Mind the Concept-Measurement Gap: the Statistical Norms Driving International Economic Mismeasurement


Daniel Mügge and Lukas Linsi
University of Amsterdam and University of Groningen

January 2020
Mind the Concept- Measurement Gap: 
the Statistical Norms Driving International Economic Mismeasurement

Daniel Mügge  
University of Amsterdam  
d.k.muegge@uva.nl

Lukas Linsi  
University of Groningen  
l.a.linsi@rug.nl

Last update: 20 December 2019

This is a preliminary draft. Please do not cite or distribute without permission of the authors.

ABSTRACT. The transnationalization and digitization of economic activity has undermined the quality of international economic data, which is still based on national territories and material production. Why do we not witness more vigorous efforts to bring statistical standards in line with present-day economic realities, or admissions that precision in economic data has become increasingly illusory? The answer, we argue, lies in the norms underpinning global statistical practice. Users expect statistics to draw on unambiguous sources, to allow for comparison over time and across countries, and they prize coherence—both internally and with holistic macroeconomic models. Yet as we show, the ambition of the transnational statistical community to meet these norms has in fact undermined the ability of economic data to represent economic life more faithfully. We base our findings on interviews with two dozen leading statisticians at international economic organizations, archival research at the International Monetary Fund, and a thorough review of debates among statistical experts.

Word count: 10,942

Keywords. Economic measurement; international organizations; constructivism; balance of payments; norms; statistics profession.
“There is a growing appreciation that the statistical compilation tools and accounting frameworks designed and developed over the last 60 years ... may reflect a world that no longer exists”

Nadim Ahmad, Head of Trade and Competitiveness Statistics, OECD

(Ahmad 2018, p. 1)

Introduction

Statistics are the bedrock of economic policy-making and debate. They allow computation, comparison, historical analysis, and future forecasting. Without such data, “the economy” would remain an intractable abstraction for policy-makers, citizens, and analysts alike.

Yet, the quality of ubiquitous economic data is much worse than their users typically acknowledge (Morgenstern, 1963; International Monetary Fund, 1987, 1992; UNECE, Eurostat and OECD, 2011; Damgaard and Elkjaer, 2017). If economic data fail to capture what they purportedly claim to represent, public deliberation, economic policy, and academic analysis drawing on them all suffer.

Statistical quality has deteriorated because of a widening gap between the concepts international economic data claim to capture and the measurements that find their way into official databases—a phenomenon we call the concept-measurement gap. Indicators had been devised for economic structures clustered in national territories and focused on material production—the industrial economies that we associate with the decades following...
World War Two. Today, these structures are transnationally integrated and intangible production and assets—services, derivatives, knowledge, licenses, and so on—are central. But while the transnationalization and digitization of economic activity has undermined the conceptual validity of key economic indicators (Lipsey, no date; Ahmad, 2018), our statistical concepts have hardly changed. This is true for many macroeconomic figures, yet it particularly affects Balance of Payment (BOP) statistics, which measure cross-border flows of goods and capital, collected following the Balance of Payments Manual (BPM) issued by the International Monetary Fund (IMF).

Statisticians who craft the standards for BOP statistics are keenly aware of the problems an increasingly transnational and intangible economy poses (Bloch and Fall 2015; Moulton and van de Ven 2018; UNECE et al. 2011; Ahmad 2018). Yet their attempts to address the concept-measurement gap have thus far been remarkably ineffective. A priori, we might expect statisticians to respond in two possible ways. They could overhaul statistical standards to match the new economic structures. Or they could incorporate ambiguity in their published statistics, for example by using uncertainty margins, or by simply admitting that we lack meaningful figures. But we observe neither. The production of data continues largely unchanged, leaving most data users with the erroneous impression of high quality figures. Why, we ask, is the widening concept-measurement gap neither narrowed by reforming standards nor reflected in the data itself? What explains the skewed statistical representations that surround us and guide economic debates and policy?

We argue that the stickiness of statistical standards in the face of profound structural changes stems from the norms that underpin macroeconomic statistics as a field of transnational knowledge production. Our analysis highlights four statistical norms that together create a strong conservative bias in international statistical standards. We call them comparability (the desire to compare statistics across countries), continuity (the ambition to build time-series datasets), certitude (the predilection for reliably quantifiable data), and coherence (the aspiration to integrate separate statistical domains into one overarching representation of “the economy”). Considered in isolation, these norms seem commonsensical. But ironically, adhering to them damages the economic numbers that populate our databases, politics, and news.

Our argument builds on a growing scrutiny in International Relations (IR) of the production and use of quantitative information in global politics—for example in the form of rankings or indicators (Davis et al., 2012; Broome and Quirk, 2015; Cooley and Snyder, 2015; Kelley and
From this literature we take the question why indicators are produced the way they are, and specifically, in our case, why Balance of Payments statistics continue to suffer from the concept-measurement gap we identify. At the same time, our emphasis on the norms underlying statistics production harks back to the sociology of professional norms (Fourcade, 2006; Halliday and Carruthers, 2009; Block-Lieb and Halliday, 2017; Kentikelenis and Seabrooke, 2017) and sociologically informed analyses of international institutions more generally (Barnett and Finnemore, 2004; Babb, 2007; Chorev and Babb, 2009; Kentikelenis and Seabrooke, 2017; Murdoch et al., 2018).

Our empirical investigation centers on the evolution of the IMF’s authoritative Balance of Payments Manuals and the key economic indicators defined therein, in particular for trade, foreign direct investment, and portfolio capital flows. We draw on a range of sources to show how deeply ingrained norms skew the production of macroeconomic statistics: specialized reports from statistical agencies and international organizations that produce statistics or oversee standards reveal the extent to which the transnational and intangible economy has dented statistical quality; two dozen interviews with leading statisticians in Paris, Frankfurt, The Hague, London, Geneva, New York, and Washington offer insights into the concerns, trade-offs, and norms as experienced by central figures in international economic statistics; and documents from the IMF archives in Washington allow us to trace these norms backwards through time, at times to the very beginning of systematic BOP statistics.

In the remainder of this article, we first situate our research within broader social science understandings of economic statistics. We then detail how the rise of the transnational and intangible economy has widened the concept-measurement gap to the point where official statistics grossly misrepresent economic relationships and dynamics. Finally, we show how the norms of comparability, continuity, certitude, and coherence structure the statistical field and explain how they help to reproduce inadequate statistical standards.
Macroeconomic statistics in an IR perspective

Macroeconomic statistics have been an international success story. The global spread of GDP as the universal metric to gauge economic prowess has been thoroughly documented (Fioramonti, 2013; Lepenies, 2013; Philipsen, 2015; Masood, 2016). But international organizations such as the United Nations, the International Labour Organization, the World Bank and the International Monetary Fund have promulgated a much wider set of economic statistics (Ward, 2004), including for example poverty measures (Clegg, 2010), government finance statistics, balance of payments statistics, and internationally harmonized unemployment statistics (author 2019). More recently, international organizations as well as NGOs have proactively crafted new indicators and rankings to nudge, name and shame governments towards different policies (Broome and Quirk, 2015; Cooley and Snyder, 2015; Kelley and Simmons, 2015, 2019; Broome, Homolar and Kranke, 2018; Honig and Weaver, 2019).

Economic concepts such as unemployment (Salais, Baverez and Reynaud, 1986; Baxandall, 2004), growth (Schmelzer, 2016; Pilling, 2018), inflation (Mackie and Schultze, 2002; Stapleford, 2009) or debt (Bloch and Fall, 2015) – just like many non-economic ones – defy straightforward definition and measurement, however. Depending on which measurement approach we ultimately use, statistics therefore highlight or obscure specific dimensions of the phenomenon in question and thereby direct attention one way or the other. Indeed, through delineating what (literally) counts as growth, inflation or unemployment, and what does not, these standards delineate “the economy” as a governance object itself (Allan, 2017). This political baggage of macroeconomic indicators raises the question of who decides what is quantified.
Over the decades, a tightly-knit transnational epistemic community has emerged that dominates global standard setting for macroeconomic statistics (Ward, 2004). Central hubs include the Statistics and Data Directorate at the OECD, the economic statistics branch at the United Nations Statistics Division (UNSD), Eurostat as a focal point for European statistical expertise and, for finance statistics in particular, the statistics department of the IMF. In reforms of the System of National Accounts (SNA)—the how-to guide for building GDP statistics—all these organizations, together with the World Bank, cooperate in the Intersecretariat Working Group on National Accounts (ISWGNA), the de facto global headquarters of GDP measurement. Comparable and linked processes exist for other macroeconomic statistics, for example through the Intersecretariat Working Group on Price Statistics. Overlapping and rotating membership and leadership of these initiatives has generated a small but highly integrated transnational epistemic community (Haas, 1992) of statistical experts in charge of setting and reforming international standards.

In the wake of the seminal International Organization special issue on epistemic communities in 1992 (Adler and Haas, 1992), much IR scholarship has highlighted how ideas and beliefs can become institutionalized in international organizations and diffused through them. In the economic realm, the focus has been on the IMF and the World Bank in their promulgation of Washington Consensus-inspired policies (Barnett and Finnemore, 2004; Woods, 2006; Weaver, 2007; Chorev and Babb, 2009; Chwieroth, 2009; Broome, 2010; Kentikelenis and Seabrooke, 2017). While we take inspiration from this work, our case is somewhat different: statistical standard setting has much less immediate and obviously distributional effects than, say, IMF lending decisions. They are therefore less politicized, leaving more room for expert deliberation. In addition, while global statistical standards have been widely adopted around the world—at least on paper (Jerven, 2013; Kerner, Jerven and Beatty, 2017)—they have no legal force. That again increases the scope for expert deliberation. Finally, while much constructivist IR scholarship has focused on explaining policy change, in our case it is the stickiness of statistical standards in the face of profound economic transformations that calls for explanation. What keeps statistical standards in place when, as we outline below, they face such challenges to their solidity?
The professional norms structuring the production of economic statistics

Our answer to this question that emerges in the empirical analysis below emphasizes the importance of four professional norms in particular: certitude, comparability, continuity and coherence. While our investigation focuses on the measurement of cross-border flows of trade and capital, the relevance of these statistical norms is not limited to the transnational or the economic realm. It applies to the production of statistics more widely. This section therefore first lays out our conceptualization of these four norms and what motivates them in general terms.

Alongside their apparent simplicity, numbers are attractive in politics and policy-making due to their air of objectivity (Porter 1995; Dorling and Simpson 1999; Sætnan et al. 2011). Arguments backed up by numbers carry authority, even if the figures rest on shaky foundations. This emphasis on numbers in public policy has only grown as new public management has introduced corporate practices such as auditing and cost-benefit analysis into the public sector (Power, 1997; Knafo, 2019).

Claims to objectivity require reliable techniques to gather and aggregate data: assessments based on individual judgement and experience must give way to indicators that are readily reproducible by building on unambiguously quantifiable information. Thereby statistics introduce a countability bias into public policy, systematically privileging information that can be entered into spreadsheets (student numbers and awarded diplomas, prices of goods, number of people with full-time jobs) at the expense of things that are hard to quantify (student learning, value-creation, job security and satisfaction) (Muller, 2018). Because reliability has a specific meaning in scientific measurement (Krippendorff, 2008), we use the label certitude to describe the associated norm: statistics should contain as little information as possible that requires subjective interpretation. Although the norm of certitude is aligned with statisticians’ mandate to produce “objective” information, it becomes a problem when the properties we want an indicator to capture grow resistant to straightforward quantification.

To be sure, economic statistics never speak for themselves, even if we ignore how they were put together. They need to be narrated and put in context for us to make sense of them (Muniesa, 2014; Beckert, 2016; Leins, 2018). Rising unemployment may signal the malfunctioning of stifled labor markets (the liberal interpretation) or cyclical gyrations of
unchecked capitalism (the critical one). Statistics require policy goals, programs, and interpretative frames to unleash their full force (Abolafia, 2010).

Political and policy narratives that use statistics often have a comparative dimension: they compare units with each other (countries, provinces, schools, and so on) and track the evolution of indicators over time. For statistics to function in this way, measurement standards must be harmonized between units and stay constant over time. If we use different yardsticks in different places or adapt them from one year to the next, observations are no longer directly comparable.

Two norms follow: comparability (the inter-unit comparison) and continuity (the constancy of measurement standards over time). Both limit the adaptation of statistical standards to changing economic circumstances. Users who demand continuity—be they policy-makers or academics—will object to frequent breaks in time series. Comparability works differently. Once countries have agreed to a shared measurement standard, it will take collective agreement to adapt it, and the harder the object of measurement is to capture in updated standards, the longer such agreement will take.

Certitude, comparability and continuity are three of the four professional norms we highlight. They are, indeed, norms. Whether they achieve their goals is a different matter. As has often been noted in measurement theory, concept validity and reliability can be at loggerheads. If the concept validity of a reliable proxy is poor, the ultimate measure may contain little information about the concept it purportedly captures. Certitude as a norm can generate mock-accuracy.

Applying similar standards to very different countries can generate data that make meaningful comparison impossible, just as sticky measures may fail in the face of societal transformation. Poverty indicators are a good example: many countries define household poverty relative to median household income. As societies grow more affluent, the material deprivation associated with poverty may vanish: at the same time, the condition of occupying the bottom rungs of the societal ladder remains constant. Depending on which dimension of poverty one highlights, keeping the measure constant may frustrate or aid over-time comparison. Either way, it is far from obvious that keeping measures constant aids comparability if the object being measured changes.

The final fourth norm that obstructs the adaptation of statistical standards is what we call coherence. In macroeconomic statistics, individual measures do not exist in isolation.
Instead, theory or common sense tell us how they relate to each other as variables in a model or representation of “the macroeconomy” (Mankiw, 2017). According to economic theory, GDP, for instance, can be compiled in three ways, based on production, expenditure, or income (Lequiller and Blades 2006). For the sake of theoretical coherence, the total market value of goods and services (production-side) should be identical to the sum of consumption, investment, government purchases, and net exports (expenditure-side), and to the sum of labor and capital income (income-side). Macroeconomic theory specifies how the different quantities relate to each other, and because they interact like cogs in a clockwork, the definition and measurement of one cannot be changed without affecting the others.

All four norms not only act as brakes on adapting measures to new circumstances; they also limit the ability of statistical representation to incorporate ambiguity. Certitude clearly privileges specific numbers, just as comparability across units and over time breaks down without them. Coherence, too, demands exact quantification: admitting to ambiguity in one area would infect the whole interlinked system.

In our empirical argument we outline how statisticians’ pursuit of these four norms has fueled a growing concept-measurement gap in balance of payments statistics: measures of international economic transactions are ever less aligned with the theoretical constructs they purport to measure. By adopting this standpoint, we imply that statistical standards can generate figures that fit more or less well with economic reality. For skeptics, this is controversial. The whole point of social statistics is to make the world legible (Foucault, 1974), and as economic interactions are themselves shaped by how we interpret the economic world—including through its representation in statistics (Searle, 1995)—it may seem futile to ask how well statistics represent a world external to them.

The performative effects of statistics notwithstanding, we believe that some ways of measuring for example trade give better or worse impressions of cross-border transactions than their alternatives. For example, statistics that count all goods that enter Germany through the port of Rotterdam as “imports from the Netherlands” would ignore that the vast majority of these goods were produced in third countries. A standard that uses the country of origin to designate the source of imports is better because it corresponds more closely to the definition of “trade.” Likewise, a GDP standard that simply adds up the value of products leaving a country’s factories, without discounting the value of intermediate imported goods, would offer a skewed impression of the productive capacity of that country.
Although statistics do not provide a one-to-one representation of external economic reality, some give a better indication of the world around us than others. We reject the extreme position that all statistics, no matter how they are compiled, are equally defective. Their ability to capture a phenomenon is akin to what Karl Popper has called “truthlikeness”—proximity to the truth—and what others have called verisimilitude. Although it is difficult to judge the relative importance of different defects contained within statistics, we are confident in diagnosing a growing concept-measurement gap when the defects unambiguously loom larger over time, as is the case in BOP statistics.

Balance of payment statistics and their discontents

The monitoring of imports and exports were an obsession in mercantilist Europe in the 16th and 17th centuries (Lipsey, no date; Studenski, 1958; Morgenstern, 1963; McCormick, 2009). These efforts only intensified as governments systematized their economic records in subsequent centuries. The first attempt to collect international BOP statistics involved the League of Nations in the 1920s and 1930s. In the aftermath of the Second World War, the responsibility shifted to the IMF (Alves, 1967). As the guardian of international financial and economic stability in the Bretton Woods era, the Fund was responsible for identifying unsustainable imbalances in global financial flows (IMF 1948, p. 1). Towards this end, the IMF strove for international conventions on how to collect data on cross-border payments. The first Balance of Payments Manual (BPM1) issued in 1948 (IMF 1948) offered standardized templates for member countries to fill out each year. A slightly expanded version, with more detail about what to include and exclude, followed two years later (IMF 1950). Since then, the IMF’s BOP Statistics enterprise has only grown in ambition.

Although the users of BOP statistics typically assume their accuracy (authors 2019), insiders in the statistical community have been quietly voicing doubts since the 1980s. A 1987 IMF report, for instance, found that:

[in] the period after 1979, the available statistics on the world current account began to show a large negative discrepancy…. Concern that such discrepancies could lead to inappropriate policy reactions was heightened in 1982, when the excess of reported debits exceeded $100 billion…. [I]mproving the world’s data on current account transactions will be a
formidable task, especially in an environment where the capacity for statistical measurement is challenged by rapid changes in the technology and forms of international transactions and by budgetary constraints. (IMF 1987)

Five years later, a similar report on capital account discrepancies reached even starker conclusions, finding the “world capital accounts system” to be “in a state of crisis” (IMF 1992). The stakes were clear: “there are strong indications that this body of information on which good economic management depends is undergoing a serious and progressive deterioration” (IMF 1992).

Our analysis of mirror trade statistics (authors 2019) compared the trade or capital flows one country reports sending to another country with the figure this second country reports for incoming flows. In principle, the two should match; in practice, they do not. In 2014, the value of exports of merchandise goods from the Netherlands destined to neighboring Germany was estimated at US$ 165.6 billion by the Dutch authorities, while official figures from Germany valued imports from the Netherlands at US$ 96.6 billion; the USA estimated importing goods from China worth US$ 466.7 billion, while Chinese sources indicated their value to be US$ 397.1 billion, and so on (own calculations based on IMF DOTS database).

Such discrepancies are the norm. A comprehensive analysis of a global dataset of bilateral merchandise trade flow numbers found that mirror records differ, on average, by no less than a factor of 1.7 (Schultz 2015, p. 138). The situation is even worse for capital flows which are harder to measure than merchandise trade. An IMF analysis of discrepancies in their own bilateral Foreign Direct Investment (FDI) data reported that

for 44 percent of the 1,805 published bilateral economy pairs ... one economy’s number is at least twice as high as the counterpart economy’s number, and for almost 10 percent of the pairs, one number is at least 10 times higher than the mirror number. (Damgaard and Elkjaer 2017, pp. 5–6)

Despite sustained efforts by international organizations to harmonize statistical standards, measurement inaccuracies have not decreased over time (authors 2019). Put simply, there is a lot of noise in BOP statistics—so much that statisticians highlight that BOP statistics are “human-made estimates according to an internationally agreed methodology” (anonymous
interview, August 22, 2017) and that they should be used to “compare broad trends, not absolute levels” (anonymous interview, April 25, 2017).

The roots of these large and persistent measurement inaccuracies can be broken down into two clusters. First, a range of factors can lead countries—even if they formally adhere to the same global statistical standards—to assign a different value to the same transaction, such as cross-national differences in data collection practices, differing levels of statistical capacity, the use of different versions of statistical manuals, and so on. Massive increases in the volume and complexity of international economic transactions have multiplied the potential for measurement errors, the probability that a transaction will escape the nets of statistical measurement, or that it will be misattributed in the national accounts.

Second, official statistics map less and less well onto the economic complexity they purport to capture—the concept-measurement gap pur sang. They implicitly model the world economy as an interconnected system of semi-closed national economies (Lepenies, 2013; Masood, 2016). Yet this conceptualization is less and less appropriate to capture economic activities in an ever more integrated global economy, in which trade and capital flows crisscross national borders in enormously complex patterns (Oatley, 2019).

The collection of merchandise trade statistics faces serious difficulties to distinguish confidently between the places in which cargo is (un-)loaded and the locations where it was actually produced or consumed (Ahmad, 2018). And as global production chains deepen, the statistical blend of such conceptually distinct flows increasingly distorts interpretations of the data (interview with Fabienne Fortanier, Head of Trade Statistics at OECD Statistics Directorate, Paris, June 6, 2017). Statistics on trade in services raise additional questions (Giovannini and Cave 2005), not least when they struggle to separate actual cross-national transactions from mere MNE-internal accounting procedures. The growing division between the geography of corporate activities and the associated accounting practices can lead to situations in which companies’ domestic sales are counted as services “trade” merely because they are registered abroad for tax purposes—a phenomenon that the renowned economist Robert Lipsey (2006, p.37) refers to as “phantom flows of trade”. Such issues pose a serious challenge for the validity of trade statistics. If left unaddressed, established indicators risk to gradually “lose their meaning” (Lipsey, no date).

While global companies’ use of off-shore structures can severely distort trade statistics, the implications for capital flow statistics are even graver. To minimize tax payments, MNEs
commonly create special purpose vehicles in low-tax offshore jurisdictions and “book” profits on intellectual property there (Shaxson 2012; Finér and Ylönen 2017; Tørsløv et al. 2018). As Maria Borge, Head of Foreign Direct Investment Statistics at the OECD, put it in a paper co-authored with Cecilia Caliandro:

FDI statistics can ... reflect other factors, such as fiscal optimisation to reduce tax burdens and the increasing sophistication in MNEs’ capital structures. This can make it difficult to interpret FDI statistics, in the sense that they are not ‘real’ and no longer represent ‘long-term’ investments in a country. (Borge and Caliandro 2018)

To distinguish between long-term productive investments and short-term speculative capital flows (itself a questionable dichotomy, see de Goede, 2005), BPM defines cross-border acquisitions of at least 10 percent of a company as FDI, with all smaller investment being classified as FPI (Foreign Portfolio Investment). At the same time, the U.S. Bureau of Economic Analysis (Ibarra-Caton and Mataloni 2014) and Eurostat (2016) have estimated that between one-half and two-thirds of total BOP FDI in- and outflows come from or go to off-shore special purpose entities (SPEs) rather than an identified parent or subsidiary company. Determining the purpose or ultimate destination of these flows cannot be established with the current tools available to BOP statisticians. Such impenetrable ownership structures make it challenging to distinguish between long-term investments and speculative investments by for example private equity or hedge funds (Blanchard and Acalin 2016; interview with U.S. BEA economists, Washington, September 20, 2017), or between genuinely ‘foreign’ investments and corporate inversions.

FPI statistics face similar problems. Short-term capital flows are channeled through opaque structures of financial intermediaries, which BOP data is unable to track. As a result, official figures are biased towards custodian centers such as Liechtenstein, Luxembourg or Switzerland (Bertaut, Griefer and Tryon, 2006; Bryan, Rafferty and Wigan, 2017; Tørsløv, Wier and Zucman, 2018), and national statisticians (and tax authorities) struggle to estimate the equity and debt positions of residents who park their assets and liabilities in offshore financial centers (Fichtner, 2017). In a global financial system in which “nationality” is a “tradable attribute of an asset” (Bryan, Rafferty and Wigan, 2017) rather than a physical location, attempts to measure the “national” holdings of an asset can fundamentally mislead. For example: a second look at US FPI figures by the Treasury Department in the early 2000s (Griefer, Lee and Warncok, 2001) revealed that more than half of short-term
capital “outflows” registered in the US national accounts in the 1990s were not driven by US investors deciding to move their capital abroad. Instead, they represented investments by US investors in (formerly) US firms, which were taken over by foreign owners. In other words, what looked like FPI “outflows” in the national accounts were conceptually more akin to FDI inflows. In the words of the chairman of the IMF BOP committee in 1989,

In view of our responsibilities for the good working of the international monetary system, I think we were in a not so comfortable situation with a state of knowledge of the international capital flows which was, well, correct and sufficient before the financial revolution the world has witnessed during the eighties. But now the picture is totally different, and the enormous loophole in our knowledge of this phenomenon was a real problem including for our credibility and for the credibility of our work in the [World Economic Outlook] field and in many other spheres. (International Monetary Fund, 1989)

In short, national accounting templates that assume simplistic economic relationships capture our current economic realities less and less well. Denationalized production and opaque corporate and financial structures have undermined the validity and hence usefulness of BOP statistics.

Statisticians are well aware of these problems (e.g., Damgaard and Elkjaer 2017; IMF 1987, 1992; OECD 2016; UNECE et al. 2011). Indeed, they have been discussing them since the 1950s (International Monetary Fund, 1956; Smith, 1966). One way the international statistical organizations have sought to reduce asymmetries in BOP figures is through facilitating bilateral meetings between national compilers (interview with IMF statisticians, Washington, September 19, 2017), while recent standards (e.g. BPM6) seek to better capture “merchanting” or “goods sent abroad for processing.” Additional efforts are being pursued to update standards to better reflect nowadays realities: The IMF is exploring ways to get a better grip on the measurement of Special Purpose Entities in global financial flows (IMF 2016a), and the OECD has created a Trade in Value Added (TiVA) database that aims to disentangle gross trade flows from actual value creation. But even this has clear limits:

TiVA offers an interesting complementary perspective. But it is built on data which are not very precise, as compilation involves many imputations and data modeling to fill in the gaps. Essentially, it is
‘modelled’ data, not real data. (Interview with IMF statisticians, Washington, September 19, 2017; cf. Ahmad 2018)

Eurostat statisticians feel that

We are only at the very beginning of getting a grip on properly measuring globalisation in a systematic cross-country way in practice. Which parts of the production activities of MNEs are actually ‘taking place’ on the domestic territory of any given country? ... How can we distinguish between movements in GDP or its components which are relevant for the domestic economy and those which are driven by the worldwide activities of multinational companies? (Stapel-Weber et al., 2018)

On balance, these initiatives have failed to stem the deterioration of BOP measurement quality—they are "plasters on the holes of a sinking ship," in the words of one statistician we interviewed (anonymous interview, April 25, 2017).

If the data are as bad as we have outlined and if statisticians are aware of the problems, why do outdated international economic statistics still dominate representations of the global economy? What makes statistical standards so sticky when they increasingly fail to fulfill their goals?

Statistical norms and conservative bias

While we appreciate the practical challenges of producing high-quality statistics, these do not tell the whole story. Clearly, statisticians and the users of statistics could have reformed them to better suit changed circumstances and new perspectives. Examples from other areas in the statistical field show that this is possible. The World Bank and United Nations have adapted their “development” measures over time (Finnemore, 1996) while many governments have revisited ethnic categories in their censuses (Marquardt and Herrera 2015; Petersen 1987). This demonstrates that measurement systems can adapt in the face of social and political change.

Were economic statisticians truly unable to adapt statistical standards, they could have highlighted measurement deficiencies more forcefully. For starters, they could have refused to report deceptively precise point estimates—single figures—for FDI or FPI. The reporting of
data ranges is common in forecasting, for example for different climate change scenarios. While the use of confidence intervals might seem more intuitive for future projections, inaccuracies in the measurement of past economic transactions are frequently substantial enough to warrant similar caution. Were this impossible, statisticians could have abandoned obsolete measures, admitting that we simply do not know the investment relationship between two far-flung countries.

Why is none of this happening in BOP statistics? We argue that to understand the stickiness of statistical representations, we have to untangle the professional norms that underlie them. Based on our analysis, the four statistical norms introduced in general terms above (comparability, continuity, certitude, and coherence) stand out. The following sections illustrate their working in the production of BOP statistics.

**Comparability**

From its outset, the global statistical enterprise emphasized the need for comparable numbers. In the early days of BOP statistics, the League of Nations tried various strategies to encourage countries to report uniform figures with limited success (Alves, 1967). BOP statistics released by the League were thus not directly comparable across countries. In the eyes of a statistician involved in the elaboration of statistical standards at the IMF in the post-war period, this undermined the whole enterprise: “Because the attempt to achieve uniformity was only partially successful, the usefulness of the figures in the League’s publications is severely limited” (Alves, 1967).

The scope for cross-country comparison is the key attraction of multi-country databases. Indeed, we commonly think of international standards and best practices as improving data quality because they promote inter-agency learning and facilitate expert debate across borders. These dynamics notwithstanding, we typically underestimate the difficulties of achieving the requisite uniformity in numbers, collected as they are by disparate national agencies. The harmonization of accounting-technical standards is challenging, costly, and time-consuming. And because the “sunk costs” that harmonization demands are so high, the aim of comparability unwittingly retards change. Even when good reform ideas abound, countries struggle to agree on acceptable and implementable standards. To better grasp the diversity of cross-border investment flows, the OECD’s 4th edition of the Benchmark Definition of FDI (developed over several years with the IMF and the UN) asked countries to
compile separate figures for greenfield, merger and acquisition, and special purpose entity inward FDI flows since 2008 (OECD, 2017). But despite the obvious improvements promised by this distinction, progress has been frustratingly slow (interview with OECD statistician, phone call, May 30, 2017). So long as only a few countries report such data, they cannot be included in cross-national databases.

Harmonization as a statistical norm also limits the sensible adaptation of standards to national circumstances, a tension the IMF has noted since the inception of BOP data collection. In an internal letter dated June 22, 1953, A.B. Hersey from the Board of Governors of the Federal Reserve System wrote to the IMF: “Though flexibility is desirable, so is uniformity. The Fund’s problem is how best to reconcile the two objectives” (Hersey, 1953). Rules for FDI statistics offer countries alternative options to determine the value of inward FDI stocks, such that countries can pick the one that best fits their situation. But national statisticians may privilege convenience over quality of conceptual fit when they choose among the alternatives. As top-level Eurostat officials recently argued regarding national accounts, “any new indicator or breakdown, particularly in a European context, should be comparable across countries and not be seen as a GDP or GNI ‘a la carte’ for each country to choose from under specific circumstances” (Stapel-Weber et al., 2018).

Data compilation is done by national authorities with their own organizational structures and legal traditions. Even when “the concepts are exactly the same, […] the ways in which they are measured can be different” (Interview with Fabienne Fortanier, Head of Trade Statistics at OECD Statistics Directorate, Paris, June 6, 2017). Hence, if the comparability of figures is the goal, room for national discretion must shrink. As the authors of a bilateral asymmetry study of Germany and Portugal point out, to eliminate measurement errors, “harmonization of theoretical concepts is not sufficient. Essential is a common approach to the practical application and interpretation of concepts and definitions” (Deutsche Bundesbank, 1997).

Surveys on data collection practices by national statistical offices (IMF and OECD, 2003; United Nations Statistics Division, 2006) and bilateral reconciliation exercises have highlighted a large number of factors that can undercut the cross-national comparability of figures, such as at-odds currency conversions, the use of dissimilar valuation techniques, or differences in classification decisions for transactions that fall into a grey area.
In response, international organizations have pushed further to narrow national compilers’ room for interpretation and discretion in data gathering and reporting. But the most recent BPM compilation guide concedes that there are real limits: “Articulating balance of payments and International Investment Position compilation methodology is difficult because economies have developed procedures independently, and each national methodology may be considered unique. Some patterns emerge, but different national experiences have created different approaches as to the most appropriate methodology. Consequently, it is not possible to present a single methodology suitable in all cases. Instead, the Guide outlines various options that may be available” (IMF 2014). Adherence to the harmonization norm thus means that statistical standards change slowly and that when it happens, it may entail standards that are at odds with national circumstances. Given the potential for misuse, flexible standards are no solution. At the same time, even harmonized standards do not guarantee even superficially comparable data.

**Continuity**

While the norm of comparability requires countries to have a common yardstick, the norm of continuity seeks to ensure that we can capture developments over time. One of the great attractions of statistics is their claim to track macrosocial or economic developments (Trewin, 2007). If measurement approaches change and we are unable to retrospectively adjust past measurements, diachronic comparability is lost.

The IMF already worried in the 1950s: “[the] Fund should ... insure that continuity of the series is not disturbed” (International Monetary Fund, 1956). Even after the gradual switch to BPM6 in the 2000s, the IMF continued to receive requests for continuous time series data covering the past decades up to the present day (Shrestha et al. 2016, p. 6). Indeed, series continuity is a key argument in statistical disputes. As the global financial system shed its post-war shackles in the late 1960s, IMF statisticians agonized over what they called “empty shell” holding companies. The UK’s central statistical office concluded that

> while ... we would not dissent from the view which the I.M.F. [sic] say they have expressed that, in principle the statistics should ideally relate to the final origin or destination of investment, we do not believe that this goal is obtainable in practice. Any partial move in this direction would involve
serious discontinuities, from which we might lose more than we should gain. (Stanton 1967; emphasis added)

One way to meet the norm for continuity is to develop parallel statistics: to begin a new series while continuing with the old one for the time being. Although the figures from the TiVA initiative offer more verisimilitude than conventional trade statistics, they are not directly integrated into the BOP system to avoid breaking the series. Continuity over time is more important for some series and users than others. Academics using regression analysis, for example, typically rely on temporally extended series to disentangle the effects of multiple variables or to observe delayed effects. While the importance of a break in the series will vary from case to case, the need to continually adapt indicators to changing economic realities clearly diminishes the comparability of data over time—to the extent that authorities may stick with indicators even when they are becoming obsolete.

Certitude

Statistical systems have in-built preferences for measures that minimize the scope for subjective judgments or gross manipulation, an intuition that dovetails with good statistical practice. Statistics’ claim to objectivity—and their status as neutral arbiters in public affairs—hinges on reliable measures that follow the same routines. To denote the resulting penchant towards “hard” measurement procedures we use the label certitude. The norm of certitude in turn privileges elements that can be unambiguously quantified, that are directly observable and require no further interpretation.

Curtailing statisticians’ room for subjective judgment and maneuver has costs in terms of validity (cf. Schedler 2012). Even where informed estimates might generate the best data, they may be eschewed in favor of measurement procedures that rely on hard, unambiguous data. FDI statistics, for example, hang on the “nationality” of domestic firms’ foreign owners. But corporate “nationality” is a complex construct, especially when several owners from various jurisdictions channel investment through multi-country tax structures. In such instances, subjective judgment would be useful, for example by showing the holdings of Amazon Luxembourg to be mostly American investments. In practice, however, statistical standards opt for an unambiguous but ultimately misleading classification: the “legal residence” of an investor, which makes Amazon Europe a Luxembourg company.
Similar problems apply to capital flow statistics that try to distinguish between predominantly “financial” (FPI) and “productive” investments (FDI). Earlier versions of the Balance of Payments Manual relied on the qualitative judgement of national accountants (IMF 1961, p. 120; IMF 1977a, p. 138). The IMF subsequently abandoned this approach in favor of an unambiguous threshold: all foreign investments that involve at least ten percent of a company’s voting stock are to be counted as FDI (author 2018). Although this rule was always arbitrary, it has become more so in recent decades. Activist hedge funds increasingly buy and sell large corporate stakes for quick financial gain—directly contravening the assumption that large investments are automatically also long-term (interview with U.S. BEA economists, Washington, September 20, 2017). What makes this rule attractive despite its shortcomings is that it can be uniformly applied. It is reliable: a different person repeating the same procedure would get similar figures. But consistent rules are less flexible than qualitative judgments and risk ignoring changing circumstances. In a trade-off familiar to social scientists, reliability comes at the cost of validity.

These examples are indicative of a broader trend: globalization, digitization, and financialization have reduced the number of points at which we can more or less directly gauge economic quantities of interest. With globalized production, the complexity derives from ever longer and increasingly intermeshed production chains. In other instances, corporations erect complex legal facades to shape outsiders’ perceptions, irrespective of how these facades relate to productive activities on the ground or financial connections between the ultimate beneficiaries. The more globalized, digitized, and financialized our economies become, the wider the gap will be between the economic connections statistics purport to capture and what the statistics themselves suggest.

Certitude as a norm not only entails unambiguous measurement procedures; it also means a preference for point estimates—data in the form of single numbers—that obscure the uncertainty underlying statistics. Rather than claiming that, say, the trade deficit of the United States with Mexico in 2017 was 70.952 billion USD (United States Census Bureau 2017), it might be more honest to say that “We, the U.S. Census Bureau, have a sense that last year the deficit was somewhere between 65 and 75 billion USD.” But such a presentation of economic statistics is currently not considered acceptable; to retain credibility as social facts, they need to continue the pretense of certitude. As was mentioned in a IMF conversation in 1956: “We shall give you the figures even though we might not like them ourselves” (International Monetary Fund, 1956).
Point estimates are also necessary for interpolation, imputations, and other statistical operations, including regressions. At the same time, as the former deputy-director of the Dutch statistical office put it:

Statistical institutions have to guard the authority of their statistics. Therefore they will be reluctant to emphasise the shortcomings or to develop competing (conflicting) information. The authority of a set of statistics grows with the duration of its use. This encourages official statistical institutes to maintain existing statistics, and thus to be conservative in developing substitutes. (van Tuinen 2007, p. 267)

This need to safeguard the incontrovertible image of statistics is also appreciated by Eurostat statisticians: “Given the potential impacts on macroeconomic statistics across countries, and the adverse reaction of users to ‘surprises’ in data, this [globalization] presents a major challenge to official statisticians” (Stapel-Weber et al. 2018, p. 3; emphasis added). Peter van de Ven, head of national accounts at the OECD, and Brent Moulton, former head of national accounts at the U.S. Bureau of Economic Analysis (2018, p. 18) equally worry that quirks in the data—for example the on-paper relocation of economic activity that led to a jump in Irish GDP—can be abused to disqualify statistics more generally.

Certitude as a statistical norm thus puts macroeconomic statistics in a double bind. On the one hand, it is felt necessary to sustain the figures’ credibility. On the other, it stands in the way of both a more creative and flexible adaptation of statistical standards to new economic circumstances and more open admission of the increasing uncertainty that underlies macroeconomic figures.

Coherence

The final norm is coherence: individual economic measures should fit into a larger, coherent whole that depicts the entire economy. Individual components of BOP statistics are meant to offer an encompassing image of inter-country economic exchanges. FDI and FPI statistics, for example, are direct complements that together are meant to capture cross-border investment in its entirety.

Statisticians and econometricians played key roles in the elaboration and practical implementation of John Maynard Keynes’ ideas, especially that of the national economy as
a system of logically inter-related parts. Jacques Polak, director of the IMF’s research department from 1958 to 1979, developed the “Polak model” relating key domestic macroeconomic variables such as GNP growth and domestic credit of the banking system to cross-border economic variables such as foreign exchange reserves and trade (Polak, 1997; Woods, 2006). Later theoretical refinements formalized the relationship between the balance of payments, changes in the domestic money supply, and developments in the real economy (IMF 1977b). Rather than isolated macroeconomic quantities to be observed individually, the constituent elements of the balance of payments came to be seen as building blocks of a larger integrated whole. This suggested for example that a net surplus or deficit in cross-border flows must imply a depletion or increase in net foreign reserves, such that the latter could be imputed from knowledge of cross-border financial flows.

Due to these developments, the BOP system does not stand on its own but is a part of the System of National Accounts. The “linkage of the international investment position and balance of payments accounts to the rest of the world account in the System of National Accounts (SNA) is strengthened and harmonized to the maximum extent possible” (IMF 2009). The different sectoral accounts are communicating vessels: a change in one account must be accounted for elsewhere—a trade deficit comes with a capital account surplus, while export revenues in the BOP are also someone’s income in the SNA. Revisions of the BOP and the SNA thus progressed in parallel in the 1980s: the “need for compatibility between the two standards is one reason both are being revised” (IMF 1992, p. 16). Such linkages complicate efforts to update statistical definitions and procedures: changes in one place have (potentially undesirable) knock-on effects elsewhere. It triggers a “train of adjustment” (International Monetary Fund, 1956).

Thinking of national economic and BOP statistics as an integrated whole also means that statistical concepts are often deductively defined. The conceptual coherence of accounts tempts us to impute values for concepts that are not directly observable. If in an equation such as $a + b = X$ we have measures for $a$ and $X$, we might impute $b$ and then report this as a known quantity. But in the process, all kinds of measurement problems with $a$ and $X$ disappear. The quality of $b$ as a data point is going to be no better than those of $a$ and $X$: here imputation creates a false sense of balance, more often than not due to statistics being forced into balance (through the delegation of imbalances to residual categories such as “errors and omissions”) rather than a “natural” balancing of items.
The theoretical elegance of the models underlying the collection of national accounts data stands central in today’s BOP statistics. While the IMF’s original efforts to collect macroeconomic data merely sought to assemble national statistics from various sources, the project has evolved into an intellectual enterprise to integrate the figures into a theoretically coherent whole. The style and substance of the IMF’s Balance of Payments Manuals mirror this development: the pioneering BPM1 (IMF 1948), less than 50 pages long, simply provided a set of tables to be filled out by national statisticians. In stark contrast, the most recent version, BPM6 (IMF 2009), is a highly didactic document of almost 400 pages, accompanied by a separate 600-page Compilation Guide (IMF 2014). As the authors of the preceding BPM5 highlighted, the manual

not only defines and describes the content of the categories employed but also attempts to explain their rationale. … With these amendments, the Manual has become as much an introduction to the principles of balance of payments accounting as a guide to reporting. (IMF 1995, p. 2)

Statisticians rightly take pride in the sophistication of the models they have developed over the years. Most national accountants are trained economists, and in view of the mathematical fetish that dominates the discipline (Fourcade, 2010), the theoretical models underlying contemporary national accounts arguably play an important role in granting legitimacy to the statistical profession. As such, statisticians’ modeling of the world economy as a logically coherent, internally balancing system performs an important social function. But the same ambition simultaneously represents a monumental obstacle for attempts to reform statistical standards since the adaptation of standards for only one item may require the overhaul of the system as a whole to ensure its coherence since “solving asymmetries in one item may create new asymmetries in another one” (IMF 2016b, p. 21). In this way, the norm of coherence further reinforces statisticians’ bias towards sticking with theoretically elegant but outdated statistical standards.

Conclusion

BOP data are our window on the global economy, used to gauge and analyze trends in trade, international financial flows, and cross-border investment patterns. We normally take this
data for granted, and analysts of the global economy rarely spend much time dissecting the data itself.

The statisticians we interviewed were painfully aware of the problems discussed in this article. But unless one pores over footnotes in statistical yearbooks, the published numbers continue to project a level of accuracy that sits uneasily with the ambiguities in the data. Headline figures take little heed of the growing problems, for example by overhauling definitions or replacing outdated concepts.

This growing concept-measurement gap has severe ramifications for the quality of economic policy, public debate, and academic analysis. So why do outdated international economic statistics continue to dominate representations of the global economy? Our analysis has focused on the norms guiding the transnational epistemic community dominating statistical standards. These norms, we have argued, damage the verisimilitude of international economic statistics.

To be sure the norms themselves are intuitive and plausible enough. Dennis Trewin, former head of the Australian Bureau of Statistics, argued that to “be useful, international statistics must be relevant, of good quality and consistent across countries and across time” (Trewin 2007, p. 308). Here we find, in a nutshell, the four norms we have discussed: what we have called certitude is seen as a way to quality, while consistency across countries and time is what we have called comparability and continuity. From a policy perspective, relevance is derived from statistics’ commensurability with macroeconomic concepts and models as used by policymakers.

But as we have argued, molding statistical standards after these norms in fact damages the measurement quality of those statistics. In this sense, we go further than many others who criticize the quantification of social and political life. The contradictions between the norms that guide economic statistical practice mean that – given contemporary economic dynamics – statisticians are doomed to fall ever shorter of their own aspirations. Unless we sacrifice one or more of the expectations users commonly have of economic data, we are destined to live with economic representations that fail by their own standards.
References


Mind the Concept-Measurement Gap: the Statistical Norms Driving International Economic Mismeasurement

www.fickleformulas.org


