequality." *Journal of Economic Perspectives* 30, no. 2: 3 – 28.

Azevedo, J. P. 2006. "AINEQUAL: Stata module to compute measures of inequality." Statistical Software Components S456748, Boston College Department of Economics.

Backiny-Yetna, P., D. Steele, and I. Y. Djima. 2017. "The Impact of Household Food Consumption Data Collection Methods on Poverty and Inequality Measures in Niger." *Food Policy* 72: 7 – 19.

Balcazar, C. F., L. Ceriani, S. Olivieri, and M. Ranzani. 2017. "Rent-Imputation for Welfare Measurement: A Review of Methodologies and Empirical Findings." *Review of Income and Wealth* 63: 881 – 898.

Banks, J., R. Blundell, and A. Lewbel. 1997. "Quadratic Engel Curves and Consumer Demand." *Review of Economics and Statistics* 79, no. 4: 527 – 539.

Barnett, V. and T. Lewis 1994. *Outliers in Statistical Data*. 3rd edition. New York: John Wiley & Sons.

Batana, Y. M. and J. Cockburn. 2018. "Do Demographics Matter for African Child Poverty?" World Bank Policy Research Working Paper no. 8426.

Bee, A., B. D. Meyer, and J. X. Sullivan. 2012. "The Validity of Consumption Data: Are the Consumer Expenditure Interview and Diary Surveys Informative?" NBER Working Paper Series no. 18308.

Beegle, K., L. Christiaensen, A. Dabalen, and I. Gaddis. 2016. *Poverty in a Rising Africa*. Washington, DC: The World Bank.

Beegle, K., J. De Weerdt, J. Friedman, and J. Gibson. 2012. "Methods of Household Consumption Measurement Through Surveys: Experimental Results from Tanzania." *Journal of Development Economics* 98: 3 – 18.

Belotti, F., G. Mancini, and G. Vecchi. 2021. "Outlier Detection for Welfare Analysts Using Stata." Unpublished manuscript, Università di Roma "Tor Vergata."

Berry, F., B. Graf, M. M. Stanger, and M. Ylä-Jarkko. 2019. *Price Statistics Compilation in 196 Economies: The Relevance for Policy Analysis*. Washington, DC: International Monetary Fund.

Bethlehem, J., F. Cobben, and B. Schouten. 2011. *Handbook of Nonresponse in Household Surveys*. New York: John Wiley & Sons.

Bhattacharya, S., V. L. Falcao, and R. Puri. 2017. "The Public Distribution System in India: Policy Evolution and Program Delivery Trends." In Bhattacharya, S., V. L. Falcao, and R. Puri (eds.) *The 1.5 Billion People Question: Food, Vouchers or Cash Transfers?*: 43 – 106. Washington, DC: The World Bank.

Blackorby, C. and D. Donaldson. 1987. "Welfare Ratios and Distributionally Sensitive Cost-Benefit Analysis." *Journal of Public Economics* 34, no. 3: 265 – 290.

Blackorby, C. and D. Donaldson. 1988. "Money Metric Utility: A Harmless Normalization?" *Journal of Economic Theory* 46, no. 1: 120 – 129.

Blinder, A. S. 1985. Comment on "Measuring Income: What Kind Should Be In?" by Ellwood, D. T., and H. Lawrence, and L. H Summers, in Bureau of the Census, "Conference on measurement of noncash benefits: Proceedings vol. 1." Washington, DC: Bureau of the Census.

Bodvarsson, Ö. B., N. B. Simpson, and C. Sparber. 2015. "Migration Theory." In Chiswick, B, and P. Miller (eds.) *Handbook of the Economics of International Migration*. Amsterdam, New York and Oxford: North-Holland.

Borlizzi, A., M. E. Delgrossi, and C. Cafiero. 2017. "National Food Security Assessment through the Analysis of Food Consumption Data from Household Consumption and Expenditure Surveys: The Case of Brazil's Pesquisa de Orçamento Familiares 2008/09." *Food Policy* 72: 20–26.

Bouis, H. E. 1994. "The Effect of Income on Demand for Food in Poor Countries: Are our Food Consumption Databases Giving Us Reliable Estimates?" *Journal of Development Economics* 44, no. 1: 199–226.

Bouis, H. E., L. Haddad, and E. Kennedy. 1992. "Does It Matter How We Survey Demand for Food? Evidence from Kenya and the Philippines." *Food Policy* 17, no. 5: 349–360.

Box, G. E. and D. R. Cox. 1964. "An Analysis of Transformations." *Journal of the Royal Statistical Society: Series B (Methodological)* 26, no. 2: 211–243.

Bradburn, N. 2010. "Recall Period in Consumer Expenditure Surveys Program." National Opinion Research Center University of Chicago unpublished manuscript.

Brandolini, A. and G. D'Alessio. 2001. "Measuring Well-Being in the Functioning Space." In Chiappero-Martinetti, M. and A. Sen (eds.) *Debating Global Society: Reach and Limits of the Capability Approach*. Milano: Feltrinelli.

Brandolini, A., S. Magri, and T. M. Smeeding. 2010. "Asset-Based Measurement of Poverty." *Journal of Policy Analysis and Management* 29: 267–284.

Brandolini, A. and J. Micklewright. 2020. "Tony Atkinson's New Book, Measuring Poverty Around the World: Some Further Reflections." ECINEQ Working Paper no. 518.

Brewer, M. and C. O'Dea. 2012. "Measuring Living Standards with Income and Consumption: Evidence from the UK." The Institute for Fiscal Studies Working Papers no. W12/12.

Brewer, M., B. Etheridge, and C. O'Dea. 2017. "Why Are Households that Report the Lowest Incomes So Well-off?" *The Economic Journal* 127: F24–F49.

Brick, J. M. 2013. "Unit Nonresponse and Weighting Adjustments: A Critical Review." *Journal of Official Statistics* 29, no. 3: 329–353.

Brick, J. M. and J. M. Montaquila. 2009. "Nonresponse and Weighting." In Rao C.R. (ed.) *Handbook of statistics* 29, part A: 163–185. Elsevier.

Browning, M., T. F. Crossley, and J. Winter. 2014. "The Measurement of Household Consumption Expenditures." *Annual Review of Economics* 6, no. 1: 475–501.

Brueckner, J. K., J. F. Thisse, and Y. Zenou. 1999. "Why Is Central Paris Rich and Downtown Detroit Poor?: An Amenity-Based Theory." *European Economic Review* 43, no. 1: 91–107.

Brzozowski, M., Crossley, T. F., and Winter, J. K. 2017. "A comparison of recall and diary food expenditure data." *Food Policy* 72: 53–61.

Buhmann, B., L. Rainwater, G. Schmaus, and T. M. Smeeding. 1988. "Equivalence Scales, Well-being, Inequality, and Poverty: Sensitivity Estimates Across Ten Countries Using the Luxembourg Income Study (LIS) Database." *Review of Income and Wealth* 34, no. 2: 115–142.

Caeyers, B., N. Chalmers, and J. De Weerdt. 2012. "Improving Consumption Measurement and Other Survey Data through CAPI: Evidence from a Randomized Experiment." *Journal of Development Economics* 98, no. 1: 19–33.

Canberra Group. 2001. *Expert Group on Household Income Statistics: Final Report and Recommendations*. Ottawa: The Canberra Group.

Capéau, B. and S. Dercon. 2006. "Prices, Unit Values and Local Measurement Units in Rural Surveys: An Econometric Approach with an Application to Poverty Measurement in Ethiopia." *Journal of African Economies* 15, no. 2: 181–211.

Carletto, G., M. Tiberti, and A. Zezza. 2021. "Measure for Measure: Comparing Survey Based Estimates of Income and Consumption for Rural Households." *The World Bank Research Observer.*

Ceccarelli, C., A. Cutillo, and D. Di Laurea. 2009. "Metodologie per la stima degli affitti figurativi ed impatto sulla distribuzione del reddito." *Rivista di Statistica Ufficiale* 11, no. 1: 17—38.

Central Statistics Organization 2018, *Afghanistan Living Conditions Survey 2016—17.* Kabul, CSO.

Ceriani, L. and P. Verme. 2019. "The Inequality of Extreme Incomes." ECINEQ Working Paper n. 490.

Ceriani, L., S. Olivieri, and M. Ranzani. 2019a. "Evaluating the Accuracy of Homeowners' Self-Assessed Rent in Metropolitan Lima." World Bank Policy Research Working Paper no. 8983.

Ceriani, L., S. Olivieri, and M. Ranzani. 2019b. "Housing, Imputed Rent, and Households' Welfare." World Bank Poverty and Equity Global Practice Working Paper, no. 213.

Chakravarty, S. R. (ed.) 2019. *Poverty, Social Exclusion and Stochastic Dominance.* Springer Nature.

Chaudhuri, S. and C. Paxson. 2002. "Smoothing Consumption under Income Seasonality: Buffer Stocks Vs. Credit Markets," Discussion Paper 0102-54, Columbia University, Department of Economics.

Chaudhuri, S. and M. Ravallion. 1994. "How Well Do Static Indicators Identify the Chronically Poor?" *Journal of Public Economics* 53, no. 3: 367—394.

Chen, C. N., T. W. Tsaur, and T. S. Rhai. 1982. "The Gini Coefficient and Negative Income." *Oxford Economic Papers* 34, no. 3: 473—478.

Chen, S. and M. Ravallion. 1996. "Data in Transition: Assessing Rural Living Standards in Southern China." *China Economic Review* 7, no. 1: 23—56.

Chen, W. H. and J.-Y Duclos. 2011. "Testing for Poverty Dominance an Application to Canada." *Canadian Journal of Economics* 44, no. 3: 781—803.

Chen, X., R. Mungai, S. Nakamura, T. Pearson, A. E. Wambile, and N. Yoshida. 2020. "How Useful Is CPI Price Data for Spatial Price Adjustment in Poverty Measurement? A Case from Ghana." World Bank Policy Research Working Paper no. 9388.

Chiang, A. C. 1984. *Fundamental Methods of Mathematical Economics.* 3rd edition. McGraw-Hill.

Chiappori, P. A. and C. Meghir. 2015. "Intrahousehold Inequality." In Atkinson A.B. and F. Bourguignon (eds.) *Handbook of Income Distribution,* vol. 2: 1369—1418. Elsevier.

Christensen, G., J. Freese, and E. Miguel. 2019. *Transparent and Reproducible Social Science Research: How To Do Open Science*. University of California Press.

Christiano, L. J. 1987. "Is Consumption Insufficiently Sensitive to Innovations in Income?" *The American Economic Review* 77, no. 2: 337—341.

Clark, R. L., A. Lusardi, O. S., and Mitchell. 2021. "Financial Fragility During the COVID-19 Pandemic." University of Pennsylvania Working Paper no. 25.

Clark, S., R. Hemming, and D. Ulph. 1981. "On Indices for the Measurement of Poverty." *The Economic Journal* 91, no. 362: 515—526.

Comin, F., M. Qizilbash, and S. Alkire. (eds.) 2008. *The Capability Approach. Concepts, Measures, and Applications.* Cambridge: Cambridge University Press.

Conforti, P., K. Grünberger, and T. Troubat. 2017. "The Impact of Survey Characteristics on the Measurement of Food Consumption." *Food Policy* 72: 43—52.

Coondoo, D., A. Majumder, and S. Chattopadhyay. 2011. "Estimating Spatial Consumer Price Indices Through Engel Curve Analysis." *Review of Income and Wealth* 57, no. 1: 138—155.

Costa, D. L. 2001. "Estimating Real Income in the United States from 1888 to 1994: Correcting CPI Bias Using Engel Curves." *Journal of Political Economy* 109, no. 6: 1288—1310.

Coudouel, A., J. S. Hentschel, and Q. T. Wodon. 2002. *Poverty Measurement and Analysis. A Sourcebook for Poverty Reduction Strategies.* Washington, DC: The World Bank.

Coulter, F. A., F. A. Cowell, and S. P. Jenkins. 1992a. "Equivalence Scale Relativities and the Extent of Inequality and Poverty." *The Economic Journal* 102, no. 414: 1067 – 1082.

Coulter, F. A., F. A. Cowell, and S. P. Jenkins. 1992b. "Differences in Needs and Assessment of Income Distributions." *Bulletin of Economic Research* 44, no. 2: 77 – 124.

Cowell, F. A. and E. Flachaire. 2007. "Income Distribution and Inequality Measurement: The Problem of Extreme Values." *Journal of Econometrics* 141, no. 2: 1044 – 1072.

Cowell, F. A. and Flachaire, E. 2015. "Statistical methods for distributional analysis." In Atkinson A. B. and F. Bourguignon (eds.) *Handbook of Income Distribution,* 2: 359 – 465. Elsevier.

Cowell, F. A. and M. P. Victoria-Feser. 1996a. "Robustness Properties of Inequality Measures." *Econometrica* 64, no. 1: 77 – 101.

Cowell, F. A. and M. P. Victoria-Feser. 1996b. "Poverty Measurement with Contaminated Data: A Robust Approach." *European Economic Review* 40, no. 9: 1761 – 1771.

Cowell, F. A. and M. P. Victoria-Feser. 2002. "Welfare Rankings in the Presence of Contaminated Data." *Econometrica* 70, no. 3: 1221 – 1233.

Cowell, F. A. and M. P. Victoria-Feser. 2006. "Distributional Dominance with Trimmed Data." *Journal of Business and Economic Statistics* 24, no. 3: 291 – 300.

Cowell, F. A. and M. P. Victoria-Feser. 2007. "Robust Stochastic Dominance: A Semi-Parametric Approach." *The Journal of Economic Inequality* 5, no. 1: 21 – 37.

Cutler, D. M. and L. F. Katz. 1992. "Rising Inequality? Changes in the Distribution of Income and Consumption in the 1980s." NBER Working Paper Series no. 3964.

D'Alessio, G. 2017. "Benessere, contesto socio-economico e differenze di prezzo: il divario tra Nord e Sud." Banca d'Italia Quaderni di Economia e Finanza, n. 385/2017.

D'Alessio, G. 2020. "A Comparative Evaluation of Poverty Measures in the Italian Survey of Household Income and Wealth." *Politica Economica* 36, no. 1: 101 – 140.

D'Souza, A. and D. Jolliffe. 2012. "Rising Food Prices and Coping Strategies: Household-Level Evidence from Afghanistan." *Journal of Development Studies* 48, no. 2: 282 – 299.

D'Souza, A. and S. Tandon. 2019. "Intrahousehold Nutritional Inequities in Rural Bangladesh." *Economic Development and Cultural Change* 67, no. 3: 625 – 657.

Dabalen, A., I. Gaddis, and N. T. V. Nguyen. 2016. "CPI Bias and Its Implications for Poverty Reduction in Africa." World Bank Policy Research Working Paper no. 7907.

Dang, H. A., D. Jolliffe, and C. Carletto. 2019. "Data Gaps, Data Incomparability, and Data Imputation: A Review of Poverty Measurement Methods for Data-Scarce Environments." *Journal of Economic Surveys* 33, no. 3: 757 – 797.

Davidson, R. 2012. "Statistical Inference in the Presence of Heavy Tails." *The Econometrics Journal* 15, no. 1: C31 – C53.

Davidson, R. and Duclos, J.-Y. 2000. "Statistical Inference for Stochastic Dominance and for the Measurement of Poverty and Inequality." *Econometrica* 68, no. 6: 1435 – 1464.

Davies, L. and U. Gather. 1993. "The Identification of Multiple Outliers." *Journal of the American Statistical Association* 88, no. 423: 782 – 792.

De Leeuw, E. and W. De Heer. 2002. "Trends in Household Survey Nonresponse: A Longitudinal and International Comparison." In Groves, R. M., D. A. Dillman, J. L. Eltinge, and R. J. A. Little (eds.) *Survey Nonresponse*: 41 – 54. New York: John Wiley & Sons.

De Vos, K. and M. A. Zaidi. 1997. "Equivalence Scale Sensitivity of Poverty Statistics for the Member States of the European Community." *Review of Income and Wealth* 43, no. 3: 319 – 333.

De Vreyer, P. and S. Lambert. 2018. "By Ignoring Intra-Household Inequality Do We Underestimate the Extent of Poverty?" PSE Working Papers halshs-01724194, HAL.

De Waal, T., J. Pannekoek, and S. Scholtus. 2011. *Handbook of Statistical Data Editing and Imputation.* New York: John Wiley & Sons.

De Weerdt, J., K. Beegle, J. Friedman, and J. Gibson. 2016. "The Challenge of Measuring Hunger through Survey." *Economic Development and Cultural Change* 64, no. 4: 727 – 758.

Deaton, A. 1987. "Estimation of Own- and Cross-Price Elasticities from Household Survey Data." *Journal of Econometrics* 36, no. 1 – 2: 7 – 30.

Deaton, A. 1988. "Quality, Quantity and Spatial Variation of Prices." *The American Economic Review* 78, no. 3: 418 – 430.

Deaton, A. 1990. "Price Elasticities from Survey Data." *Journal of Econometrics* 44, no. 3: 281 – 309.

Deaton, A. 1992. "Saving and Income Smoothing in Côte d'Ivoire." Research Program in Development Studies, Woodrow Wilson School, Princeton University.

Deaton, A. 1997. *The Analysis of Household Surveys: A Microeconometric Approach to Development Policy*. Washington, DC: The World Bank.

Deaton, A. 1998. "Getting Prices Right: What Should Be Done?" *Journal of Economic Perspectives* 12, no. 1: 37 – 46.

Deaton, A. and O. Dupriez. 2011. "Purchasing Power Parity Exchange Rates for the Global Poor." *American Economic Journal: Applied Economics* 3, no. 2: 137 – 66.

Deaton, A. and Grosh, M. 2000. "Consumption." In Grosh, M. and P. Glewwe (eds.) *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study. Volume One*: 91 – 134. Washington, DC: The World Bank.

Deaton, A. and J. Muellbauer. 1980. *Economics and Consumer Behavior*. Cambridge: Cambridge University Press.

Deaton, A. and A. Tarozzi. 2005. "Prices and Poverty in India." In Deaton, A. and V. Kozel (eds.) *The Great Indian Poverty Debate:* 381 – 41. New Delhi, India: MacMillian.

Deaton, A. and S. Zaidi. 1999. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." Woodrow Wilson School Development Studies Paper no. 192.

Deaton, A. and S. Zaidi. 2002. "Guidelines for Constructing Consumption Aggregates for Welfare Analysis." Working Paper no. 135.

Decancq, K. M., M. Fleurbaey, and E. Schokkaert. 2015. "Inequality, Income, and Well-Being." In Atkinson A. B. and F. Bourguignon (eds.) *Handbook of Income Distribution,* 2A: 67 – 140. Elsevier.

Demertzis, M., M. Domínguez-Jiménez, and A. Lusardi. 2020. "The financial fragility of European households in the time of COVID-19." Policy Contribution no. 2020/15.

Dempster, A. and D. Rubin. 1983. "Incomplete Data in Sample Surveys." *Sample Surveys* 2: 3 – 10.

Dercon, S. (ed.) 2005a. *Insurance Against Poverty*. New York: Oxford University Press.

Dercon, S. 2005b. "Risk, Poverty and Vulnerability in Africa." *Journal of African Economies* 14, no. 4: 483 – 488.

Dewald, W. G., J. G. Thursby, and R. G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit and Banking Project." *The American Economic Review* 76, no. 4: 587 – 603.

Diewert, W. E. 1976. "Exact and Superlative Index Numbers." *Journal of Econometrics* 4, 2: 115 – 145.

Diewert, W. E. 2001. "The Consumer Price Index and Index Number Purpose." *Journal of Economic and Social Measurement* 27, no. 3 and 4: 167 – 248.

Diewert, W. E. 2008. "OECD Workshop on Productivity Analysis and Measurement: Conclusions and Future Directions." In *Proceedings from the OECD Workshops on Productivity Measurement and Analysis*: 11-36. Paris: OECD.

Diewert, W. E. 2013. "Irving Fisher and Index Number Theory." *Journal of the History of Economic Thought* 35, no. 2: 199 – 232.

Diewert, W. E., J. Greenlees, and C. R. Hulten (eds.) 2010. *Price index concepts and measurement*. Chicago: University of Chicago Press.

Diewert, W. E. and C. Shimizu. 2019. "Measuring the Services of Durables and Owner Occupied Housing." Vancouver School of Economics Discussion Paper no. 19-04.

Diewert, W. E. and H. Wei, H. 2017. "Getting Rental Prices Right for Computers: Reconciling Different Perspectives on Depreciation." *Review of Income and Wealth* 6: S149—S168.

Dillon, A., C. Carletto, S. Gourlay, P. Wollburg, and A. Zezza. 2021. "Agricultural Survey Design: Lessons from the LSMS-ISA and Beyond." Washington, DC: World Bank.

Doemeland, D. and J. Trevino. 2014. "Which World Bank Reports Are Widely Read?" World Bank Policy Research Working Paper no. 6851.

Duan, N. 1983. "Smearing Estimate: A Nonparametric Retransformation Method." *Journal of the American Statistical Association* 78, no. 383: 605—610.

Duclos, J.-Y. and A. Araar. 2006. *Poverty and Equity: Measurement, Policy and Estimation with DAD*. Springer.

Dupriez, O. 2007. "Building a Household Consumption Database for the Calculation of Poverty PPPs." World Bank technical note.

Dupriez, O. and E. Boyko. 2010. "Dissemination of Microdata Files: Principles Procedures and Practices." International Household Survey Network Working Paper no. 5.

ECASTD. 2016. ECAPOV 2.0 Harmonization Guidelines. World Bank Europe and Central Asia Team for Statistical Development. January 2019. Internal document.

Ellwood, D. T. and L. H. Summers. 1985."Measuring Income: What Kind Should Be In," in Bureau of the Census, Conference on the Measurement of Noncash Benefits, Proceedings, vol. 1: 8—27.

Éltetõ, Ö. and É. Havasi. 2002. "Impact of Choice of Equivalence Scale on Income Inequality and on Poverty Measures," *Review of Sociology of the Hungarian Sociological Association* 8: 137—48.

Fafchamps, M., D. McKenzie, S. Quinn, and C. Woodruff. 2012. "Using PDA Consistency Checks to Increase the Precision of Profits and Sales Measurement in Panels." *Journal of Development Economics* 98, no. 1: 51—57.

FAO. 2005a. "Equivalence Scales. General Aspects." EASYPol Series no. 032.

FAO. 2005b. "Equivalence Scales. Subjective Methods." EASYPol Series no. 033.

FAO. 2005c. "Equivalence Scales. Objective Methods." EASYPol Series no. 034

FAO and The World Bank. 2018. *Food Data Collection in Household Consumption and Expenditure Surveys. Guidelines for low- and middle-income countries.* Rome: FAO.

Farfán, G., M. E. Genoni, and E. Vakis. 2017. "You Are What (and Where) You Eat: Capturing Food Away from Home in Welfare Measures." *Food Policy* 72: 146—156.

Farfán, G., K. R. McGee, J. Perng, and R. Vakis. 2019. "Poverty Measurement in the Era of Food Away from Home: Testing Alternative Approaches in Vietnam." World Bank Policy Research Working Paper no. 8692.

Fehr, D., R. Hakimov, and D. Kübler. 2015. "The Willingness to Pay—Willingness to Accept Gap: A Failed Replication of Plott and Zeiler." *European Economic Review* 78: 120—128.

Fenoaltea, S. 1976. "Real Value Added and the Measurement of Industrial Production." *Annals of Economic and Social Measurement* 5, no. 1: 111—137.

Fenoaltea, S. 2020. *Reconstructing the Past: The Measurement of Aggregate Product*. Torino: Fondazione Luigi Einaudi.

Ferreira, F. and M. Ravallion. 2011. "Poverty and Inequality: The Global Context." In Nolan, B., W. Slaverda, and T. M. Smeeding (eds.) *Oxford Handbook of Economic Inequality*. Oxford: Oxford University Press.

Filzmoser, P., J. Gussenbauer, and M. Templ. 2016. "Detecting Outliers in Household Consumption Survey Data." Vienna University of Technology.

Finn, A. and V. Ranchhod. 2015. "Genuine Fakes: The Prevalence and Implications of Data Fabrication in a Large South Africa Survey." *The World Bank Economic Review* 31, no. 1: 129 – 157.

Fisher, I. 1922. *The Making of Index Numbers: A Study of Their Varieties, Tests, and Reliability*. Boston: Houghton Mifflin Company.

Fisher, M., J. J. Reimer, and E. R. Carr. 2010. "Who Should be Interviewed in Surveys of Household Income?" *World Development* 38, no. 7: 966-973.

Fleurbaey, M. and F. Maniquet. 2011. *A Theory of Fairness and Social Welfare*. Cambridge: Cambridge University Press.

Foster, J. E. and A. F. Shorrocks. 1988. "Poverty Orderings." *Econometrica* 56, no. 1: 173 – 177.

Foster, J., J. Greer, J., and E .Thorbecke. 1984. "A Class of Decomposable Poverty Measures." *Econometrica* 52, no. 3: 761 – 766.

Foster, J., J. Greer, and E. Thorbecke. 2010. "The Foster – Greer – Thorbecke (FGT) Poverty Measures: 25 Years Later." *Journal of Economic Inequality* 8, no. 4: 491 – 524.

Foster, J., S. Seth, M. Lokshin, and Z. Sajaia. 2013. *A Unified Approach to Measuring Poverty and Inequality: Theory and Practice.* Washington, DC: The World Bank.

Frick, J. R., N. M. Grabka, T. M. Smeeding, and P. Tsakloglou. 2010. "Distributional Effects of Imputed Rents in Five European countries." *Journal of Housing Economics* 19, no 3: 167 – 179.

Friedline, T., R. D. Masa, and G. A. Chowa. 2015. "Transforming Wealth: Using the Inverse Hyperbolic Sine (IHS) and Splines to Predict Youth's Math Achievement." *Social science research*, 49: 264 – 287.

Friedman, M. A. 1957. *Theory of the Consumption Function*. Princeton, N.J.: Princeton University Press.

Friedman, J., K. Beegle, J. De Weerdt, and J. Gibson. 2017. "Decomposing Response Error in Food Consumption Measurement: Implications for Survey Design from a Randomized Survey Experiment in Tanzania." *Food Policy*, 72: 94 – 111.

Frisch, R. 1933. "Editor's Note." *Econometrica* 1, no. 1: 1 – 4.

Gaddis, I. 2016. "Prices for Poverty Analysis in Africa." World Bank Policy Research Working Paper no. 7652.

Gallardo, M. 2018. "Identifying Vulnerability to Poverty: A Critical Survey." *Journal of Economic Surveys* 32, no. 4: 1074 – 1105.

Gambacorta, R., A. Rosolia, and F. Zanichelli. 2022. "Finances of European Households Throughout the Pandemic." In Bandyopadhyay, S. (ed.) Research on Economic Inequality: Poverty, Inequality and Shocks, vol. 29: 247 – 265. Emerald Publishing.

García-Gómez, C., A. Perez, and M. Prieto-Alaiz. 2019. "A Review of Stochastic Dominance Methods for Poverty Analysis." *Journal of Economic Surveys* 33, no. 5: 1437 – 1462.

Gather, U. and C. Becker. 1997. "Outlier Identification and Robust Methods." In Maddala, G. S. and C. R. Rao (eds.) *Handbook of Statistics*, vol. 15: 123 – 142. Amsterdam: North-Holland.

Genoni, M. E. 2012. "Health Shocks and Consumption Smoothing: Evidence from Indonesia." *Economic Development and Cultural Change* 60, no. 3: 475 – 506.

Gentzkow, M. and Shapiro, J. M. 2014. *Code and data for the social sciences: A practitioner's guide*. Chicago, IL: University of Chicago.

Gertler, P. J., E. Rose, and P. Glewwe. 2000. "Health." In Grosh, M. and P. Glewwe (eds.) *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study. Volume One*: 177 – 216. Washington, DC: The World Bank.

Gibson, J. 2007. *A Guide to Using Prices in Poverty Analysis*. Unpublished manuscript.

Gibson, J. 2013. "The Crisis in Food Price Data." *Global Food Security* 2, no. 2: 97 – 103.

Gibson, J. 2016. "Poverty Measurement: We Know Less Than Policy Makers Realize." *Asia and the Pacific Policy Studies* 3, no. 3: 430—442.

Gibson, J., K. Beegle, J. De Weerdt, and J. Friedman. 2015. "What Does Variation in Survey Design Reveal About the Nature of Measurement Errors in Household Consumption?" *Oxford Bulletin of Economics and Statistics* 77, no. 3: 466—474.

Gibson, J. and B. Kim. 2013. "Quality, Quantity, and Nutritional Impacts of Rice Price Changes in Vietnam." *World Development* 43: 329—340.

Gibson, J. and B. Kim. 2015. "Hicksian Separability Does Not Hold Over Space: Implications for the Design of Household Surveys and Price Questionnaires." *Journal of Development Economics* 114: 34—40.

Gibson, J. and B. Kim. 2019. "Quality, Quantity, and Spatial Variation of Price: Back to the Bog." *Journal of Development Economics* 137: 66—77.

Gibson, J., T. Le, and B. Kim. 2017. "Prices, Engel Curves, and Time-Space Deflation: Impacts on Poverty and Inequality in Vietnam." World Bank Policy Research Working Paper no. 8097.

Gibson, J. and A. Romeo. 2017. "Fiscal-Food Policies are Likely Misinformed by Biased Price Elasticities from Household Surveys: Evidence from Melanesia." *Asia and the Pacific Policy Studies* 4, no. 3: 405—416.

Gibson, J. and S. Rozelle. 2002. "How Elastic is Calorie Demand? Parametric, Nonparametric, and Semiparametric Results for Urban Papua New Guinea." *Journal of Development Studies* 38, no. 6: 23—46.

Gibson, J. and S. Rozelle. 2005. "Prices and Unit values in Poverty Measurement and Tax Reform Analysis." *The World Bank Economic Review* 19, no. 1: 69—97.

Gibson, J. and S. Rozelle. 2011. "The Effects of Price on Household Demand for Food and Calories in Poor Countries: Are Our Databases Giving Reliable Estimates?" *Applied Economics* 43, no. 27: 4021—4031.

Gilbert, C. L., L. Christiaensen, and J. Kaminski. 2016. "Price Seasonality in Africa: Measurement and Extent." World Bank Policy Research Working Paper no. 7539.

Glaeser, E. L. 1998. "Should Transfer Payments Be Indexed to Local Price Levels?" *Regional Science and Urban Economics* 28, no. 1: 1—20.

Gluschenko, K. 2006. "Biases in Cross-Space Comparisons Through Cross-Time Price Indexes: The Case of Russia." BOFIT Discussion paper no. 9/2006.

Goyder, J. 2019. *The Silent Minority: Nonrespondents on Sample Surveys*. Routledge.

Grimard, F. 1997. "Household Consumption Smoothing through Ethnic Ties: Evidence from Côte d'Ivoire." *Journal of Development Economics* 53, no. 2: 391—422.

Grosh, M. and P. Glewwe (eds.) 2000. *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 years of the Living Standards Measurement Study*. Washington, DC: World Bank.

Grosh, M. and J. E. Munoz. 1996. "A Manual for Planning and Implementing the Living Standards Measurement Study Survey." LSMS Working Paper no. 126.

Groves, R. M. 2006. "Nonresponse Rates and Nonresponse Bias in Household Surveys." *Public Opinion Quarterly* 70, no. 5: 646—675.

Groves, R. M., R. B. Cialdini, and M. P. Couper. 1992. "Understanding the Decision to Participate in a Survey." *Public Opinion Quarterly* 56, no. 4: 475—495.

Groves, R. M. and M. P. Couper. 1998. *Nonresponse in Household Interview Surveys*. New York: John Wiley & Sons.

Groves, R. M., F. J. Fowler Jr., M. P. Couper, J. M. Lepkowski, E. Singer, and R. Tourangeau. 2009. *Survey Methodology*. New York: John Wiley & Sons.

Grubbs, F. E. 1969. "Procedures for Detecting Outlying Observations in Samples." *Technometrics* 11, no. 1: 1–21.

Haddad, L. J. and S. R. Kanbur. 1990. "How Serious is the Neglect of Intra-Household Inequality?" *The Economic Journal* 100, no. 402: 866–81.

Hagenaars, A. J., K. De Vos, and M. Asghar Zaidi, 1994. "Poverty Statistics in the Late 1980s: Research Based on Micro-data." Statistical Office of the European Communities, Luxembourg.

Hamilton, B. W. 2001. "Using Engel's Law to Estimate CPI bias." *American Economic Review* 91, no. 3: 619–630.

Hampel, F. R. 1974. "The Influence Curve and Its Role in Robust Estimation." *Journal of the American Statistical Association* 69, no. 346: 383–393.

Hampel, F. R., E. M. Ronchetti, P. J. Rousseeuw, and W. A. Stahel. 1986. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons.

Hanemann, M. W. 1991. "Willingness to Pay and Willingness to Accept: How Much Can They Differ?" *The American Economic Review* 81, no. 3: 635–647.

Hart, E. M., P. Barmby, D. LeBauer, F. Michonneau, S. Mount, P. Mulrooney, T. Poisot, K. H. Woo, N. B. Zimmerman, and J. W. Hollister. 2016. "Ten Simple Rules for Digital Data Storage." *PLOS Computational Biology* 12, no. 10.

Haughton, J. and S. R. Khandker. 2009. *Handbook on Poverty + Inequality.* Washington, DC: The World Bank.

Hausman, J. 2003. "Sources of Bias and Solutions to Bias in the Consumer Price Index." *Journal of Economic Perspectives* 17, no. 1: 23–44.

Heckman, J. J. 1979. "Sample Selection Bias as a Specification Error." *Econometrica* 47, 1: 153–161.

Heeringa, S. G., B. T. West, and P. A. Berglund. 2017. *Applied Survey Data Analysis.* Chapman and Hall.

Heijink, R., K. Xu, P. Saksena, and D. Evans. 2011. "Validity and Comparability of Out-of-Ppocket Health Expenditure from Household Surveys: A Review of the Literature and Current Survey Instruments." Geneva: World Health Organization.

Hentschel, J. and P. Lanjouw. 2000. "Household Welfare Measurement and the Pricing of Basic Services." *Journal of International Development* 12, no. 1: 13–27.

Heston, A. and A. O. Nakamura. 2009. "Questions About the Equivalence of Market Rents and User Costs for Owner Occupied Housing." *Journal of Housing Economics* 18: 273–279.

Hlasny, V. 2020. "Nonresponse Bias in Inequality Measurement: Cross-Country Analysis Using Luxembourg Income Study Surveys." *Social Science Quarterly* 101, no. 2: 712–731.

Hlasny, V. and P. Verme. 2018a. "Top Incomes and the Measurement of Inequality in Egypt." *World Bank Economic Review* 32, no. 32: 428–455.

Hlasny, V. and Verme, P. 2018b. "Top Incomes and Inequality Measurement: A Comparative Analysis of Correction Methods Using the EU SILC data." *Econometrics* 6, no. 2: 1-21.

Hlasny, V., L. Ceriani, and P. Verme. 2020. "Bottom Incomes and the Measurement of Poverty and Inequality." GLO Discussion Paper no. 519.

Houthakker, H. S. and S. J. Prais. 1952. "Les Variations de Qualité dans les Budgets de Famille." *Economie Appliquée*, 5, 62.

Huber, P. J. 1981. *Robust Statistics*. New York: John Wiley & Sons.

Huber, P. J. 1996. *Robust Statistical Procedures*. Siam.

Hurd, M. and S. Rohwedder. 2009. "Methodological Innovations in Collecting Spending Data: The HRS Consumption and Activities Mail Survey." *Fiscal Studies* 30, no. 3–4: 435–459.

Hurst, E., Li, G. and B. Pugsley. 2014. "Are Household Surveys Like Tax Forms? Evidence from Income Underreporting of the Self-employed." *Review of Economics and Statistics* 96, no. 1: 19 – 33.

ILO. 2013. *Sources and Methods: Labour Statistics, 2013 Update*. Geneva: ILO.

ILO. 2004. *Consumer Price Index Manual: Theory and Practice*. International Monetary Fund.

Jalan, J. and M. Ravallion. 1999. "Are the Poor Less Well Insured? Evidence on Vulnerability to Income Risk in Rural China." *Journal of Development Economics*, 58, no. 1: 61 – 81.

Jenkins, S. J. 1999. "INEQDECO: Stata module to calculate inequality indices with decomposition by subgroup," Statistical Software Components S366002, Boston College Department of Economics, revised February 15, 2021.

Jehle, G. A. and P. J. Reny. 2011. *Advanced Microeconomic Theory*. Essex: Pearson Education Limited.

Johnson, D. S., T. M. Smeeding, and B. B. Torrey. 2005. "Economic Inequality through the Prisms of Income and Consumption." *Monthly Labor Review*, 128: 11.

Jolliffe, D. 2001. "Measuring Absolute and Relative Poverty: The Sensitivity of Estimated Household Consumption to Survey Design." *Journal of Economic and Social Measurement* 27, no. 1 and 2: 1 – 23.

Jolliffe, D. 2006. "Poverty, Prices, and Place: How Sensitive Is the Spatial Distribution of Poverty to Cost-of-living Adjustments?" *Economic Inquiry* 44, no. 2: 296 – 310.

Jolliffe, D., G. Datt, and M. Sharma. 2004. "Robust Poverty and Inequality Measurement in Egypt: Correcting for Spatial-Price Variation and Sample Design Effects." *Review of Development Economics* 8: 557 – 572.

Jolliffe, D. and U. Serajuddin. 2015. "Estimating Poverty with Panel Data, Comparably: An Example from Jordan." World Bank Policy Research Working Paper no. 7373.

Kaara, J. and Ramasawmy, S. 2008. "Food Data Collected Using Acquisition and Consumption Approaches with a Seven-Day Recall Method in Kenya's KIHBS 2005/2006. Deriving Food Security Information from National Household Budget Survey: Experiences, Achievements, Challenges." Roma: Food and Agriculture Organization.

Kakwani, N. and R. J. Hill. 2002. "Economic Theory of Spatial Cost of Living Indices with Application to Thailand." *Journal of Public Economics* 86, no. 1: 71 – 97.

Kalton G. and D. Kasprzyk. 1986. "The Treatment of Missing Survey Data." *Survey Methodology* 12: 1 – 16.

Kalton, G. and I. Flores-Cervantes. 2003. "Weighting Methods." *Journal of Official Statistics* 19, no. 2: 81-97.

Kaminski, J., L. Christiaensen, and C. L. Gilbert. 2016. "Seasonality in Local Food Markets and Consumption: Evidence from Tanzania." *Oxford Economic Papers* 68, no. 3: 736 – 757.

Kanbur, R. 2018. "Intrahousehold Inequality and Overall Inequality," in Ocampo, J. A and J. E. Stiglitz (eds.) *The Welfare State Revisited*: 147 – 164. New York Chichester, West Sussex: Columbia University Press.

Khandker, S. R. 2009. "Poverty and Income Seasonality in Bangladesh." World Bank Policy Research Working Paper no. 4923.

Kitzes, J., D. Turek, and F. Deniz (eds.) 2017. *The Practice of Reproducible Research: Case Studies and Lessons from the Data-Intensive Sciences.* University of California Press.

Kolenikov, S. 2016. "Post-stratification or non-response adjustment." *Survey Practice* 9, no. 3.

Korinek, A., J. A. Mistiaen, and M. Ravallion. 2006. "Survey Nonresponse and the Distribution of Income." *Journal of Economic Inequality* 4: 33 – 55.

Korinek, A., J. A. Mistiaen, and M. Ravallion. 2007. "An Econometric Method of Correcting for Unit Nonresponse Bias in Surveys." *Journal of Econometrics*, 136: 213 – 35.

Lambert, P. J. 2001. *The Distribution and Redistribution of Income*. Manchester University Press.

Lanjouw, J. O. and P. Lanjouw. 2001. "How To Compare Apples and Oranges: Poverty Measurement Based on Different Definitions of Consumption." *The Review of Income and Wealth* 47, no. 1: 25 – 42.

Lanjouw, J. O. and P. Lanjouw. 1997. "Poverty Comparisons with Noncompatible Data: Theory and Illustrations." World Bank Policy Research Working Paper no. 1709.

Lanjouw, P. 2009. "Constructing a Consumption Aggregate for the Purpose of Welfare Analysis: Principles, Issues and Recommendations Arising from the Case of Brazil." Unpublished manuscript.

Lanjouw, P. and Ravallion, M. 1995. "Poverty and household size." *The Economic Journal* 105, no. 433: 1415 – 1434.

Lewbel, A. 2006. "Engel curves." *The New Palgrave Dictionary of Economics.*

Li, C. and J. Gibson, J. 2014. "Spatial Price Differences and Inequality in the People's Republic of China: Housing Market Evidence." *Asian Development Review*, 31: 1.

Little, R. J. and D. B. Rubin. 2019. *Statistical Analysis with Missing Data.* New York: John Wiley & Sons.

Liu, H., T. I. Wahl, J. L. Seale Jr., and J. Bai. 2015. "Household Composition, Income, and Food-Away-From-Home Expenditure in Urban China." *Food Policy*, 51: 97 – 103.

Lohr, S. L. 2009. *Sampling: Design and Analysis.* Nelson Education.

Lokshin, M. 2018. "Data Quality in Research: What if We're Watering the Garden while the House Is on Fire?", World Bank blog.

Long, J. S. 2009. *The Workflow of Data Analysis Using Stata*. College Station, TX: Stata Press.

Lu, C., B. Chin, G. Li, and C. J. Murray. 2009. "Limitations of Methods for Measuring Out-of-Pocket and Catastrophic Private Health Expenditures." *Bulletin World Health Organization* 87, no. 3: 238 – 44, 244A – 244D.

Luiten, A., J. Hox, and E. de Leeuw. 2020. "Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys," *Journal of Official Statistics* 36, no 3: 469 – 487.

MacLeod, W. B. and M. Urquiola. 2018. "Is Education Consumption Or Investment? Implications For The Effect Of School Competition." NBER Working Paper no. 25117.

Malpezzi, S. 2000. "Housing." In Grosh, M. and P. Glewwe (eds.) *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standards Measurement Study.* Volume 1: 293 – 314. Washington, DC: World Bank.

Martirosova, D. 2008. "Food Data Collected Using Acquisition and Consumption Approaches with Daily Diaries in Armenia's ILCS 2004." In Sibrian, S. (ed.) *Deriving Food Security Information from National Household Budget Surveys: Experiences, Achievements, Challenges*: 59 – 68. Rome: Food and Agriculture Organization.

Mas-Colell, A., M. D. Whinston, and J. R. Green. 1995. *Microeconomic Theory*. New York: Oxford University Press.

McKay, A. 2000. "Should the Survey Measure Total Household Income?" In Grosh, M. and P. Glewwe (eds.) *Designing Household Survey Questionnaires for Developing Countries. Lessons from 15 Years of the Living Standard Measurement Study*. Volume 2: 83 – 104. Washington, DC: World Bank.

McKelvey, C. 2011. "Price, Unit Value, and Quality Demanded." *Journal of Development Economics* 95, no. 2: 157 – 169.

McKnight, A. and M. Rucci. 2020. "The Financial Resilience of Households: 22 Country Study with New Estimates, Breakdowns by Household Characteristics and a Review of Policy Options." LSE Centre for Analysis of Social Exclusion, CASE Papers/219.

Mercier, M. and P. Verwimp. 2017. "Are We Counting All the Poor?: Accounting for the Intra-Household Allocation of Consumption in Burundi." *Journal of Demographic Economics* 83, no. 3: 307 – 327.

Meyer, B. D. and J. Sullivan. 2008. "Changes in the Consumption, Income, and Well-Being of Single Mother Headed Families." *American Economic Review* 98, no. 5: 2221 – 2241.

Meyer, B. D. and J. Sullivan. 2011. "Further Results on Measuring the Well-being of the Poor Using Income and Consumption." *Canadian Journal of Economics/Revue canadienne d'économique* 44, no. 1: 52 – 87.

Meyer, B. D. and J. Sullivan. 2012. "Identifying the Disadvantaged: Official Poverty, Consumption Poverty, and the New Supplemental Poverty Measure." *Journal of Economic Perspectives* 26, no. 3: 1 – 27.

Meyer, B. D., W. K. Mok, and J. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives* 29, no. 4: 199 – 226.

Milanovic, B. 2008. "Qat Expenditures in Yemen and Djibouti: an Empirical Analysis." *Journal of African Economies* 17, no. 5: 661 – 687.

Miller, L. L. and E. Aharoni. 2015. "Understanding Low Survey Response Rates Among Young US Military Personnel." RAND Corporation.

Minot, N. 2014. "Food Price Volatility in Sub-Saharan Africa: Has It Really Increased?" *Food Policy* 45: 45 – 56.

Moore, J., L. Stinson, and E. Welniak. 2000. "Income Measurement Error in Surveys: A Review." *Journal of Official Statistics* 16, no. 4: 331 – 362

Morduch, J. 1995. "Income Smoothing and Consumption Smoothing." *Journal of Economic Perspectives* 9, no. 3: 103 – 114.

Morduch, J. 1998. "Poverty, Economic Growth, and Average Exit Time." *Economics Letters* 59, no. 3: 385 – 390.

Morelli, S. and E. Munoz. 2020. "Unit Nonresponse Bias in the Current Population Survey." Unpublished manuscript.

Morelli, S. and E. Munoz. 2021. "kmr: A Command to Correct Survey Weights for Unit Nonresponse using Group's Response Rates." *Stata Journal* 2, no. 1: 206 – 219.

Moretti, E. 2010. "Poverty, Inequality and Cost-of-living Differences." University of Kentucky Center for Poverty Research Discussion Paper No. DP2010-07.

Muller, C. 2008. "The Measurement of Poverty with Geographical and Intertemporal Price Dispersion: Evidence from Rwanda." *Review of Income and Wealth* 54, no. 1: 27 – 49.

Musumeci, I. 2016. "Farm-Gate Price Monitoring in Selected Impact Countries: Lessons Learned from the 2013 – 2014 Pilot." World Food Programme.

Nago, E. S., C. K. Lachat, L. Huybregts, D. Roberfroid, R. A. Dossa, and P. W. Kolsteren. 2010. "Food, Energy and Macronutrient Contribution of Out-of-home Foods in School-going Adolescents in Cotonou, Benin." *British Journal of Nutrition* 103, vol. 2: 281 – 288.

National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. Washington, DC: The National Academies Press.

National Research Council. 1995. *Measuring Poverty: A New Approach.* Washington, DC: The National Academies Press.

National Research Council. 2002. *At What Price?: Conceptualizing and Measuring Cost-of-Living and Price Indexes.* Washington, DC: The National Academies Press.

Nicoletti, C. 2010. "Poverty Analysis with Missing Data: Alternative Estimators Compared." *Empirical Economics* 38, vol. 1: 1 – 22.

Nordhaus, W. D. and J. Tobin. 1973. "Is Growth Obsolete?" In M. Moss (ed.) *The Measurement of Economic and Social Performance*: 509 – 564 NBER Books.

Norris, S. and K. Pendakur. 2013. "Imputing Rent in Consumption Measures, with an Application to Consumption Poverty in Canada, 1997—2009." *Canadian Journal of Economics/Revue canadienne d'économique* 46, no. 4: 1537—1570.

O'Higgins, M. and S. P. Jenkins. 1990. "Poverty in the EC: Estimates for 1975, 1980 and 1985." In Teekens, R. and B. M. S. Van Praag (eds.) *Analysing Poverty in the European Community*: 187—212. Eurostat News Special Edition, Luxembourg: Publications Office for Official Publications of the European Communities.

OECD. 2013. *OECD Framework for Statistics on the Distribution of Household Income, Consumption and Wealth.* Paris: OECD Publishing.

Omoniyi, A., W. D. Vundru, and T. Kilic. 2021. "Examining the Utility of Market Price Surveys in Welfare Analysis: Evidence from Malawi." Unpublished manuscript.

Osberg, L. and K. Xu. 2001. "International Comparison of Poverty Intensity: Index Decomposition and Bootstrap Inference," *Journal of Human Resources* 35, no. 1: 51—81.

Oseni, G., J. Durazo, and K. McGee. 2017. *The Use of Non-standard Units for the Collection of Food Quantity: A Guidebook for Improving the Measurement of Food Consumption and Agricultural Production in Living Standards Surveys.* Washington, DC: World Bank.

Oseni, G., F. Huebler, and K. McGee, Amankwah, A., Legault, E. and Rakotonarivo, A. 2018. *Measuring Household Expenditure on Education: a New Guidebook on Measurement*. Washington, DC: World Bank.

Oseni, G., A. Palacios-Lopez, H. K. Mugera, and J. Durazo. 2021. *Capturing What Matters, Essential Guidelines for Designing Household Surveys.* Washington, DC: World Bank.

Oulton, N. 2012. "How to Measure Living Standards and Productivity." *Review of Income and Wealth* 58, no. 3: 424—456.

Palaniswamy, N. and T. Vishwanath. 2019. "Do household surveys face a non-response crisis? Not necessarily, according to recent experience in Jordan." The World Bank Data Blog. https://blogs.worldbank.org/opendata/do-household-surveys-face-non-response-crisis-not-necessarily-according-recent-experience-jordan

Pape, U. J. and L. Parisotto. 2019. "Estimating Poverty in a Fragile Context. The High Frequency Survey in South Sudan," World Bank Policy Research Working Paper no. 8722.

Paxson, C. 1993. "Consumption and Income Seasonality in Thailand." *Journal of Political Economy* 101, no. 1: 39—72.

Pittau, M. G., R. Zelli, and R. Massari. 2011. "Do Spatial Price Indices Reshuffle the Italian Income Distribution?" *Modern Economy* 2, no. 3: 259—265.

Platek, R. 1977. "Some Factors Affecting Non-response." *Survey Methodology* 3: 191—214.

Plewes, T. J. and R. Tourangeau. (eds.) 2013. *Nonresponse in Social Science Surveys: A Research Agenda*. Washington, DC: National Academies Press.

Ponthieux, S. 2017. "Intra-household Pooling and Sharing of Resources: A Tentative 'Modified' Equivalised Income." In Atkinson, A. B., A. C. Guio, and E. Marlier (eds.) *Monitoring Social Inclusion in Europe*. Luxembourg: Publications Office of the European Union, 175—188.

Pradhan, M. P. 2009. "Welfare Analysis with a Proxy Consumption Measure: Evidence from a Repeated Experiment in Indonesia." *Fiscal Studies* 30, no. 3/4: 391—417.

Prais, S. J. and H. S. Houthakker. 1955. *The Analysis of Family Budgets with an Application to two British Surveys conducted in 1937—39 and their Detailed Results*. New York: Cambridge University Press.

Pyatt, G., C. N. Chen, and J. Fei. 1980. "The Distribution of Income by Factor Components." *The Quarterly Journal of Economics* 95, no. 3: 451—473.

Ravallion, M. 1988. "Expected Povertv Under Risk-induced Welfare Variabilitv." *Economic Journal* 98, no. 393: 1171-1182.

Ravallion, M. 1994. *Poverty Comparisons*. London: Routledge.

Ravallion, M. 1998. "Poverty Lines in Theory and Practice." World Bank LSMS Working Paper no. 133.

Ravallion M. 2008. "Poverty Lines." In Durlauf, S. N. and L. E. Blume (eds.) *The New Palgrave Dictionary of Economics.* London: Palgrave Macmillan.

Ravallion, M. 2010. "The Troubling Tradeoffs of the Human Development Index." World Bank Policy Research Working Paper no. 5484.

Ravallion, M. 2016. *The Economics of Poverty. History, Measurement and Policy*. New York and Oxford: Oxford University Press.

Ravallion, M. 2020. "On Measuring Global Poverty." *Annual Review of Economics* 12: 167 – 188.

Ravallion, M. 2021. *Missing Top Income Recipients*. NBER Working Paper no. 28890.

Ravallion, M. and B. Bidani. 1994. "How Robust is a Poverty Profile?" *The World Bank Economic Review* 8, no. 1: 75 – 102.

Ravallion, M. and M. Huppi. 1991. "Measuring Changes in Poverty: a Methodological Case Study of Indonesia During an Adjustment Period." *The World Bank Economic Review* 5, no. 1: 57 – 82.

Ray, R. 2017. "The Role of Prices in Welfare Comparisons: Methodological Developments and a Selective Survey of the Empirical Literature." *Economic Record* 93, no. 301: 314 – 332.

Ray, R. 2018. *Household Behaviour, Prices, and Welfare: A Collection of Essays Including Selected Empirical Studies.* Singapore: Springer.

Roser, M. and E. Ortiz-Ospina. 2017. "Global Extreme Poverty." Published online at OurWorldInData.org. Accessed on 07/2019. Retrieved from https://ourworldindata.org/extreme-poverty.

Rousseeuw, P. J. and C. Croux. 1993. "Alternatives to the Median Absolute Deviation." *Journal of the American Statistical Association* 88, no. 424: 1273 – 1283.

Rousseeuw, P. and M. Hubert. 2017. "Anomaly Detection by Robust Statistics." *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery* 8, no. 2: 1236.

Rubin, D. B. 1976. "Inference and Missing Data." *Biometrika*, 63: 581 – 592.

Rubin, D. B. 1987. *Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons.

Santaeulàlia-Llopis, R. and Y. Zheng. 2017. "Why is Food Consumption Inequality Underestimated? A Story of Vices and Children." Barcelona Graduate School of Economics Working Paper 969.

Särndal, C. E. and S. Lundström. 2005. *Estimation in Surveys with Nonresponse*. New York: John Wiley & Sons..

Saunders, P. and P. Siminski. 2005. "Home Ownership and Inequality: Imputed Rent and Income Distribution in Australia." *Economic Papers: A Journal of Applied Economics and Policy* 24, no. 4: 346 – 367.

Schaeffer, N. C., J. Dykema, and D. Maynard. 2010. "Interviewers and Interviewing." In Marsden, P. V. and J. D. Wright (eds.) *Handbook of Survey Research. Second Edition*. Bingley, UK: Emerald Group Publishing, 437 – 470.

Schluter, C. 2012. "On the Problem of Inference for Inequality Measures for Heavy-tailed Distributions." *The Econometrics Journal* 15, no. 1: 125 – 153.

Schluter, C. and K. J. van Garderen. 2009. "Edgeworth Expansions and Normalizing Transforms for Inequality Measures." *Journal of Econometrics* 150, no. 1: 16 – 29.

Schündeln, M. 2018. "Multiple Visits and Data Quality in Household Surveys." *Oxford Bulletin of Economics and Statistics* 80, no. 2: 380 – 405.

Scott, C. and B. Amenuvegbe. 1991. "Recall Loss and Recall Duration: an Experimental Study in Ghana." *Inter-Stat* 4, no. 1, 31 – 55.

Sen, A. 1985. *Commodities and Capabilities*. Amsterdam: North-Holland.

Sen, A. 1987. *The Standard of Living*. Cambridge: Cambridge University Press.

Sen, A. 1993. "Capability and Well-Being." In M. Nussbaum, Sen, A. (eds.) *The Quality of Life*. Oxford: Clarendon Press.

Shorrocks, A. F. 1995. "Revisiting the Sen Poverty Index." *Econometrica* 63, no. 5: 1225—1230.

Shorrocks, A. F. 1983. "Ranking Income Distributions." *Economica* 50, no. 197: 3—17.

Shorrocks, A. F. 1980. "The Class of Additively Decomposable Inequality Measures." *Econometrica* 48, no. 3: 613—625.

Singer, E. and C. Ye. 2013. "The Use and Effects of Incentives in Surveys." *The Annals of the American Academy of Political and Social Science* 645, no. 1: 112—141.

Skoufias, E. and A. R. Quisumbing. 2005. "Consumption Insurance and Vulnerability to Poverty: A Synthesis of the Evidence from Bangladesh, Ethiopia, Mali, Mexico and Russia." *The European Journal of Development Research* 17, no. 1: 24—58.

Slesnick, D. T. 2001. *Consumption and Social Welfare: Living Standards and their Distribution in the United States.* Cambridge: Cambridge University Press.

Smeeding, T. M. and D. H. Weinberg. 2001. "Toward a Uniform Definition of Household Income," *Review of Income and Wealth* 47: 1—24.

Smith, L. C. 2015. "The Great Indian Calorie Debate: Explaining Rising Undernourishment During India's Rapid Economic Growth." *Food Policy* 50: 53—67.

Smith, L. C., H. Alderman, and D. Aduayom. 2006. "Food Insecurity in Sub-Saharan Africa: New Estimates from Household Expenditure Surveys." *Research Report* 146, International Food Policy Research Institute.

Smith, L. C., O. Dupriez, and N. Troubat. 2014. "Assessment of the Reliability and Relevance of the Food Data Collected in National Household Consumption and Expenditure Surveys." International Household Survey Network (IHSN) Working Paper 008.

Snedecor, G. W. and W. G. Cochran. 1989. *Statistical Methods*. Ames: Iowa State University Press.

Solon, G., S. J. Haider, and J. M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50, no. 2: 301—316.

Stone, R. 1954. "Linear Expenditure Systems and Demand Analysis: an Application to the Pattern of British Demand." *The Economic Journal* 64, no. 255: 511—527.

Stiglitz, J. E., A. Sen, and J. P. Fitoussi. 2009. *Report by the Commission on the Measurement of Economic Performance and Social Progress*.

Summers, R. 1973. "International Price Comparisons Based Upon Incomplete Data." *Review of Income and Wealth* 19, no. 1: 1—16.

Szulc, A. 2006. "Poverty in Poland during the 1990s: Are the Results Robust?" *Review of Income and Wealth* 52: 423—48.

Törmälehto, V. M. and H. Sauli. 2013. "The Distributional Impact of Imputed Rent in EU-SILC 2007—2010." Eurostat Methodologies and Working Papers 2013 Edition.

Tourangeau, R., L. J. Rips, and K. Rasinski. 2000. *The Psychology of Survey Response.* Cambridge: Cambridge University Press.

Tourangeau, R., R. M. Groves, and C. D. Redline. 2010. "Sensitive Topics and Reluctant Respondents: Demonstrating a Link Between Nonresponse Bias and Measurement Error." *Public Opinion Quarterly* 74, no. 3, 413—432.

Triplett, J. E. 2001. "Should the Cost-of-living Index Provide the Conceptual Framework for a Consumer Price Index?" *The Economic Journal* 111, no. 472: 311—334.

Troubat, N. and K. Grünberger. 2017. "Impact of Survey Design in the Estimation of Habitual Food Consumption. The Case of the 2007/08 Socio Economic Survey of Mongolia Applied to Urban Households." *Food Policy* 72: 132 – 145.

Tunçel, T. and J. K. Hammitt. 2014. "A New Meta-analysis on the WTP/WTA Disparity." *Journal of Environmental Economics and Management* 68, no. 1, 175 – 187.

United Nations Statistics Division. 2005. *Handbook on Poverty Statistics: Concepts, Methods and Policy Use*. New York: United Nations.

United Nations Statistics Division. 2005. *Designing Household Survey Samples: Practical Guidelines.* New York: United Nations.

United Nations. 2018. *Classification of Individual Consumption According to Purpose (COICOP) 2018.* New York: United Nations.

UNECE. 2011. *Canberra Group Handbook on Household Income Statistics, Second Edition*. Geneva: United Nations Economic Commission for Europe.

UNECE. 2017. *Guide on Poverty Measurement.* New York: United Nations.

Valliant, R. and J. A. Dever. 2018. *Survey Weights: A Step-by-step Guide to Calculation*. College Station, TX: Stata Press.

Valliant, R., J. A. Dever, and F. Kreuter. 2013. *Practical Tools for Designing and Weighting Survey Samples.* New York: Springer.

Van Kerm, P. 2001. "INEQUAL7: Stata Module to Compute Measures of Inequality." Statistical Software Components S416401, Boston College Department of Economics.

Van Kerm, P. 2007. "Extreme Incomes and the Estimation of Poverty and Inequality Indicators from EU-SILC." IRISS Working Paper Series no. 2007-01.

Van Praag, B. 1971. "The Welfare Function of Income in Belgium: An Empirical Investigation." *European Economic Review* 2, no. 3: 337 – 369.

Van Praag, B. M. and M. F. Warnaar. 1997. "The Cost of Children and the Use of Demographic Variables." In *Consumer Demand. Handbook of Population and Family Economics*, Volume 1A: 241 – 273. Amsterdam, New York and Oxford: North-Holland.

Van Veelen, M. 2002. "An Impossibility Theorem Concerning Multilateral International Comparisons of Volume," *Econometrica* 70: 369 – 375.

Van Veelen, M. 2009. "The Apples and Oranges Theorem for Price Indices," *Economics Letters* 103: 12 – 14.

Van Veelen, M. and R. van der Weide. 2008. "A Note on Different Approaches to Index Number Theory," *American Economic Review* 98: 722 – 1730.

Varian, H. R. 1992. *Microeconomic Analysis.* New York-London: Norton.

Varian, H. R. 1993. "What Use is Economic Theory?" Michigan-Center for Research on Economic and Social Theory Working Paper no. 93-14.

Varian, H. R. 2010. *Intermediate Microeconomics. A Modern Approach*. New York-London: Norton.

Vecchi, G. 2017. *Measuring Wellbeing*. New York: Oxford University Press.

Verme, P. and A. Araar (eds.) 2017. *The Quest for Subsidy Reforms in the Middle East and North Africa Region: A Microsimulation Approach to Policy Making*. Cham, Switzerland: Springer.

Wagstaff, A. 2007. "The Economic Consequences of Health Shocks: Evidence from Vietnam." *Journal of Health Economics* 26, no. 1: 82 – 100.

Wagstaff, A. 2019. "Measuring Catastrophic Medical Expenditures: Reflections on Three Issues." *Health Economics* 28, no. 6: 765 – 781.

Wagstaff, A. and E. van Doorslaer. 2003. "Catastrophe and Impoverishment in Paying for Health Care: with Applications to Vietnam 1993 – 1998." *Health Economics* 12, no. 11: 921 – 933.

Wagstaff, A., P. H. V. Eozenou, and M. F. Smitz. 2019. "Out-of-Pocket Expenditures on Health: A Global Stocktake." World Bank Policy Research Working Paper 8808.

Waid, J. L., J. R. Bogard, S. H. Thilsted, and S. Gabrysch. 2017. "Estimates of Average Energy Requirements in Bangladesh: Adult Male Equivalent Values for Use in Analyzing Household Consumption and Expenditure Surveys." Data Brief no. 14, 101—106.

Watts, H. W. 1968. "An Economic Definition of Poverty," in Moynihan D. P. (ed.) *On Understanding Poverty*, New York: Basic Books.

Weisell, R. and M. C. Dop. 2012. "The Adult Male Equivalent Concept and its Application to Household Consumption and Expenditures Surveys (HCES)." *Food and Nutrition Bulletin* 33, no. 3 (supplement): S157—S162.

WHO Expert Committee on Drug Dependence, and World Health Organization. 2003. *WHO Expert Committee on Drug Dependence: Thirty-third Report* (915). Geneva: World Health Organization.

Wilson, G., J. Bryan, K. Cranston, J. Kitzes, L. Nederbragt, and T. K. Teal. 2017. "Good Enough Practices in Scientific Computing." *PLoS Computational Biology* 13, no. 6.

Wooldridge, J. M. 2012. *Introductory Econometrics: A Modern Approach*. South-Western College Pub.

World Bank. 2012. *Responding to Higher and More Volatile World Food Prices*. Washington, DC: World Bank.

World Bank. 2014. *Iraq—The Unfulfilled Promise of Oil and Growth: Poverty, Inclusion and Welfare in Iraq, 2007—2012*. Washington, DC: World Bank.

World Bank. 2015. *A Measured Approach to Ending Poverty and Boosting Shared Prosperity: Concepts, Data, and the Twin Goals.* Policy Research Report. Washington, DC: World Bank.

World Bank. 2017. *Monitoring Global Poverty: Report of the Commission on Global Poverty*. Washington, DC: World Bank.

World Bank. 2018. *Poverty and Shared Prosperity 2018: Piecing Together the Poverty Puzzle*. Washington, DC: World Bank.

World Bank. 2020. *DIME Analytics Data Handbook.* Washington, DC: World Bank.

Xu, K., F. Ravndal, D. B. Evans, and G. Carrin. 2009. "Assessing the Reliability of Household Expenditure Data: Results of the World Health Survey." *Health Policy* 91, no. 3: 297—305.

Yates, J. 1994. "Imputed Rent and Income Distribution." *Review of Income and Wealth* 40, no. 1: 43—66.

Yeo, I. K. and R. A. Johnson. 2000. "A New Family of Power Transformations to Improve Normality or Symmetry." *Biometrika* 87, no. 4: 954—959.

Zaidi, S. 2007. "Constructing the Consumption Aggregate. Part I: Aggregation Issues." Unpublished teaching material. The World Bank: Washington, DC.

Zheng, B. 1993. "An Axiomatic Characterization of the Watts Poverty Index." *Economics Letters* 42, no. 1: 81—86.

Zheng, B. 1997. "Aggregate Poverty Measures." *Journal of Economic Surveys* 11, no. 2: 123—162.

Index

H

I

K

L

Q

R

S

T

U

V

W

Z