Stages of Diversification Revisited

Dany Bahar Sebastian Bustos Muhammed A. Yıldırım*

Abstract

This paper revisits the debate on the relationship between economic growth and specialization or, conversely, diversification of a country's production and export baskets. In earlier work, Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011) find a U-shaped relationship between income and concentration, implying that countries diversify first and then specialize as they climb the income ladders. This paper scrutinizes this finding to elicit the drivers of this result. First, the U-shaped relationship becomes an L-shaped one after excluding countries that are rich in natural resource from the sample. Second, analysis using more granular data to compute concentration finds that countries' transition from middle-income to high-income status is associated with diversification, and not with specialization. Third, when looking at the universe of countries that transitioned from low-income status to middle-income status and to high-income status, the paper finds no particular pattern of specialization or diversification. However, it finds that regardless of the concentration of countries' export baskets, the sophistication or complexity of products in the export basket progressively increases with country income.

Keywords: structural transformation, development, diversification, concentration, exports, production

* Dany Bahar is Associate Professor of the Practice of International and Public Affairs at Brown University and Senior Fellow at The Growth Lab at Harvard University. e-mail: dany_bahar@brown.edu. Sebastian Bustos is Senior Fellow at The Growth Lab at Harvard University. e-mail: sebastian_bustos@hks.harvard.edu. Muhammed A. Yildirim is the Director of Academic Research at The Growth Lab at Harvard University and an Associate Professor of Economics at Koç University. e-mail: muhammed_yildirim@hks.harvard.edu. This paper serves as a background paper to the *World Development Report 2024: The Middle-Income Trap.* The findings, interpretations, and conclusions expressed in this paper are entirely those of the authors. They do not necessarily represent the views of the International Bank for Reconstruction and Development/World Bank and its affiliated organizations, or those of the Executive Directors of the World Bank or the governments they represent. The authors thank Jean Imbs, Martin Fiszbein, Ricardo Hausmann, Somik Lall, Justin Yifu Lin, and Romain Warcziag, as well as members of the Harvard Growth Lab for invaluable feedback.

Introduction

Economic growth has been directly linked to the process of structural transformation in terms of both sectoral production and exports. The seminal work by Imbs and Wacziarg (2003), and its extension by Cadot, Carrère, and Strauss-Kahn (2011) using trade data, established a widely cited regularity on the empirical relationship between income levels and the concentration of a country's production and export baskets. In particular, these studies show that specialization follows a U-shaped curve. At low levels of income, countries' production and exports are highly concentrated in a few sectors; as income grows, countries tend to diversify. But the results of these studies indicate a level of income beyond which the sectoral distribution of production tends to concentrate again. According to these findings, sectoral diversification goes through two stages: first, diversification; and later, increasing concentration. Yet, another line of research convincingly shows a robust correlation between income levels and diversification (Hausmann and Klinger 2007; Hausmann et al. 2014). In light of this conflicting evidence, this paper revisits this question and attempts to close the gap between these results.

Exploring the tension between specialization and diversification at higher levels of countries' per capita income (GDP) is crucial to better understand the relationship between economic growth and the process of structural transformation. Understanding this relationship is essential to better inform economic policy. Does the transition from middle-income status to high-income status go along with specialization or diversification? Both paths can be explained by different theories. On one hand, even after countries diversify and grow, they become more integrated in the global economy, allowing them to reconcentrate in a particular set of production and export lines, fulfilling the gains from trade that come from comparative advantage. On the other hand, global integration also implies diffusion of knowledge, and this could also result in productivity gains across industries reflected in diversification, rather than concentration (Bahar, Hausmann, and Hidalgo 2014).

This paper empirically revisits the evidence of both Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011) that suggests a U-shaped relationship between income and concentration. The analysis sheds light on a number of important findings that complement their original results and show that there is a more nuanced relationship between the two variables.

The analysis begins by showing that the right-hand side of the U-shaped relationship between sectoral income and concentration depends crucially on natural resources. That is, the U-shaped form of the nonparametric relationship between income and concentration is not robust to the exclusion of countries that are rich in natural resources. The respecialization stage found in previous research appears to be driven by countries that are rich in natural resources, which tend to have both high incomes and highly concentrated production or export baskets, even with respect to their non–natural resource export baskets. Thus, the analysis excludes countries rich in natural resources, and a set of natural resource products, from the samples. When this is done, the relationship between income and concentration looks more like an L-shaped curve than a U-shaped one.

Next, the analysis explores the nonparametric relationship between income and concentration using different levels of sectoral aggregation. It takes advantage of the fact that trade data provide enough granularity—much more than production data—to make proper comparisons on this dimension. The analysis finds that the U-shaped relationship is much less pronounced when computing the concentration indexes at higher levels of disaggregation (for example, sector and product classifications at the detailed 6-digit level versus the aggregate 2-digit level based on the Harmonized System), implying that richer countries highly diversify within clusters. The analysis also shows that these results are not mechanically driven by certain 2-digit codes having a different number of 6-digit lines underneath them in a way that correlates with income.

Finally, instead of just looking at cross-sectional results, the analysis explores the dynamics of the handful of countries that transitioned from below-median to high income levels. It finds that there is no clear pattern in terms of diversification or specialization. However, one characteristic that seems to increase monotonically with income, regardless of diversification or specialization, is the sophistication or complexity of a country's export basket, measured by different metrics.

This paper's results speak to the important link between structural transformation and economic development. While the debate in the literature on whether economic development is related to diversification or specialization is unsettled,[1] overall empirical evidence is scarce. This paper—by revisiting the earlier seminal work of Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011)—provides some answers: the transition from low-income to middle-income status goes along with diversification of both sectors and varieties, whereas the transition from middle-income to high-income status, in the rare occasions that it occurs, involves a slight concentration in sectors but continuing diversification of lines within those sectors.

The paper rationalizes the results by developing a simple neo-Ricardian model of international trade to show the relationship between income and concentration of production and exports. When bringing the model to the data, for most countries, a productivity shock results both in increases in income and a more diversified portfolio of exports. In this simple model, the initial level of diversification is crucial for whether the shock will diversify or concentrate production.

Data and Metrics

Data Sources

For most of the empirical analysis, the main source of data comes from data on international trade collected by the United Nations global trade data platform, UN COMTRADE, through the Atlas of Economic Complexity of The Growth Lab at Harvard University (Hausmann et al. 2014). The analysis focuses particularly on export data, and uses two product classifications depending on the aim of the empirical analysis. The Standard

[1] See, for example, Imbs and Wacziarg 2003; Koren and Tenreyro 2007; Hidalgo et al. 2007; Hausmann et al. 2014; Cadot, Carrère, and Strauss-Kahn 2011.

Industry Trade Classification (SITC) Revision 2 dataset enables a look back to the mid-1960s, to as recently as 2020, with approximately 760 product categories at the 4-digit level of disaggregation. The Harmonized System 1992 (HS92) dataset has consistent data starting in 1995 but has the advantage of containing nearly 5,000 different product categories at the 6-digit level. While the SITC dataset will prove useful when exploring transitions of countries along the income distribution given its longer time dimension, the HS92 dataset provides more degrees of freedom when exploring concentration measures at different levels of aggregation (for example, 2 digits versus 6).

Most of the empirical analysis focuses on export data, instead of production data as in Imbs and Wacziarg (2003), because export data have important advantages over production data on several fronts. First, export data follow international standards and thus are better suited to cross-country comparisons. Second, export data have higher levels of disaggregation, a feature that is central to this analysis. Third, in the cases in which export data from a given country are missing, they can be imputed by looking at the imports to all partner countries, which cannot be done with production data. Fourth, trade data provide more than 60 years of disaggregated data following a fairly constant classification. In this sense, parts of this paper closely follow not only the work of Imbs and Wacziarg (2003) but also of Cadot, Carrère, and Strauss-Kahn (2011), who exclusively use export data for their analysis, finding similar results.[2]

The research also incorporates the Industrial Statistics (INDSTAT) database of the United Nations Industrial Development Organization (UNIDO), used earlier by Imbs and Wacziarg (2003).[3] These data are used to replicate the original results and to provide insights and robustness checks with it. These data sources are complemented with data on country-level per capita income from the Penn World Tables and data on rents from natural resources from the World Development Indicators. All in all, the analysis includes data for up to 177 countries for up to about 6 decades of data, depending on the product classification used in the estimation.

Concentration Metrics

The analysis uses a number of concentration metrics to measure the level of concentration of a country's production or export basket. These measures include the Gini index, the Theil index, and the Herfindahl-Hirschman index (HHI). The Gini and Theil indexes are measures of inequality. HHI is a measure of concentration. In their widely used formulations, the Gini and Theil indexes capture how far the observed distribution of productions is from a uniform distribution. In theory, the Theil index can also compare the observed distribution to other distributions, not only the uniform distribution. Another main advantage of the Theil index is its ability to be decomposed into subindexes, which will be used later in this paper. HHI, on the other hand, is a measure of concentration and effectively captures the number of industries in which production or exporting is concentrated. The analysis also calculates the share of products that a country

[2] An additional reason is that the UNIDO data require important assumptions in terms of when is it reasonable to interpret zero values as missing ones, which has important implications for the robustness of the results. This issue is discussed in detail in appendix B.

[3] The authors thank Jean Imbs and Romain Warcziag for generously sharing the data and code for replication.

produces or exports with Balassa's RCA ≥ 1, a measure of revealed comparative advantage (RCA) widely used in international economics. In this paper, this measure is labelled "active export lines." In contrast to the previous measures, the active export lines increase with the level of diversification.

All the concentration measures are based on the distribution of production or exporting in a country. Suppose that there are C countries and N industries. Let $X_{ci,t}$ denote the production or exports of the country c in industry i in year t. Therefore, the share of industry i in the production or exports of the country c is given by:

$$s_{ci,t} = \frac{X_{ci,t}}{\sum_{i=1}^{N} X_{ci,t}}.$$

The Gini index is constructed by comparing the absolute differences between any pair of industries with the total sum of exports. If the distribution of shares or levels is uniform, then the Gini index approaches 0. The theoretical maximum of the Gini index is 0.5. Mathematically, the Gini index for the country c in year t can be written as:

$$\mathrm{GINI}_{c,t} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|X_{ci,t} - X_{cj,t}\right|}{2N\sum_{i=1}^{N} X_{ci,t}} = \frac{\sum_{i=1}^{N}\sum_{j=1}^{N}\left|s_{ci,t} - s_{cj,t}\right|}{2N}.$$

The Theil index is built on the concept of entropy developed in the field of information theory. Given a sample mean denoted by $\mu_{c,t} \equiv \sum_{i=1}^{N} X_{ci,t}\,/N$, the Theil index for country c in year t can be written as:

$$T_{c,t} \equiv \frac{1}{N}\sum_{i=1}^{N}\frac{X_{ci,t}}{\mu_{c,t}}\ln\left(\frac{X_{ci,t}}{\mu_{c,t}}\right).$$

When compared to a uniform distribution, the index boils down to the Shannon entropy of shares minus a constant:

$$T_{c,t} \equiv \sum_{i=1}^{N} s_{ci,t}\ln\left(s_{ci,t}\right) - \ln(N).$$

The Hirschman-Herfindahl index (HHI) can be thought of as an expected share of industry. The inverse of HHI is often used as a measure of the effective number of industries in a country. Mathematically, the HHI of the country c in year t is given by:

$$\mathrm{HHI}_{c,t} \equiv \sum_{i=1}^{N}\left(s_{ci,t}\right)^2.$$

The paper also presents results with the share of active product or export lines with a revealed comparative advantage (RCA) greater than or equal to one, out of the total possible export lines. An RCA equal or larger

than one implies that a country's production or exports of a given product, in a given industry or sector, are above average relative to the production or exports of that same product overall in the world.

Mathematically, the share of industry i in the world production is given by:

$$S_{i,t} = \frac{\sum_{c=1}^{C} X_{ci,t}}{\sum_{c=1}^{C} \sum_{i=1}^{N} X_{ci,t}}.$$

Specifically, the RCA of country c in producing or exporting in industry i at year t is computed as:

$$\text{RCA}_{ci,t} \equiv \frac{s_{ci,t}}{S_{i,t}}.$$

With a slight adjustment of notation, the share of active lines for country c in year t is calculated as:

$$\text{Lines}_{c,t} \equiv \frac{1}{N_t} \sum_{i=1}^{N_t} (\text{RCA}_{ci,t} \geq 1).$$

where N_t denotes the number of different products in classification traded in the world in year t.

Main Results

This analysis is based on the method developed by Imbs and Wacziarg (2003) to establish a nonparametric relationship between concentration of countries' production or export baskets and their income levels. Appendix A describes these methods[4] and shows that this analysis can replicate the main results of Imbs and Wacziarg (2003) with a high level of precision (see figure A1).

The main takeaway from Imbs and Wacziarg (2003) is that empirically a U-shaped relationship between income and concentration exists. Low-income countries tend to have production baskets that are quite concentrated; as they grow their production baskets diversify; and as they continue to transition from middle-income to high-income status, their production basket concentrates again. Appendix B discusses the sensitivity of these results based on a number of data choices.

The analysis reexamines the U-shaped empirical relationship between income and concentration levels established by Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011) by incorporating important considerations regarding the data, methods, and the dynamics of the countries' actual transitions.

[4] This is the same methodology used by Cadot, Carrère, and Strauss-Kahn (2011) relying on export data, rather than production data.

First, it presents results showing that the inclusion or noninclusion of countries rich in natural resources in the analysis can significantly change the conclusions from previous studies.

Second, it shows that when using different levels of disaggregation of the data, the respecialization pattern that is seen in high-income countries becomes much less pronounced.

Third, it explores the dynamics of countries that actually transitioned across the income distribution over the past decades and find no clear pattern in terms of diversification or specialization. However, it finds that one characteristic of their export basket that seems to be progressively increasing with income, regardless of diversification or specialization, is the sophistication or complexity of products in their export basket.

Most of these results rely on export data, given the reasons discussed in the introduction.

The Role of Natural Resource–Rich Countries

Natural resource–rich countries tend to be outliers in the relationship between per capita income and concentration of exports. Moreover, countries that are rich in natural resources tend to have, as a consequence, more concentrated non-natural resource export baskets (Bahar and Santos 2018). Thus, an important consideration is whether the patterns described by Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011) can be affected when the set of countries rich in natural resources is excluded.

This analysis explores this question by reestimating the relationship between income and concentration per capita including and excluding countries that are rich in natural resources from the estimates. The definition of countries rich in natural resources includes countries that rank above the 75th percentile in either of the following two variables: (1) rents of natural resources as a share of GDP (from the World Development Indicators) *or* (2) share of natural resource exports in the country's total export basket.[5] The list of countries that classified as rich in natural resources is presented in table C1 in appendix C. The appendix also shows that the results are robust to varying the threshold defining the sample of natural resource–rich countries.

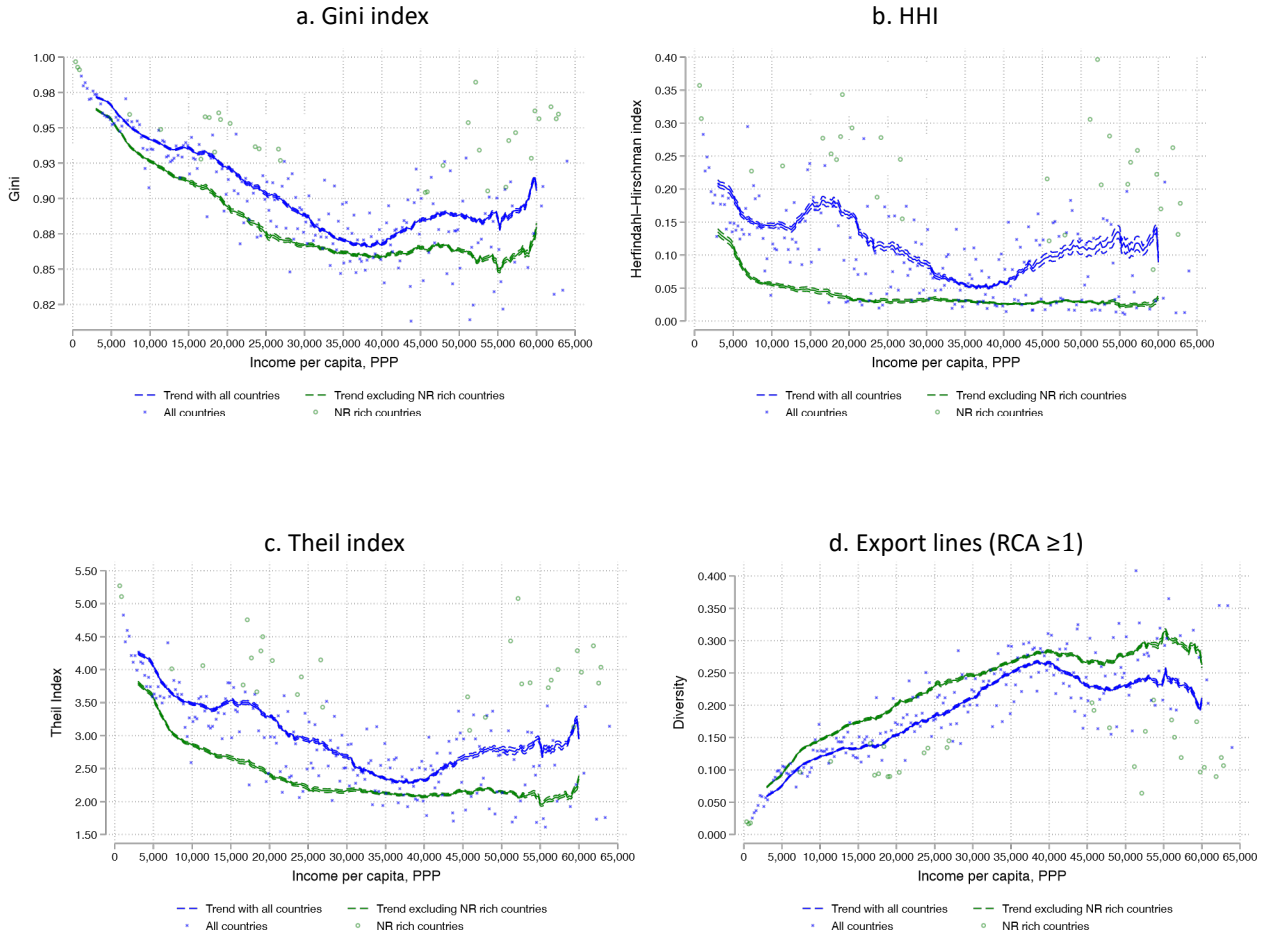
This analysis is performed with export data. Although the analysis yields robust results using the UNIDO data, the differences between the full sample and the sample excluding countries rich in natural resources are much less pronounced because there are few countries in the original sample of UNIDO data used by Imbs and Wacziarg (2003) that would qualify as rich in natural resources according to this paper's classification.

Figure 1 compares the relationship between export concentration and income for all countries and for non-natural resource–rich countries in the data. Panel a presents the results using the Gini index. Panel b uses the Herfindahl-Hirschman Index. Panel c uses the Theil index. Panel d uses the count of active export lines with an RCA equal to or greater than one.

[5] Both indicators of intensity of natural resources are calculated as averages over the period 1995–2020. Exports of natural resource products follow the list defined by Hausmann et al. (2014) using the SITC Rev. 2 classification.

The main takeaway from this exercise is that, when countries that are rich in natural resources are excluded from the sample, the U-shaped relationship becomes an L-shaped relationship: there is no pattern of respecialization in the data for countries transitioning from middle-income to high-income status. As such, countries rich in natural resources are driving the positive relationship between income and concentration at the highest levels of income.

Figure 1. The relationship between income and concentration, with and without countries rich in natural resources



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥ 1) in panel d. The blue line plots the relationship using all countries, whereas the green line plots the relationship excluding countries that are rich in natural resources (defined as countries ranking above the 75th percentile in either natural resource rents or natural resource exports as a share of GDP, on average for the period 1995 to 2020). The dots in the figures correspond to the observations upon which the nonparametric form is calculated. The green circles represent countries that are rich in natural resources, whereas the blue crosses represent countries that are not rich in natural resources. NR = natural resource; PPP = purchasing power parity.

This becomes clearer when focusing on the dots visualized in the figure, which correspond to the observations upon which the nonparametric form is calculated. The green circles represent countries that are rich in natural resources, while the blue crosses represent countries that are not rich in natural resources. These visualizations help reveal that on the extremes of the income distribution, particularly among the richest countries, there is an important presence of countries that are rich in natural resources that also have much more concentrated export baskets, driving the respecialization pattern among richer countries.

An important discussion to have regarding the importance of this distinction with respect to the original results of Imbs and Wacziarg (2003) is that in their paper they calculate the concentration of production only using manufacturing sectors. As such, it follows that the empirical finding limited to the manufacturing sector only should not be affected by excluding countries that are rich in natural resources. However, as shown by Bahar and Santos (2018), countries rich in natural resources tend to be more concentrated even with respect to their non–natural resource export baskets. This paper shows that in the context of the relationship between income and concentration, this also matters (see figure C3 in appendix C). In particular, even when natural resource lines are excluded from the calculation of the concentration measures, the relationship between per capita income and concentration is flatter at higher incomes for the sample that excludes product lines rich in natural resources than the sample that includes all lines.

The Role of Levels of Aggregation of the Data

The level of disaggregation of the data can play an important role in explaining the documented patterns. Why? Because even if respecialization occurs at higher levels of per capita income, it might only occur in highly aggregated sectors. For instance, East Asian countries are concentrated in a few clusters—electronics being the most prominent one, but within such clusters there is wide diversification. Thus, the limitations of data (as well as the conceptualization of what a sector actually is) might offer a reinterpretation of the results.

The analysis explores this issue by reestimating the nonparametric relation between the different concentration indexes and per capita income, this time using different levels of data disaggregation in terms of export lines.

This is another important reason why the analysis in this section is focused on exports instead of production, given that production data do not have enough detail beyond sectors comparable across countries, whereas export data do.

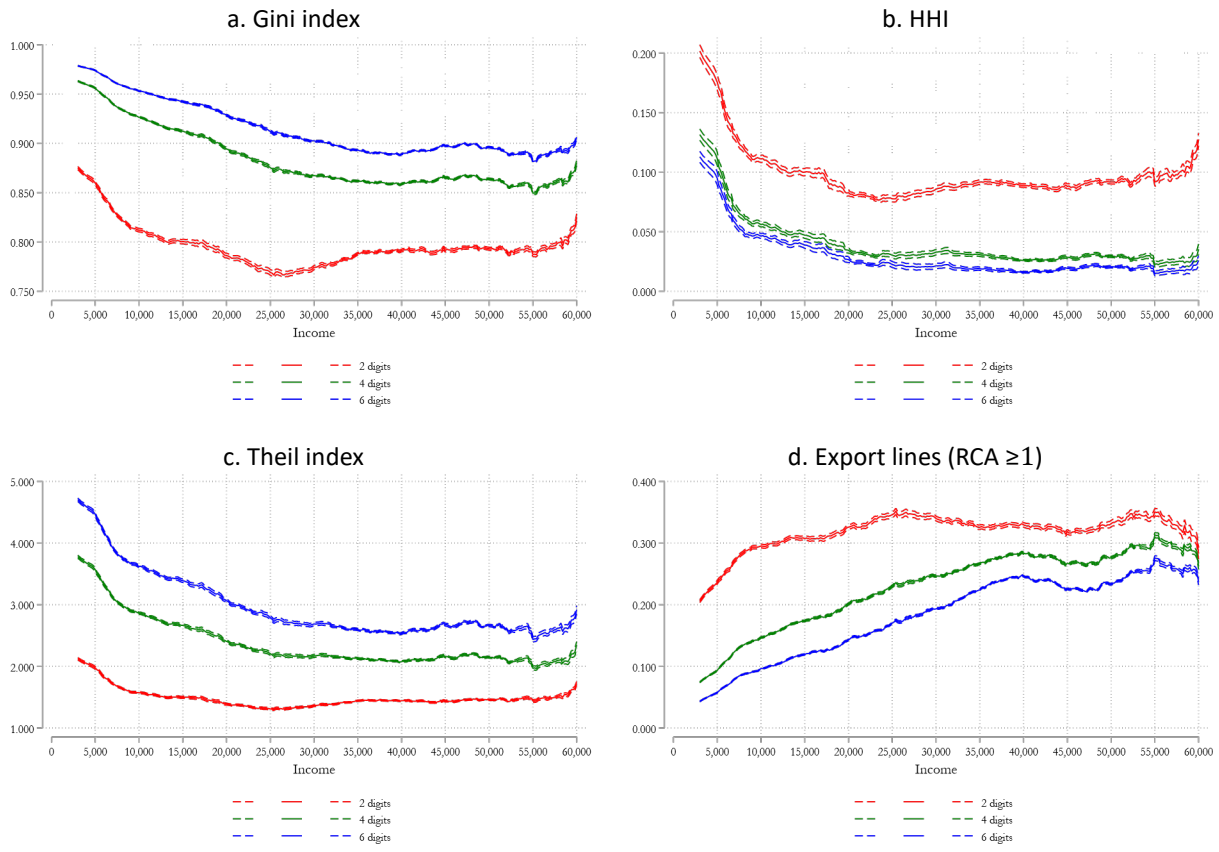
For this exercise, the use of the Harmonized System export data is crucial because it allows us to estimate that relationship using concentration measures computed based on 2-digit, 4-digit, and 6-digit codes.[6] To illustrate what this disaggregation means, consider HS 2-digit code 61, which corresponds to “Articles of clothing and clothing accessories, knitted or crocheted.” HS 4-digit code 6101 corresponds to “Men’s or boys’

[6] Going beyond 6-digits would not be appropriate for this exercise because when countries report beyond this level, there is much less certainty that the data are comparable across countries.

overcoats, carcoats, capes, cloaks, anoraks (including ski-jackets), windbreakers and similar articles, knitted or crocheted." Furthermore, the HS 6-digit code 610130 specifies that it refers to clothing made out of “man-made fibers." Given the level of specificity that is achieved with more disaggregation, it is crucial to understand whether the relationship between income and concentration is affected by the level of specificity when computing concentration.

Figure 2 presents the relationship between income and concentration per capita using export data disaggregated at the 2-digit, 4-digit, and 6-digit levels based on the Harmonized System. The figure is based on the sample of countries that excludes countries rich in natural resources (defined as having either rents or exports of natural resource both as a share of GDP that rank above the 75th percentile, on average for the period of 1995 to 2020). Across the board, the results show that the respecialization pattern at higher levels of income is more pronounced (certainly for the HHI and active export lines panels) the more aggregated the data are that are used to compute the different indexes.

Figure 2. Income and concentration, using different levels of disaggregation



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥1) in panel d. The blue line plots the relationship using concentration metrics calculated with export data disaggregated at the 2-digit level based on the Harmonized System, whereas the red line is based on 4 digits, and the green line is based on 6 digits. Estimates are based on a sample that excludes countries rich

in natural resources, defined by having rents or exports of natural resources as a share of GDP that rank above the 75th percentile, on average for the period of 1995 to 2020.

A more in-depth analysis of the data validates the visualizations. Table 1 summarizes the level of concentration at different levels of the income distribution: a low level of $2,000 GDP per capita (which corresponds approximately to Uganda's current per capita income or the 10th percentile of the income distribution), and a high level of $50,000 GDP per capita (which corresponds to approximately the income of Germany, or the 90th percentile of the income distribution).[7] In essence, this table aims to compare the two edges of the nonparametric estimation, as well as provide information about the inflection point, which is the level of income at which the nonparametric relationship between income and concentration reaches its lowest concentration level. The table also shows the information for the estimations at different levels of disaggregation (2, 4, and 6 digits based on the Harmonized System) and for all the concentration measures (Gini index, HHI, Theil index, and active export lines). It also summarizes the results when using the sample with all countries (right-hand side of the table) and for the sample excluding countries rich in natural resources, using the exclusion rule previously described (left-hand side of the table).

For example, the left-hand side of the table shows that at the low-income level using the Gini index, the concentration is 0.86 at 2 digits, 0.96 at 4 digits, and 0.97 at 6 digits, whereas for the high level of the income distribution, the Gini index is 0.79 at 2 digits, 0.86 at 4 digits, and 0.88 at 6 digits.

The results of the table yield several takeaways. First, the inflection point—the level of income at which countries move toward respecialization as their income grows—changes significantly depending on the level of disaggregation used to compute the concentration indexes. This is particularly the case when looking at the sample of non-natural resource–rich countries. For instance, when looking at the HHI at 2 digits for non-natural resource–rich countries, the inflection point is $23,400, whereas at 6 digits it more than doubles to $55,000. That means that when looking at 6 digits as compared to 2 digits, diversification is the common pattern seen in countries that evolve from low- to middle- and even high-income status, to levels above the 90th percentile of the income distribution. This result is robust in all concentration measures.

The second important takeaway of this table is that the respecialization pattern noticed in the nonparametric estimation between income and concentration is, indeed, much weaker at higher levels of disaggregation (consistently with the shift of the inflection point). For instance, the ratio of concentration between the high income and low income levels in the table becomes smaller when moving from 2 digits to 6 digits for three out of the four concentration measures: HHI, Theil index, and the number of active export lines. That is, when looking at the HHI, at high income levels there is a reconcentration that corresponds to 52 percent of the level of concentration of low income levels at the 2-digit level, but at the 6-digit level this ratio is 0.22. These calculations provide more evidence that the U-shaped relationship between income and concentration is significantly less pronounced on the right tail of the income distribution than established in previous studies, the more disaggregated the data are.

[7] The results are robust to using lower or higher percentiles for comparison points.

Table 1. Income and concentration, using different levels of disaggregation

		(1) Excluding natural resource–rich countries			(2) All countries		
		2-digit	4-digit	6-digit	2-digit	4-digit	6-digit
Gini index							
Low	$2,000	0.87	0.96	0.98	0.90	0.97	0.98
High	$50,000	0.79	0.86	0.90	0.83	0.89	0.92
Inflection point		$25,3000.77	$55,0000.85	$55,0000.88	$32,1000.80	$38,5000.87	$38,5000.89
High/Low		0.91	0.90	0.92	0.92	0.91	0.93
Low/Inflection		1.14	1.13	1.11	1.13	1.12	1.10
High/Inflection		1.03	1.02	1.02	1.04	1.03	1.02
HHI							
Low	$2,000	0.20	0.13	0.11	0.28	0.21	0.19
High	$50,000	0.09	0.03	0.02	0.20	0.11	0.10
Inflection point		$23,4000.08	$55,3000.02	$55,0000.01	$35,4000.12	$35,4000.05	$35,4000.04
High/Low		0.45	0.22	0.18	0.71	0.53	0.55
Low/Inflection		2.64	5.95	7.63	2.28	4.22	4.92
High/Inflection		1.20	1.33	1.39	1.62	2.24	2.71
Theil index							
Low	$2,000	2.12	3.78	4.70	2.44	4.26	5.26
High	$50,000	1.46	2.15	2.66	1.86	2.78	3.43
Inflection point		$25,3001.30	$55,3001.96	$55,0002.44	$35,4001.57	$38,0002.28	$38,0002.78
High/Low		0.69	0.57	0.57	0.76	0.65	0.65
Low/Inflection		1.64	1.92	1.93	1.56	1.86	1.89
High/Inflection		1.12	1.09	1.09	1.19	1.22	1.23
Active export lines (RCA>=1)							
Low	$2,000	0.21	0.07	0.04	0.17	0.06	0.03
High	$50,000	0.33	0.28	0.23	0.28	0.23	0.19
Inflection point		$25,4000.35	$55,0000.31	$55,0000.28	$38,4000.32	$38,4000.27	$38,4000.23
High/Low		1.61	3.73	5.42	1.60	3.89	5.87
Low/Inflection		0.59	0.24	0.16	0.55	0.22	0.14
High/Inflection		0.94	0.88	0.85	0.88	0.86	0.83

Source: Original calculations for the *World Development Report 2024.*

Note: The table summarizes the estimated level of concentration at different levels of the income distribution: a low level of $2,000 GDP per capita (which corresponds to the 10th percentile of the income distribution), and a high level of $50,000 GDP per capita (which corresponds to approximately the 90th percentile of the income distribution). It presents information for estimations at different levels of disaggregation (2, 4, and 6 digits based on the Harmonized System) and for all concentration measures (Gini index, HHI, Theil index, and active export lines), including the concentration value as well as the inflection point (the lowest point of concentration across the income distribution for each estimation). The right-hand side of the table presents information for estimations using the sample with all countries, and the left-hand side uses the sample that excludes countries rich in natural resources (defined as countries ranking above the 75th percentile in either natural resource rents or natural resource exports as a share of GDP, on average for the period 1995 to 2020). HHI = Herfindahl-Hirschman Index; RCA = revealed comparative advantage.

When looking at export concentration computed using 2-digit categories, countries in the 1st percentile of the income distribution have, on average, an HHI of 0.23. Countries in the median of the distribution have an HHI of 0.135, while countries in the 99th percentile of the distribution have an HHI of 0.185, implying there is respecialization at this level. The level of concentration for countries in the 99th percentile of the income distribution is about 53 percent of that in the 1st percentile. A similar pattern occurs with non-natural

resource–rich countries. The respecialization in the upper end of the income distribution reaches levels are about 84 percent of the concentration in the bottom of the distribution.

What are the implications of these results? They imply that while high-income countries tend to concentrate in particular sectors, they remain highly diversified in terms of industries or products within those sectors—much more so than low-income countries. The results confirm that as more disaggregated data are used to compute the level of concentration, the respecialization pattern documented by both Imbs and Wacziarg (2003) and Cadot, Carrère, and Strauss-Kahn (2011) weakens and often disappears. Thus, it might be the case that the respecialization pattern previously documented, in fact, reflects the pattern that the process of income growth is associated with the development of highly diversified clusters of economic activity.

An important consideration for this exercise is whether this result is driven, to some extent, artificially by different disaggregation levels of 2-digit industries more prevalent in high-income countries than in low-income countries. That is, if low-income countries are, for instance, concentrated in 2-digit sectors—such as agriculture—that the Harmonized System disaggregates into fewer 6-digit industries than other 2-digit sectors more prevalent in higher-income countries, such as manufacturing, then the results would be driven by this irregularity in the data. Appendix D discusses this in detail and shows that this irregularity does not affect the main findings.

Exploring the Transition from Low-Income to Middle-Income to High-Income Status

Whereas the preceding estimation takes into account only within-country variation, the shape of the curve might not necessarily reflect the path from low-income status to middle-income status and to high-income status for all countries. In particular, the nonparametric estimation only requires enough countries—and several observations for each country—along certain ranges of the income distribution.

Thus, this subsection presents findings from an exploratory exercise that aims to understand how common the transition is from low-income to middle-income and to high-income status in the data, and how that transition looks in the concentration-specialization space, with the inflection point of the U-shaped curve as the benchmark.

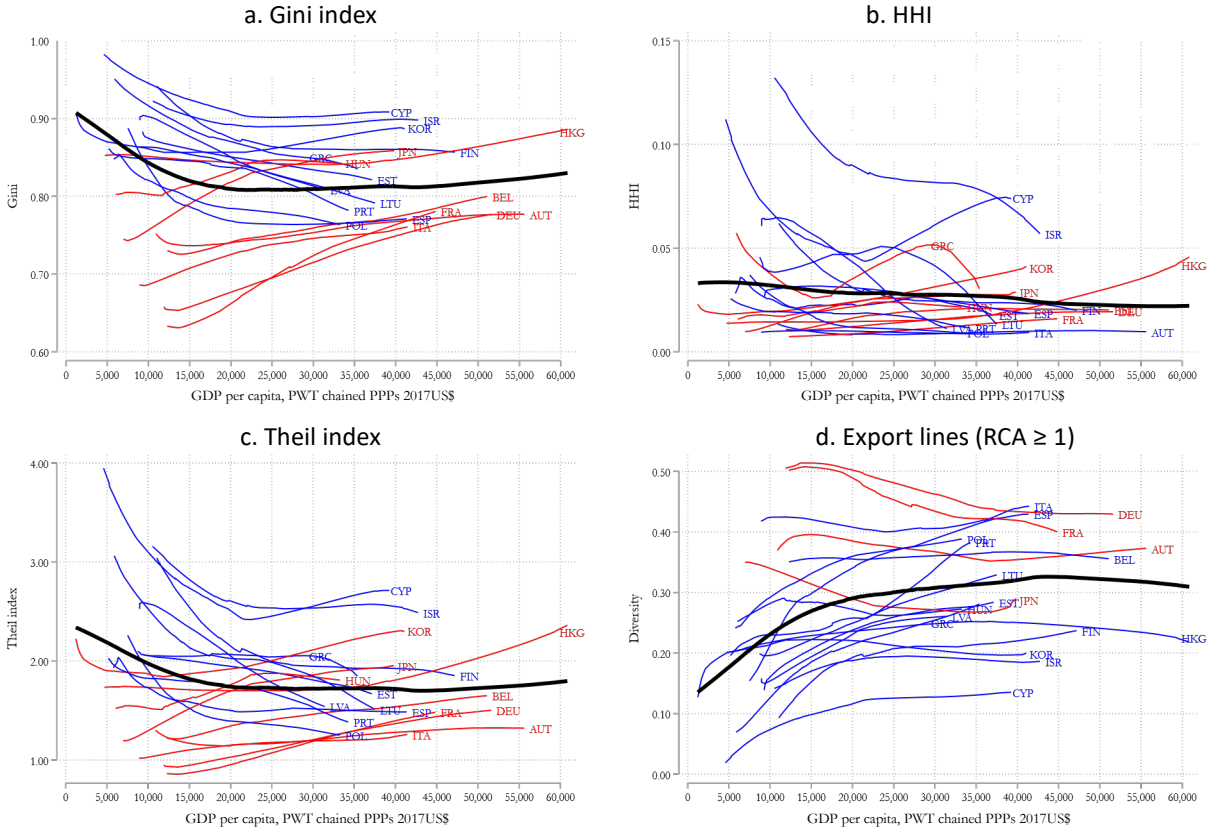
For this exercise to be meaningful, the HS export data present an important limitation given their relatively short time range (from the mid-1990s until the present) to study income dynamics. Hence, the focus here is on the main results using exports from the same source—UNCOMTRADE and Hausmann et al. (2014), but using a different classification, the Standard Industry Trade Classification (SITC), which has data from 1964 to the present.[8]

[8] The downside of the SITC classification is that it disaggregates products up to 4 digits, and thus is not as suitable as the Harmonized System to study in detail the results regarding disaggregation in the previous subsection. Nonetheless, the disaggregation results are robust using the SITC dataset, too.

A look at the data reveals that in the 60-year period from the 1960s to 2020, only a handful of countries—23—transitioned income levels from below the 50th percentile to above the 75th percentile. These economies are Austria; Belgium; Cyprus; Estonia; France; Finland; Germany; Greece; Hong Kong SAR, China; Hungary; Israel; Italy; Japan; Republic of Korea; Latvia; Lithuania; the Netherlands; Poland; Portugal; Slovak Republic; Spain; Trinidad and Tobago; and the United Kingdom.

Figure 3 plots the trajectory of these countries in terms of income and diversification of their export baskets with respect to the four diversification measures used in the previous sections: the Gini index, HHI, the Theil index, and the number of active export lines. The blue lines indicate the trajectories of those countries that ended up with more diversified export baskets. The red lines indicate the trajectories of countries that ended up with more specialized export baskets.

Figure 3. Trajectories of countries using different measures of concentration



Source: Original calculations for the *World Development Report 2024.*
Note: The panels plot nonparametric trends (lowess) for each country that transitioned from income (GDP) per capita of less than US$13,000 to more than US$31,000 (50th and 75th percentile, respectively, in year 2019). In addition, in each panel, the nonparametric trend is shown by the black line. Each panel splits the sample of countries between those that diversified (shown in blue: that is, a country's final trend is more diversified than its starting point) or those whose production became more concentrated (shown in red). The figure uses International Organization for Standardization (ISO) country codes. PWT = Penn World Table; PPP = purchasing power parity; RCA = revealed comparative advantage.

The figure presents a number of interesting facts. First, there is no clear trend: about half the countries in that sample diversify their export baskets as they become richer, whereas the other half become more specialized as their income grows, according to all the four measures.[9]

While this examination is not definitive, it does challenge the view that there are standard "stages of diversification" broadly speaking: the sample of countries that transition from middle-income to high-income status is small, and among those, there is no clear pattern of export basket concentration.

Sophistication

One aspect that remains unexplored in this setting is *what type* of product lines—rather than *how many*—are associated with a country's growth trajectory. This line of inquiry follows the empirical evidence shown by Hausmann, Hwang, and Rodrik (2007), who find that there is a robust empirical relationship between the sophistication of products that compose a country's export baskets and their levels of per capita income.

The analysis next brings this test into our setting to study for the countries that actually transitioned from low-income to middle-income and then to high-income status. To do so, the analysis uses three different proxies of product sophistication.[10]

- The first is the share of products in their export basket that can be classified as "differentiated"—as opposed to homogenous or reference-priced goods, using the classification by Rauch (1999).
- The second is the weighted average of the Product Complexity Index (PCI) of the export basket (Hausmann et al. 2014).
- The third is the weighted average of the trade elasticity of substitution of a country's exports, using Broda and Weinstein (2006). If the elasticity is high (low), it means consumers can easily switch from one good to another when the price of the first good rises (drops) relative to the second. In essence, elasticity is high for products that are less specialized and less substitutable for other products in the market.

The results for this exercise are shown in figure 4. It plots the transition of all countries identified in figure 3 (the only countries that transition from below the 50th to above the 75th percentile of the income distribution since the mid-1960s until 2020). The figure uses the same color legend as figure 3 for consistency: Blue represents countries that diversified their export baskets during this transition, while red represents countries that experienced further concentration in their export baskets.

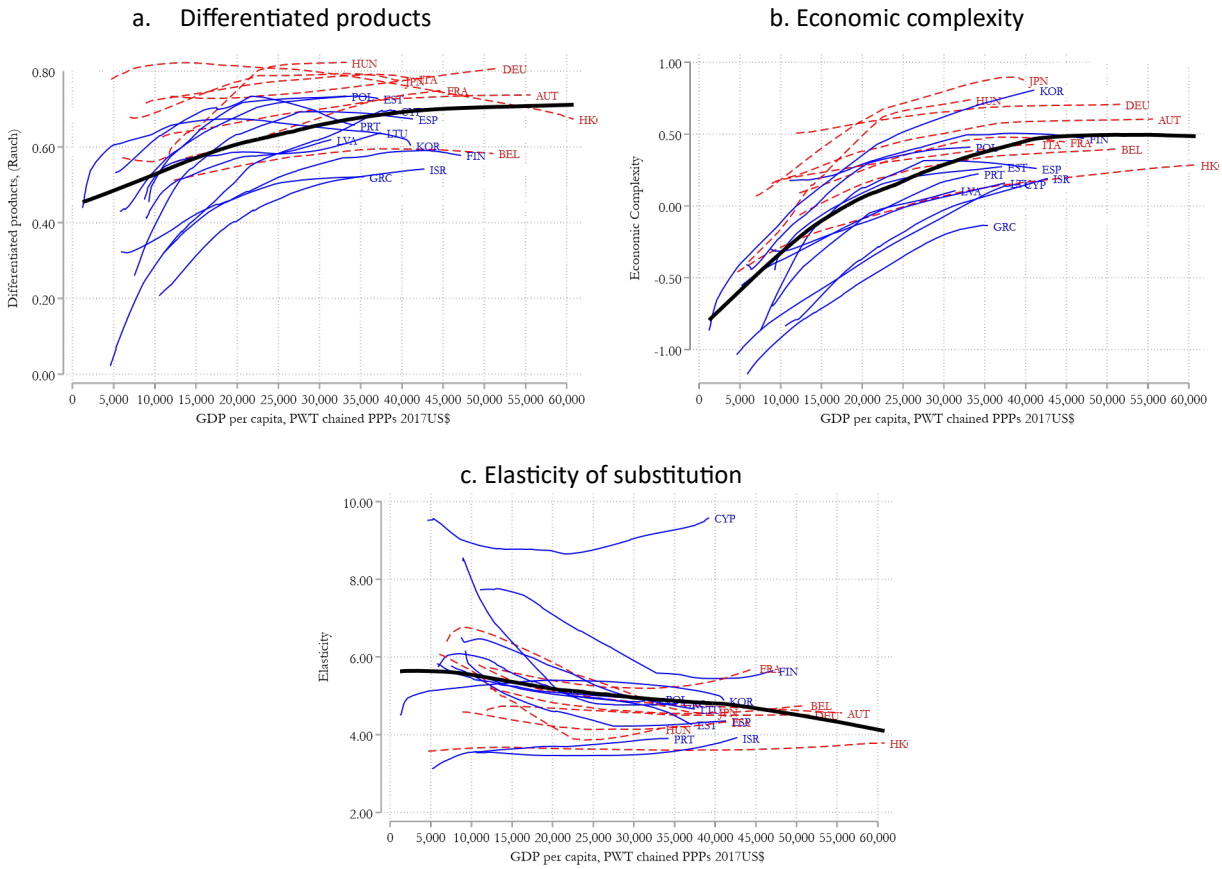
Using all measures of sophistication, a robust increasing monotononic relationship between sophistication and income is apparent for all countries that made the transition, whether their export baskets became more or less diversified during their trajectory.

---

[9] There seems to be some sort of convergence to an average level of diversification: countries that start highly specialized, diversify; and countries that start highly diversified, specialize.

[10] The analysis refrains from using the income/productivity level (PRODY) computed by Hausmann, Hwang, and Rodrik (2007) simply because it uses income as input, to avoid results being driven by an obvious circular relationship.

Figure 4. Sophistication of export baskets



Source: Original calculations for the *World Development Report 2024.*
Note: The panels plot nonparametric trends (lowess) for each country that transitioned from GDP per capita of less than US$13,000 to more than US$31,000 (50th and 75th percentile, respectively, in year 2019). Each panel plots the nonparametric relationship between income and three variables (shown in black line): in panel a, the share of a country's export basket that can be classified as differentiated products; in panel b, the weighted average of the Product Complexity Index of its exports; and in panel c, the weighted average of the trade elasticities of its exports. Each panel plots the sample of countries between those that diversified according to the Gini index (shown in blue/solid line—that is, a country's final trend is more diversified than its starting point) and those whose production became more concentrated (shown in red/dashed). The figure uses International Organization for Standardization (ISO) county codes. PPP = purchasing power parity; PWT = Penn World Table.

Linking Income and Diversification via a Ricardian Model

To better understand the link between diversification and income, the analysis next builds on a neo-Ricardian model of international trade, in which the industrial composition and income of a country are primarily determined by differences in productivity between countries. Hence, both income and diversification are driven by the changes in productivity of countries in different industries. The analysis uses a simplified version of the model in Costinot, Donaldson, and Kumunjer (2012), a multiindustry extension of celebrated model of Eaton and Kortum (2002). Within this simplified setup, the analysis shows how productivity changes lead to changes in concentration, as measured with HHI, and income.

This model assumes that there are many countries indexed by c, and many goods indexed by i, with each good consisting of an infinite number of varieties, ω. Labor is the single factor of production, which is assumed to be freely mobile across industries but not mobile across countries. Let L_c denote the fixed supply of labor in country c, with L_{ci} workers employed in industry i in country c. The labor employed in variety ω of this industry is given by $L_{ci}(\omega)$.

Production of variety ω in each industry i is linear in the number of workers assigned to the industry, and it is given by:

$$y_{ci}(\omega) = z_{ci}(\omega)L_{ci}(\omega),$$

where $z_{ci}(\omega)$ is the productivity level. It is assumed to be a random variable drawn independently for each triplet (c, i, ω) from a Frèchet distribution, whose cumulative distribution function, $F_{ci}(.)$, follows:

$$F_{ci}(c) = \exp\left[-\left(\frac{z}{z_{ci}}\right)^{-\theta}\right],$$

where $z_{ci} > 0$ is the location parameter that governs the productivity level, and $\theta > 1$ governs the intra-industry heterogeneity.

The wage for workers in country c is w_c, and is the result of assigning workers across industries and varieties until the labor markets clear. To simplify the discussion and without loss of generality, the model assumes that there are no trade costs. The heterogeneity in costs—that is, $w_c/z_{ci}(\omega)$—generates scope for cross-industry Ricardian comparative advantage. In any country m, the price of variety ω in industry ω, $p_{i,m}(\omega)$, is determined by the minimum cost producer of this variety globally:

$$p_{i,m}(\omega) = \min\left[\frac{w_c}{z_{ci}(\omega)}\right].$$

The demand of households in the country m is governed by a two-tier consumption utility function (nested-CES utility function). In the upper level, consumers decide what share of their income they will spend in industry i. It is assumed that these shares, denoted by α_i, are the same in all countries. In the lower tier, consumers decide among the varieties, with a constant elasticity of substitution function with elasticity σ_i. Therefore, the expenditure of country m in industry i is given by:

$$x_{i,m}(\omega) = \left[\frac{p_{i,m}(\omega)}{p_{i,m}}\right](\alpha_i w_m L_m) \quad \text{with} \quad p_{i,m} = \left[\int p_{i,m}(\omega)^{1-\sigma_i}\right]^{1/(1-\sigma_i)}$$

where $p_{i,m}$ is the CES price index, $\sigma_i < (1 + \theta)$ is the elasticity of substitution, and $0 \leq \alpha_i \leq 1$ is the Cobb-Douglas share.

Due to the uniform price assumption, the price index for industry i is the same between countries. With these assumptions, and by using the properties of the Frèchet distribution, the exports of country c in industry i can be written as:

$$x_{ci} = \sum_m \frac{(w_c/z_{ci})^{-\theta}}{\sum_{c'}(w_{c'}/z_{c'i})^{-\theta}} \alpha_i w_m L_m = \sum_m \frac{(w_c/z_{ci})^{-\theta}}{\sum_{c'}(w_{c'}/z_{c'i})^{-\theta}} \alpha_i \mathrm{GDP}_W,$$

where $\mathrm{GDP}_W = \sum_m w_m L_m$ is the total world GDP. Assume that each country is small relative to the world such that:

$$\frac{\alpha_i \mathrm{GDP}_W}{\sum_{c'}(w_{c'}/z_{c'i})^{-\theta}} \equiv \Phi_i \quad \Longrightarrow \quad x_{ci} = (w_c/z_{ci})^{-\theta}\Phi_i,$$

Φ_i is constant for a small change in each country's increase in the productivity. With this simplification, the share of industry i in country c's production can be written as:

$$s_{ci} = \frac{x_{ci}}{\sum_{i'} x_{ci'}} = \frac{(w_c/z_{ci})^{-\theta}\Phi_i}{\sum_{i'}(w_c/z_{ci'})^{-\theta}\Phi_{i'}} = \frac{z_{ci}^{\theta}\Phi_i}{\sum_{i'} z_{ci'}^{\theta}\Phi_{i'}}.$$

Given the shares, changes in the concentration of production/exports can be related to changes income relative to a shock in the productivity of an industry with the following proposition:

Proposition 1. *For a small increase in the productivity of industry j in country c, the change in HHI is given by:*

$$\frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\log z_{cj}} = 2\theta s_{cj}\left(s_{cj} - \mathrm{HHI}_c\right),$$

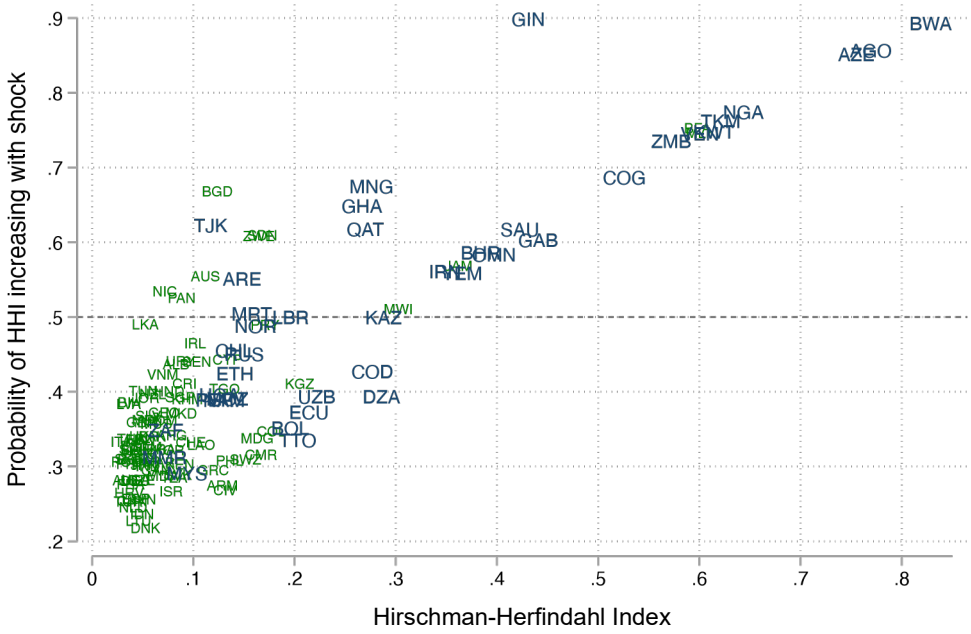
and change in wage (income) is given by:

$$\frac{\mathrm{d}\log w_c}{\mathrm{d}\log z_{cj}} = \frac{\theta}{1+\theta} s_{cj}.$$

The proof of the proposition is given in Appendix E. Intuitively, the proposition implies that any technological shock results in an increase in income. But the income increase is largest if the shock affects the dominant industries of the country. On the other hand, concentration, as measured by HHI, will increase only if a productivity shock occurs in an industry j such that $s_{cj} > \mathrm{HHI}_c$.

To quantify how likely this is to happen, a back-of-envelope calculation is performed using current export data. The exercise is to randomly select a product with a probability proportional to the product's weight in the export basket of a country, and then determine whether HHI increases. Assuming that shocks are correlated with industry size, assume that research and development (R&D) efforts and management enhancements are more intensively allocated to the predominant sectors of a country. Figure 5 shows the results of the plot of dynamics derived from the model. The horizontal axis plots the initial concentration of each country's export basket. The vertical axis plots the probability of the shock resulting in increased concentration. The figure displays natural resource–rich countries in blue. Figure 5 shows that for most countries, a productivity shock results in a more diversified portfolio of exports because the likelihood of concentration is below the 50 percent threshold. Most countries above the 50 percent threshold—for which a random productivity shock is likely to increase concentration—are rich in natural resources are unusually highly concentrated to begin with.

Figure 5. Productivity shocks and concentration



Source: Original calculations for the *World Development Report 2024.*
Note: This figure plots probability of a productivity shock resulting in higher concentration (vertical axis) versus the initial concentration level (horizontal axis) for each country in the sample. Countries in blue are natural resource-rich countries and countries in green are non-natural resource–rich countries. The figure uses International Organization for Standardization (ISO) country codes. HHI = Hirschman-Herfindahl Index.

Concluding Remarks

What is the relationship between structural transformation and economic development? This paper explores this question by revisiting some of the stylized facts that have been established in the literature. Although economic theory suggests that specialization is a result of openness to trade, it is less clear empirically how

this plays out in general equilibrium, where so many other factors beyond integration play a role in this relationship.

Diversification of a country's export basket could be both a cause and a consequence of the process of economic growth and development. The evidence in this paper, however, suggests that after taking into account measurement peculiarities, respecialization is not necessarily the norm in the transition from middle-income to high-income status.

As such, it is not the level of diversification or concentration that holds the explanatory power to explain economic growth when looking at a handful of countries that have transitioned from middle-income to high-income status over the past few decades. Rather—as suggested by Hausmann, Hwang, and Rodrik (2007), the content and the sophistication of their production and export basket seems to have a clear increasing monotonic relationship with income. This relationship is not seen when exploring the dynamics of concentration or diversification of countries' export baskets.

Appendix A. Nonparametric estimation and replication of Imbs and Warcziag (2003)

To explore the relation between economic diversification and income levels, this paper follows the nonparametric “lowess" inspired method of Imbs and Wacziarg (2003), which imposes as little structure as possible. In a nutshell, the method consists of running regressions using a local sample defined over a bandwidth of income levels; and the procedure repeats, moving the window by a fixed dollar amount until completing the distribution of income level under study. The result is a figure of the fitted diversification values (and its confidence intervals) for the midpoint of each income bandwidth. As opposed to running a polynomial regression over the whole sample, this method ensures that the estimates in each income window (such as low-income) are independent of the estimates of other windows (such as high-income). Because the data used is from a panel of countries over time, each country could appear several times in each income window. To control for multiple observations of the same country, the regression includes country fixed effects. Hence, each estimate is equivalent to using a flat or rectangular weighting scheme, as opposed to the lowess method, which generally uses triangular weights.

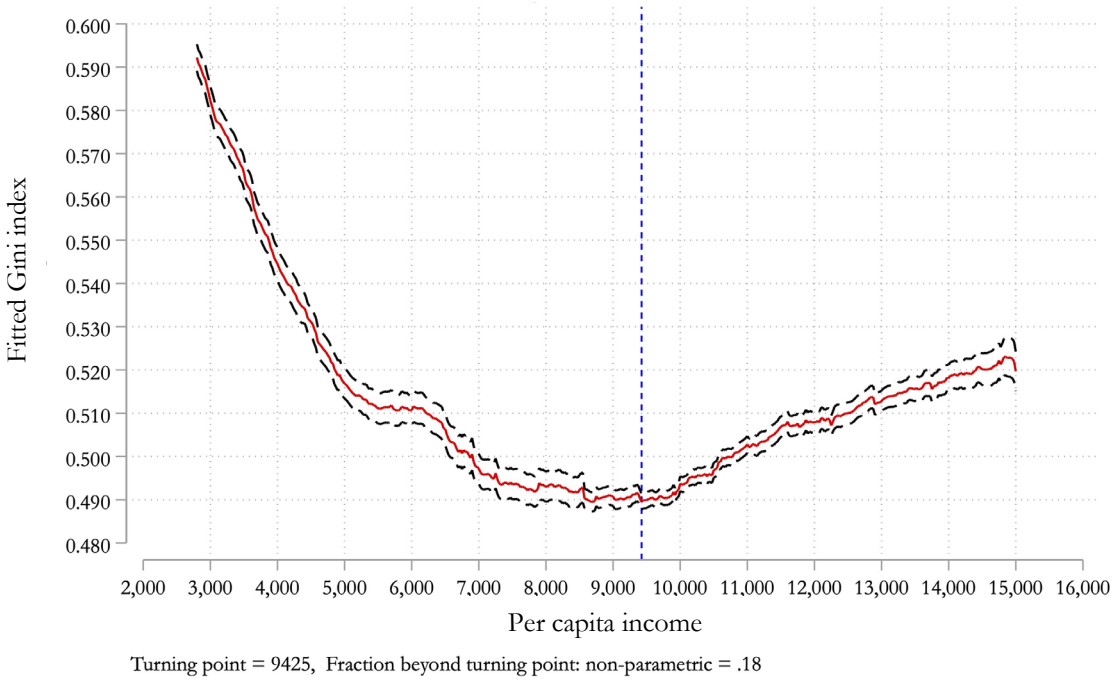
The method requires two parametric choices: the range of the income window; and the size of the step increments when moving the sample. Following the implicit decisions of Imbs and Wacziarg (2003), this analysis set the bandwidth at $8000, which is approximately equivalent to the standard deviation of income levels in the full sample. Increments of $100 dollars were set, which is approximately equivalent to the choice of $25 used by Imbs and Wacziarg (2003) once adjustments are made for purchasing parity. As Imbs and Wacziarg (2003) note, there are many ways of choosing the bandwidths and setting increments, but these do not qualitatively change the results. For instance, choosing smaller bandwidths could result in a potentially more volatile figure when the results are plotted because each estimate is dependent on a smaller local sample. On the other hand, choosing larger bandwidths and step increments would result in a smoother figure. As in Imbs and Wacziarg (2003), the parameters chosen allow this analysis to achieve the goal of recovering clear diversification patterns across income levels.

Using this methodology this analysis has been able to replicate with a high degree of precision the results of Imbs and Wacziarg (2003). Figure A1 presents the results graphically using the Gini index as a measure of concentration of countries’ production baskets (based on value added) and per capita income.

One of the features of the data is that the density of observations (country-year combinations) at low levels of income is significantly larger than at high levels of income. This is not a salient feature, as shown by the curves relating diversification and income. For this reason, to show how observations scatter over the income distribution, figure A1 adds binscatters version of the data in the background. That is, the income distribution was divided in segments of $200 and the average of each diversification measure used in the analysis was plotted in the vertical axis. Dots are differentiated depending on the proportion of the underlying observations that are classified as natural resource–rich countries or not. This procedure avoids

overcrowding the figure and simplifies the visualization. The resulting figure emphasizes the distribution of observations and the concentration of natural resource–rich countries along the income distribution.

Figure A1. Replicating the Main Result of Imbs and Warcziag (2003)



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration of the production basket (using data from the United Nations Industrial Development Organization, UNIDO) based on the Gini index, following the estimation methodology of Imbs and Wacziarg (2003) (shown in red line). The dashed black lines indicate the 95% confidence interval.

Appendix B. Deep dive into the original results of Imbs and Warcziag (2003)

After replicating the original results by Imbs and Wacziarg (2003) with a high degree of precision, this paper further examines the sensitivity of the results to different data choices.

This appendix focuses on two main issues: the treatment of missing values in the data; and the exclusion of countries with a large number of missing values. The two issues are intertwined. This appendix, unless otherwise noted, focuses on the United Nations Industrial Development Organization (UNIDO) data that report value added statistics (the findings are consistent if employment statistics are used).

Overall, missing values pose an important empirical challenge in every respect. In this context, if some countries have an unusually large number of industries with presumably missing values, then it would make sense to exclude some of those countries from the analysis due to lack of data.

However, the analysis finds that the sensitivity of the results to what constitutes a "large" number of industries with missing data is high. This could generate different results based on sample selection.

For example, in the original study, the authors include country-year observations for which there is reporting (that is, nonmissing values) for at least 27 out of 28 industry classifications used in the data. Under this condition, for example, France would be excluded from the calculations that use UNIDO data because in many instances France reports two industries with missing values: oil (ISIC 354), and nonferrous metals (ISIC 361). Thus, France reports nonzero values in a total of 26 industries—and thus falls below the threshold of 27.

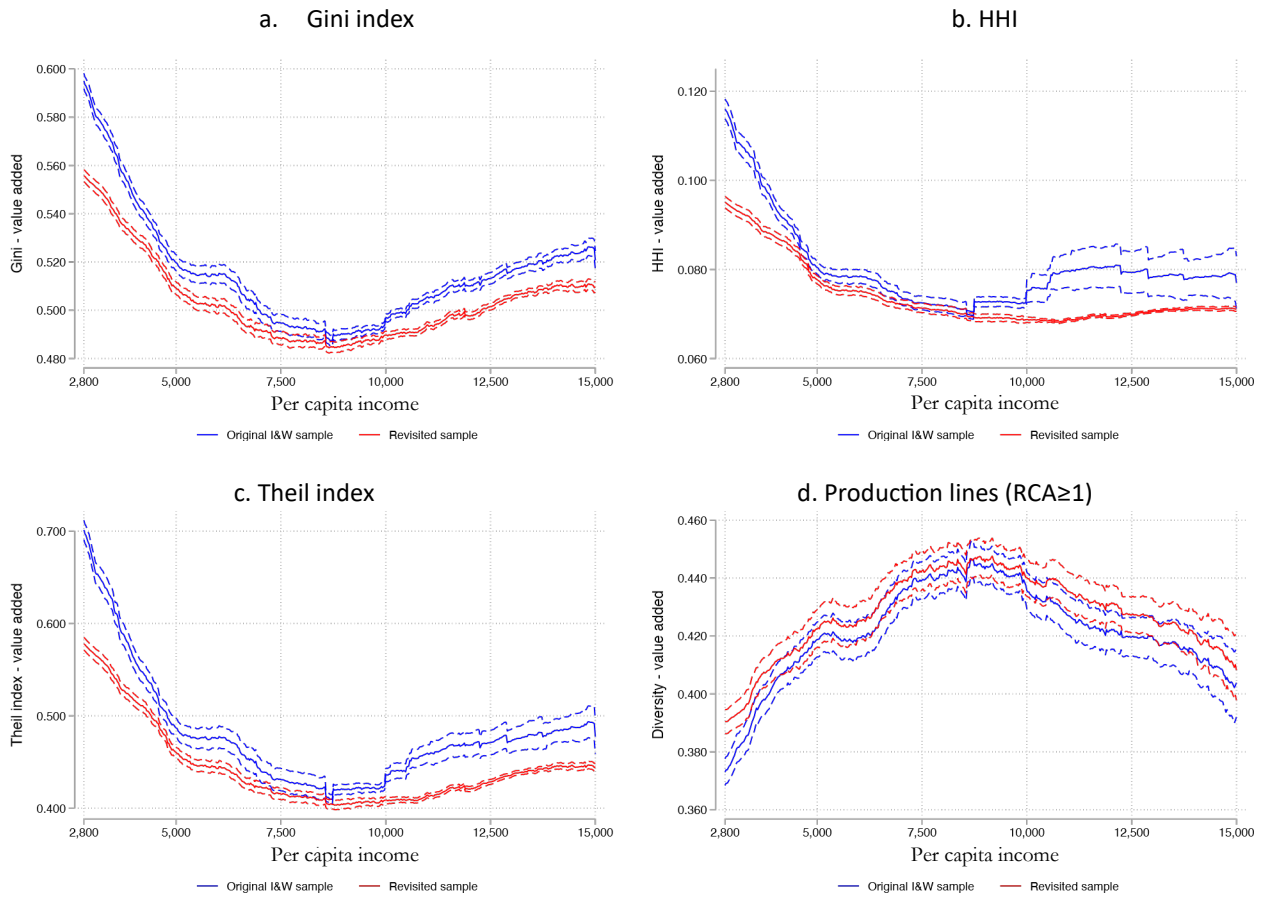
On the other hand, Argentina in 1975 does not have any missing observations. However, it has positive entries (nonzero values) for only 4 industries and zeros (indicating no production) for the remaining 24 industries. Accordingly, Argentina in 1975 would be included in the sample. However, most countries produce at least some product in every industrial classification; it is thus hard to believe that the frequent number of zeros for Argentina represent actual zeros, instead of missing values.

Given such discrepancies, the analysis performs a number of robustness checks that vary the sample of country-year observations along two margins. First, the analysis assumes that all zeros in the data indicate missing values because it is rare for a country to produce nothing at all (zero) in highly aggregated sectors; hence, zeros could simply represent a missing value. This would naturally exclude many more country-year observations than the original study (for example, Argentina in 1975 would be excluded because it reports only 4 industries with nonzero values. This paper refers to this refined sample as the "revised" sample.

Second, the analysis varies the threshold of the minimum number of industries required to have nonmissing values to include in the calculations. The original study retained this threshold at 27.

Figure B1 summarizes the results regarding the treatment of missing values. It computes the nonparametric relationship between the different concentration measures and income for the original and the revised sample. In both cases, the original rule of including country-year observations for which 27 or more industries are nonmissing is used. As the figure shows, the trend of the resided sample, while still U-shaped for the Gini index, the Herfindahl-Hirschman Index (HHI), and the Theil index (and with an inverted U-shape for production lines), is much less pronounced in both ends than the original result.

Figure B1. Original sample versus revised sample (treating zero values as missing values)



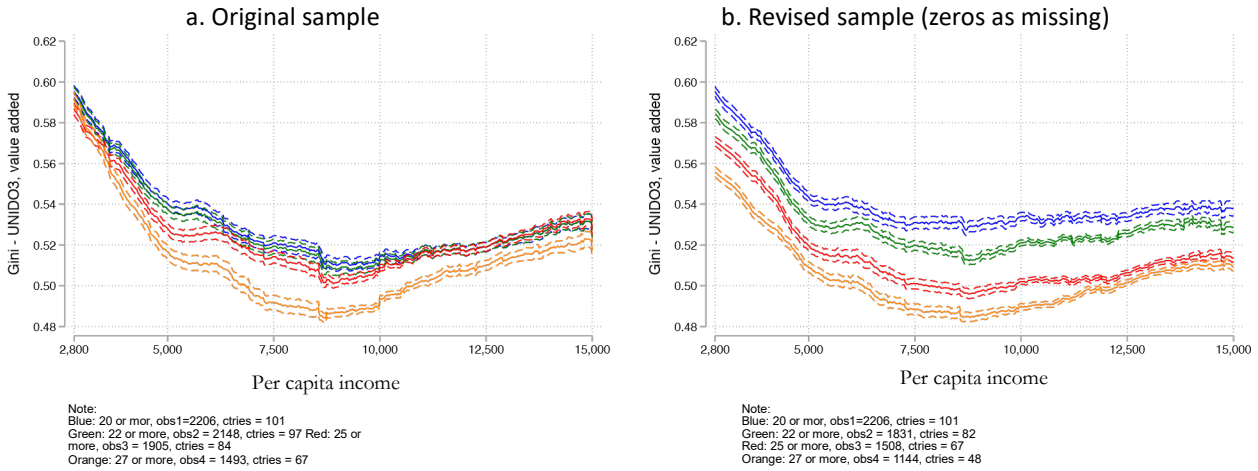
Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active production lines with revealed comparative advantage greater than or equal to one (RCA ≥1) in panel d. Following the original paper (Imbs and Wacziarg (2003), both lines include countries that have nonmissing values for at least 27 industries to compute each concentration index. The blue line presents results using the original data, while the red line presents results using a revised sample that treats zero values as missing values for the value added of industries for country-year observations. I&W = Imbs and Wacziarg.

Figure B2 summarize results regarding the exclusion of countries with a large number of missing values. In particular, it plots the nonparametric relationship between concentration (using the Gini index) and income both using the original sample (shown in panel a) and the revised sample (shown in panel b), varying the nonmissing industries threshold upon which country-year observations are included in the sample.

The lines represent different rules of exclusion of country-year observations: the original rule of 27 or more industries with nonmissing values (nonmissing industries) (in orange); a second threshold of 25 or more nonmissing industries (in red); a third threshold of 22 or more nonmissing industries (in green); and a fourth threshold of 20 or more nonmissing industries (in blue). Each plot includes the number of countries included in each variation of the threshold; naturally, the higher the threshold, the fewer countries are used. The difference between the more conservative (27) and less conservative (22) thresholds results in a significant difference in the number of countries included in the calculations, often by about one-third.

The main takeaway of figure B2 is that the choice of threshold can result in less pronounced U-shaped relationship between income and concentration. In the case of the revised sample (panel b), the U-shaped relationship is less robust overall.

Figure B2. Varying thresholds of exclusion of country-year observations based on nonmissing industries



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using the Gini index. Panel a presents plots using the original data, while panel b presents plots using a revised sample that treats zero values as missing values for the value added of industries for country-year observations. Each line represents the aforementioned computed nonparametric relationship using different exclusion thresholds for country-year observations with missing industry data needed to compute the concentration index. The four lines present the cases of countries that have nonmissing values for: at least 27 industries (as in the original paper) (in orange), 25 industries (in red), 22 industries (in green), and 20 industries (in blue).

Appendix C. Natural Resources–Rich Countries

Table C1 lists all countries considered rich in natural resources and excluded from the analysis when reestimating the relationship between income and concentration. The list includes countries that rank above the 75th percentile in terms of their 1995 to 2020 average in either rent of natural resources as a share of GDP (based on data from the World Development Indicators) or share of natural resource exports in the country's total export basket (based on data from UN COMTRADE and the classification provided by Hausmann et al. 2014).

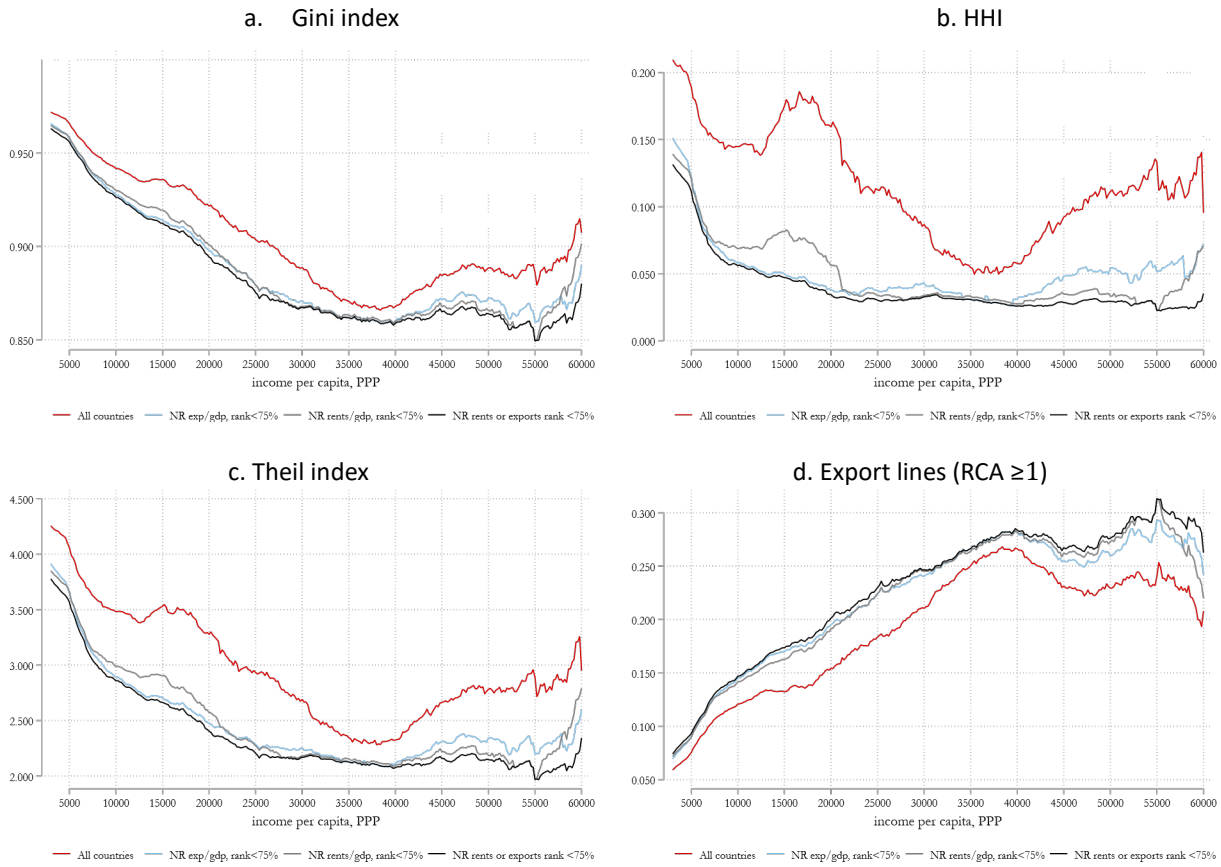
Table C1. List of Natural Resource–Rich Countries

	Natural resource rents		Natural resource exports			Natural resource rents		Natural resource exports	
Country	% GDP	Percentile	% GDP	Percentile	Country	% GDP	Percentile	% GDP	Percentile
Algeria	21.3	0.91	75.2	0.91	Liberia	21.2	0.89	31.8	0.78
Angola	34.4	0.95	96.3	0.99	Malaysia	11.2	0.76	10.4	0.75
Azerbaijan	27.2	0.93	62.1	0.92	Mauritania	11.8	0.78	54.2	0.87
Burundi	21.2	0.90	32.2	0.53	Mongolia	14.9	0.82	65.8	0.93
Bahrain	19.2	0.87	15.8	0.70	Mozambique	11.5	0.77	39.2	0.79
Bolivia	6.6	0.63	59.3	0.83	Namibia	1.8	0.41	44.6	0.84
Botswana	2.0	0.43	84.3	0.93	Nigeria	15.6	0.83	90.6	0.89
Central African Republic	10.3	0.75	38.5	0.60	Norway	7.6	0.68	56.5	0.85
Chad	18.9	0.87	49.9	0.75	Oman	35.9	0.96	78.6	0.98
Chile	7.1	0.66	48.5	0.81	Peru	5.8	0.61	53.5	0.77
Congo, Dem. Rep.	23.2	0.91	83.8	0.82	Qatar	30.9	0.95	78.9	0.97
Congo, Rep.	40.2	0.98	81.4	1.00	Russian Federation	13.1	0.80	50.0	0.81
Ecuador	10.2	0.74	38.9	0.77	Saudi Arabia	36.1	0.97	76.7	0.95
Equatorial Guinea	37.9	0.97	70.1	0.97	Sierra Leone	12.2	0.80	58.6	0.80
Ethiopia	18.1	0.86	4.3	0.08	South Africa	4.4	0.58	40.2	0.76
Gabon	28.3	0.94	80.8	0.95	Tajikistan	1.6	0.37	49.4	0.79
Ghana	11.1	0.76	38.7	0.73	Trinidad and Tobago	12.0	0.79	31.7	0.85
Guinea	14.7	0.82	84.5	0.91	Turkmenistan	40.2	0.99	59.4	0.86
Guinea-Bissau	16.9	0.85	12.8	0.50	United Arab Emirates	20.7	0.89	58.7	0.94
Iran, Islamic Rep.	24.4	0.92	74.3	0.83	Uzbekistan	16.2	0.84	30.7	0.63
Iraq	44.5	1.00	93.6	0.96	Venezuela, RB	19.7	0.88	66.9	0.88
Kazakhstan	17.6	0.85	64.0	0.90	Yemen, Rep.	25.4	0.93	79.3	0.87
Kuwait	43.0	0.99	73.4	0.99	Zambia	11.6	0.78	68.0	0.89

Source: Original calculations for the *World Development Report 2024.*
Note: The table lists all countries considered rich in natural resources in the baseline results. It includes countries that rank above the 75th percentile in terms of their rent of natural resources or natural resource exports as a share of GDP, on average for the period 1995 to 2020.

Using different concentration measures, figure C1 visualizes how results excluding natural resource–rich countries are robust to different rules for their exclusion. The red line represents the nonparametric relationship between income and concentration using all countries. The light blue line excludes countries ranking above the 75th percentile in either natural resource rents or natural resource exports both as a share of GDP and both averaged between 1995 to 2020. The grey line excludes countries ranking above the 75th percentile in terms of their share of natural resource exports as a share of GDP, averaged between 1995 to 2020. The black line excludes countries above the 75th percentile in either natural resource rents as a share of GDP, averaged between 1995 to 2020. The results show that our choice of criteria to define (and exclude) natural resource–rich countries is robust to other criteria, too.

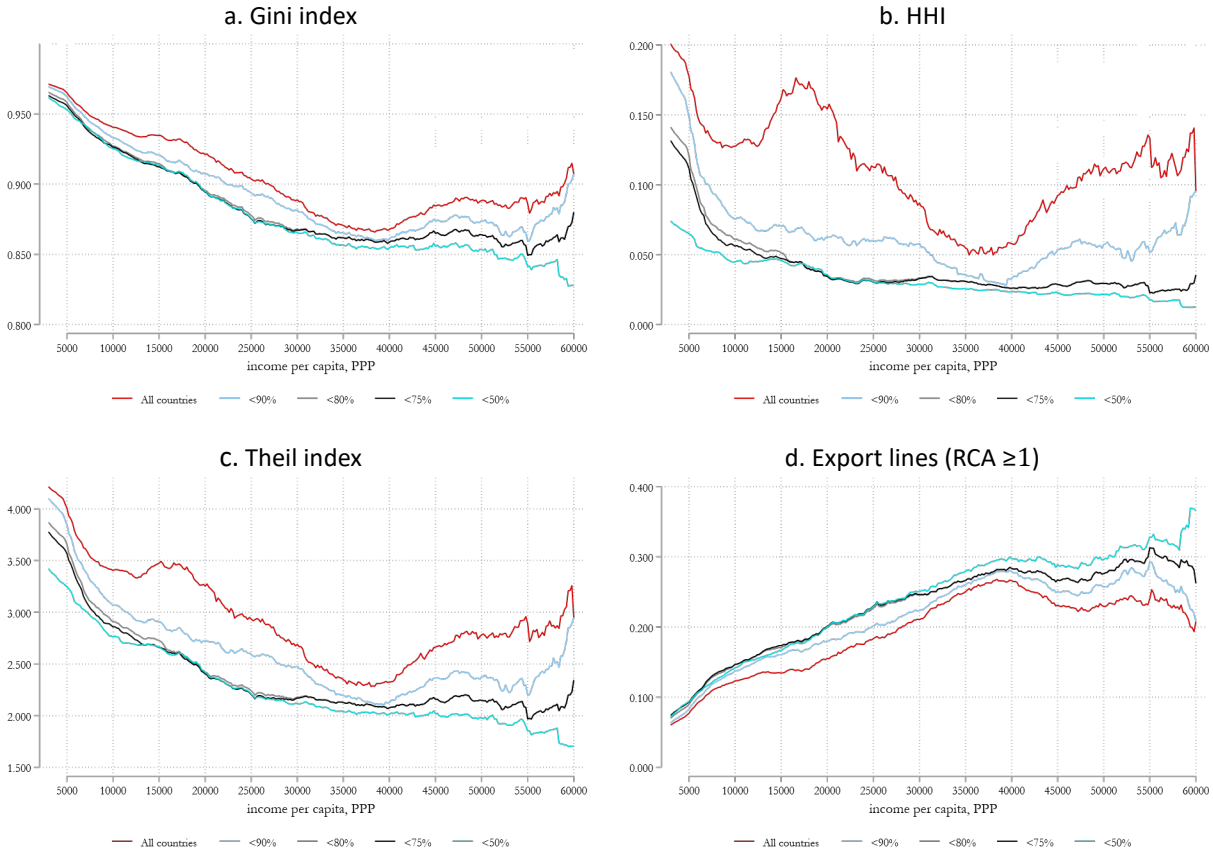
Figure C1. Varying exclusion criteria for natural resource–rich countries



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥ 1) in panel d. The red line represents the nonparametric relationship between income and concentration using all countries. The light blue line excludes countries ranking above the 75th percentile in either natural resource rents or natural resource exports as a share of GDP, averaged between 1995 to 2020. The grey line excludes countries ranking above the 75th percentile in terms of their natural resource exports as a share of GDP, averaged between 1995 to 2020. The black line excludes countries above the 75th percentile in either natural resource rents as a share of GDP, averaged between 1995 to 2020. NR = natural resource; PPP = purchasing power parity; RCA = revealed comparative advantage.

Figure C2 shows that the results are robust to using different thresholds to define countries rich in natural resources, from above the median in terms of rents or exports of natural resources as share of GDP to above the 90th percentile.

Figure C2. Varying threshold criteria for identifying natural resource–rich countries

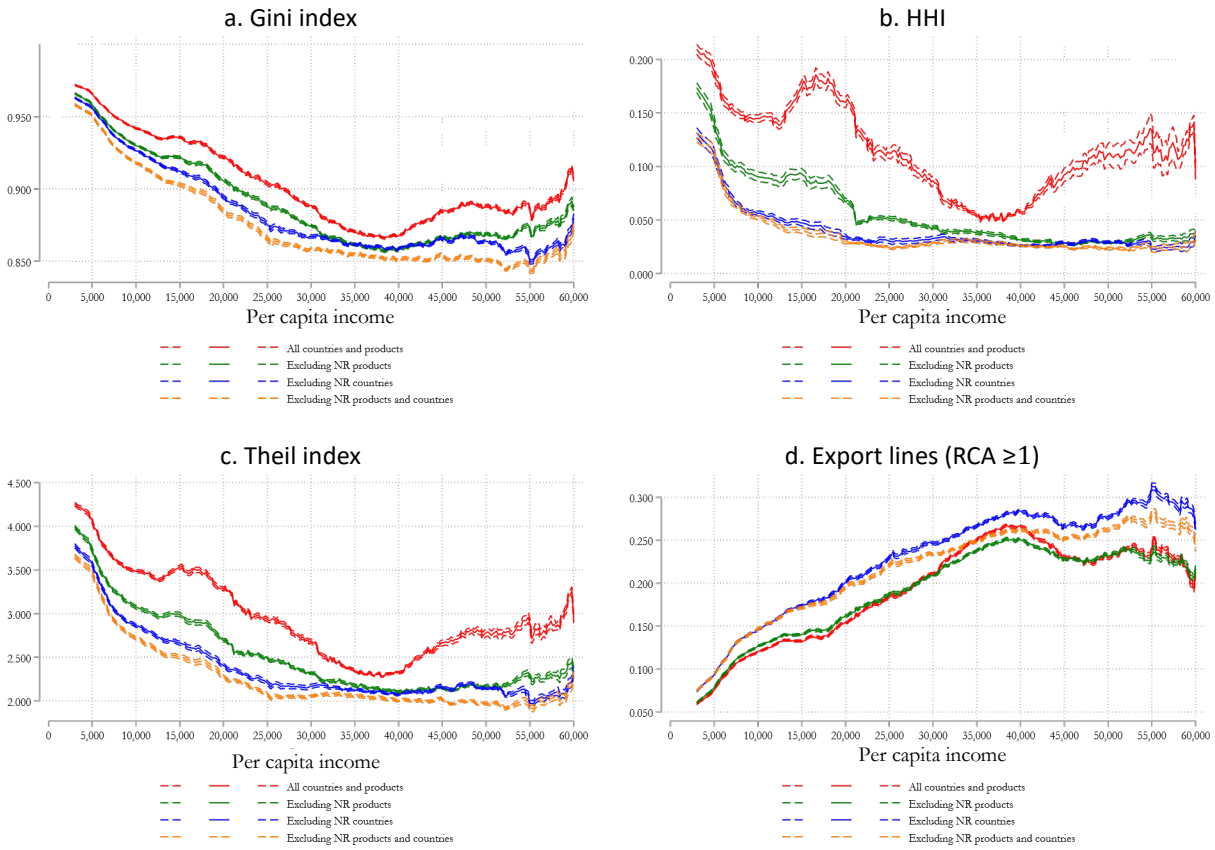


Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the proportion of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥1) in panel d. The red line represents the nonparametric relationship between income and concentration using all countries. The light blue line excludes countries ranking above the 90th percentile in either natural resource rents or natural resource exports as share of GDP, and both averaged between 1995 to 2020. The grey line excludes countries ranking above the 80th percentile. The black line excludes countries ranking above the 75th percentile (our baseline results). The blue line excludes countries above the median. Each panel uses a different concentration measure. PPP = purchasing power parity.

Finally, figure C3 presents an important empirical result that explains why it is important in this exercise not only to compute concentration measures using only manufacturing or the basket of non–natural resource products—as done by Imbs and Wacziarg (2003), but rather to exclude countries that are rich in natural resources. In the figure, each line computes plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥ 1) in panel d. The red line represents the nonparametric relationship between income and concentration using all countries. The light blue line excludes countries ranking above the 75th percentile in either natural resource rents or natural resource exports both as a share of GDP, averaged between 1995 to 2020. The grey line excludes countries ranking above the 75th percentile in terms of their natural resource exports as a share of GDP, averaged between 1995 to 2020. The black line

excludes countries above the 75th percentile in either natural resource rents as a share of GDP, averaged between 1995 to 2020. The relationship between income and concentration per capita becomes even flatter at higher levels of income when excluding countries that are rich in natural resources (blue and orange lines) than when simply excluding export lines that correspond to natural resources (green line). This is consistent with the work by Bahar and Santos (2018), who show that countries that are rich in natural resources, and thus more prone to Dutch Disease, have more concentrated non-natural resource–rich export baskets.

Figure C3. Excluding natural resource lines and natural resource–rich countries

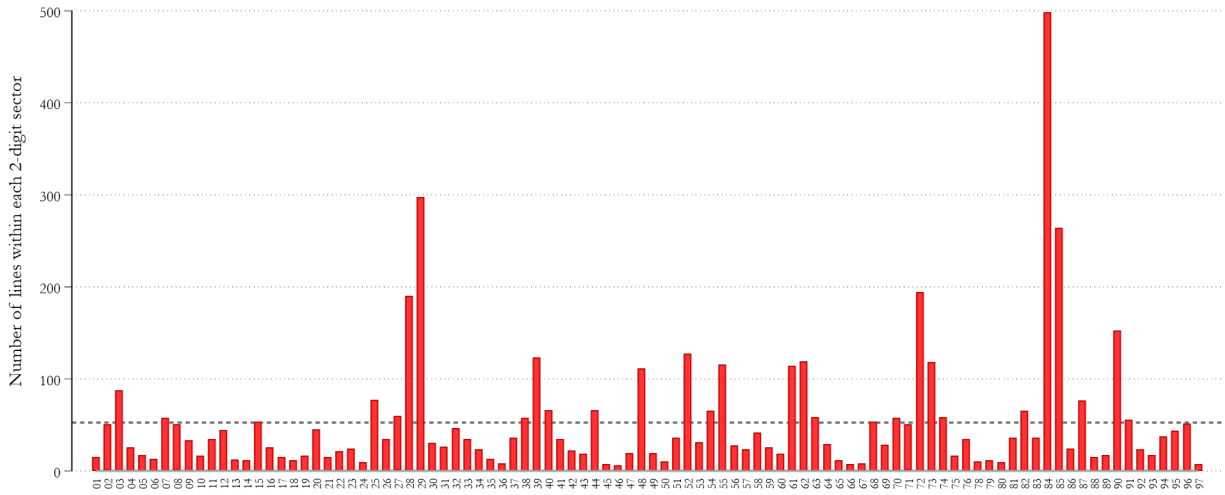


Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the nonparametric relationship between per capita income and concentration using a number of metrics: the Gini index in panel a; the Herfindahl-Hirschman Index (HHI) in panel b; the Theil index in panel c; and the number of active export lines with revealed comparative advantage greater than or equal to one (RCA ≥ 1) in panel d. The red line represents the nonparametric relationship between income and concentration using all countries. The green line excludes from the concentration measures export lines that correspond to natural resources (using the definition provided by Hausmann et al. (2014). The blue line excludes countries that are rich in natural resources (defined as countries ranking above the 75th percentile in either natural resource rents or natural resource exports as a share of GDP, on average for the period 1995 to 2020). The orange line excludes the natural resource–rich countries (using the same criteria as the blue line) and for those countries it excludes from the concentration measures export lines that correspond to natural resources. NR = natural resources.

Appendix D. Disaggregation considerations

The considerable variation of 6-digit codes within each 2-digit code of the Harmonized System (HS) dataset used in this analysis has important implications for this analysis. As figure D1 shows, some 2-digit HS codes, such as 29, having nearly 300 6-digit varieties, whereas the average 2-digit sector has 52.5 varieties. Of the 96 2-digit sectors, only 29 exceed the average. Thus, it is important to consider the extent to which the results are being driven artificially by this irregularity.

Figure D1. Number of HS 6-digit codes within 2-digit sectors



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the number of 6-digit lines within each one of the 96 2-digit sectors for the Harmonized System (HS). The dashed line marks the average number of 6-digit sectors (52.5).

To deal with this, the Theil index is decomposed to allow the nonparametric relationship between income and concentration to be estimated in a way that is not affected by the varying number of 6-digit lines across the different 2-digit sectors.

Decomposing the Theil index

Suppose there are N agents indexed by i with characteristic x_i. Suppose that sum of all x_i is S and their average is μ such that:

$$S \equiv \sum_{i=1}^{N} x_i = N\mu.$$

Then the Theil's T index is defined as:

$$
\begin{aligned}
T &\equiv \frac{1}{N}\sum_{i=1}^{N}\frac{x_i}{\mu}\ln\left(\frac{x_i}{\mu}\right) \\
&= \sum_{i=1}^{N}\frac{x_i}{S}\left[\ln\left(\frac{x_i}{S}\right)-\ln\left(\frac{\mu}{S}\right)\right] \\
&= \sum_{i=1}^{N}\frac{x_i}{S}\ln\left(\frac{x_i}{S}\right)-\ln\left(\frac{\mu}{S}\right) \\
&= \sum_{i=1}^{N}\frac{x_i}{S}\ln\left(\frac{x_i}{S}\right)+\ln N.
\end{aligned}
$$

Note that if we define:

$$p_i = x_i/S,$$

Then

$$T = \ln(N) - \sum_{i=1}^{N} p_i \ln(p_i).$$

The second term is the entropy and $\ln(N)$ is the maximum entropy that the system has. Hence, Theil's T index captures the difference between the maximum entropy and the system entropy.

Suppose that instead of a single level, there are disaggregated sections (shown with D) within aggregated classes (shown with A).

Let N_A denote number of aggregated classes and N_D total number of disaggregated sections. For an aggregate class a, denote the disaggregated sections belonging to this class with I_a. The number of disaggregated sections in this class is shown by $N_A \equiv |I_a|$.

Define the total and the mean for an aggregate class a as:

$$S_a = \sum_{i\in I_a} x_i \equiv N_a \mu_a.$$

Define within-region inequality with t_a:

$$t_a \equiv \ln N_a + \sum_{i\in\mathcal{I}_a}\frac{x_i}{S_a}\ln\left(\frac{x_i}{S_a}\right),$$

and across-region inequality with T_A:

$$T_A = \ln N_A + \sum_a \frac{S_a}{S} \ln\left(\frac{S_a}{S}\right)$$

The Theil's T index calculated using the disaggregated sections can be decomposed into aggregate classes with:

$$\begin{aligned}
T_D &= \sum_{i=1}^{N_D} \frac{x_i}{S} \ln\left(\frac{x_i}{S}\right) + \ln N_D \\
&= \sum_a \frac{S_a}{S} \sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \left[\ln\left(\frac{x_i}{S_a}\right) + \ln\left(\frac{S_a}{S}\right)\right] + \ln N_D \\
&= \sum_a \frac{S_a}{S} \left[\sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{x_i}{S_a}\right) + \sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{S_a}{S}\right)\right] + \ln N_D \\
&= \sum_a \frac{S_a}{S} \left[\sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{x_i}{S_a}\right) + \frac{\sum_{i \in \mathcal{I}_a} x_i}{S_a} \ln\left(\frac{S_a}{S}\right)\right] + \ln N_D \\
&= \underbrace{\sum_a \frac{S_a}{S} \ln\left(\frac{S_a}{S}\right)}_{T_A - \ln N_A} + \ln N_D + \sum_a \frac{S_a}{S} \underbrace{\sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{x_i}{S_a}\right)}_{t_a - \ln N_a} \\
&= T_A + \sum_a \frac{S_a}{S} t_a - \sum_a \frac{S_a}{S} \ln N_a + \ln N_D - \ln N_A \\
&= T_A + \sum_a \frac{S_a}{S} t_a - \sum_a \frac{S_a}{S} \ln\left(\frac{N_a}{N_D/N_A}\right)
\end{aligned}$$

The first term captures the inequality at the regional level. The second term aggregates the within-region inequalities. The third term corrects for the regional sizes. The term N_D/N_A gives the average region size. If all regions are of equal size, the contribution from the last term will be exactly 0.

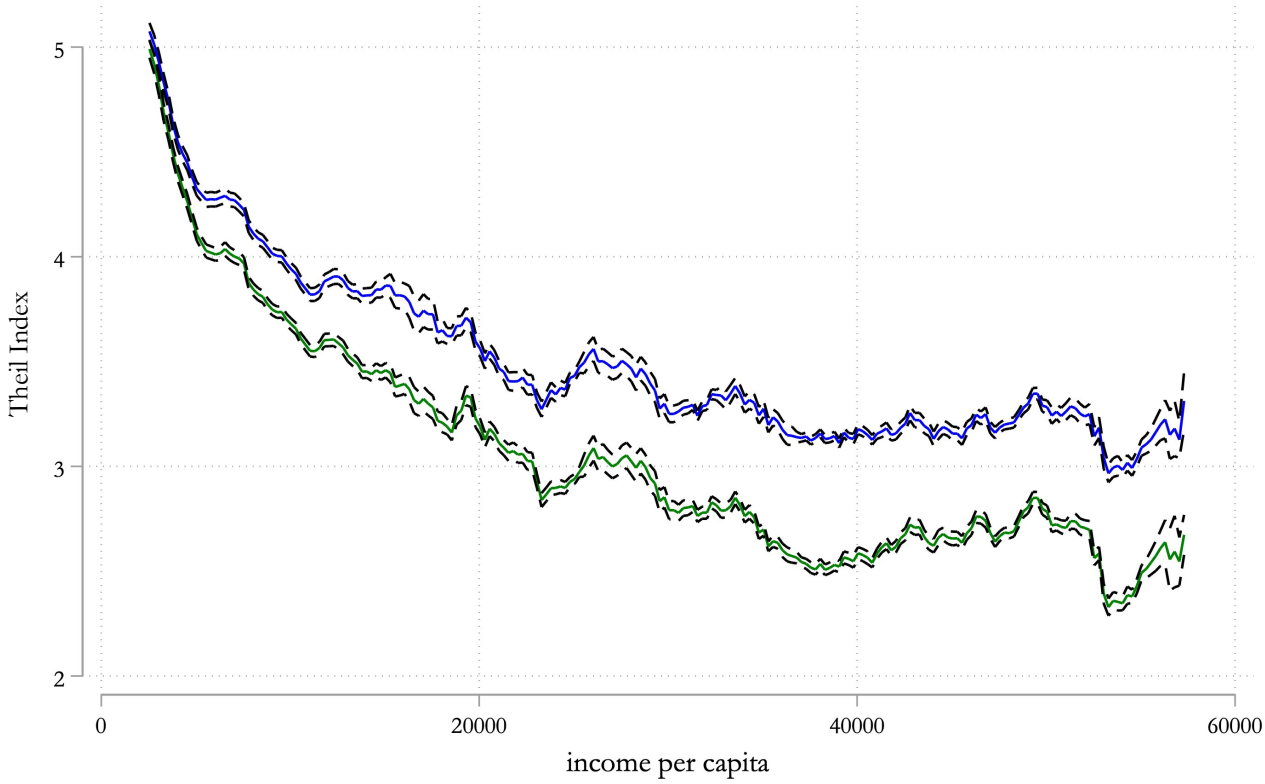
Equivalently, the index can be written using only the regional inequalities as:

$$\begin{aligned}
T_D &= \sum_a \frac{S_a}{S} \ln\left(\frac{S_a}{S}\right) + \ln N_D + \sum_a \frac{S_a}{S} \sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{x_i}{S_a}\right) \\
&= \sum_a \frac{S_a}{S} \sum_{i \in \mathcal{I}_a} \frac{x_i}{S_a} \ln\left(\frac{x_i}{S_a}\right) + \sum_a \frac{S_a}{S} \ln N_a - \sum_a \frac{S_a}{S} \ln N_a + \sum_a \frac{S_a}{S} \ln\left(\frac{S_a}{S}\right) + \ln N_D \\
&= \sum_a \frac{S_a}{S} t_a + \sum_a \frac{S_a}{S} \ln\left(\frac{\mu_a}{\mu}\right)
\end{aligned}$$

Estimation

The nonparametric relationship between income and concentration is then reestimated using the Theil index, with and without the adjustment for the different number of lines within each 2-digit sector, using the Harmonized System export data. The results are presented in figure D2. The figure plots the estimation of the nonparametric relationship between income and concentration using the Theil index based on 6-digit industries for the sample that exclude non-natural resource–rich countries. The blue line represents the estimation using the regular Theil index, whereas the green light represents the adjusted Theil index that accounts for the irregularity in the data of certain 2-digit sectors having a different number of 6-digit lines underneath them, as shown in figure D1. The figure shows that when adjusting for this, at higher level of income, there is no strong evidence of a respecialization pattern; the evidence is even weaker when using the regular Theil index. Thus, the findings are not driven mechanically by the irregularity in the data.

Figure D2. Relationship between income and concentration using the regular Theil index and adjusted for the number of 6-digit lines



Source: Original calculations for the *World Development Report 2024.*
Note: The figure plots the estimation of the nonparametric relationship between income and concentration using the Theil index based on 6-digit industries for the sample that exclude non-natural resource–rich countries (defined as countries ranking above the 75th percentile in either natural resource rents or natural resource exports as a share of GDP, on average for the period 1995 to 2020). The blue line represents the estimation using the regular Theil index. The green line represents the adjusted Theil index that accounts for the irregularity in the data of certain 2-digit sectors having a different number of 6-digit lines underneath them. The dotted black lines represent 95% confidence intervals.

Appendix E. Proofs

The proof of Proposition 1:

Shares are given by:

$$s_{ci} = \frac{x_{ci}}{\sum_{i'} x_{ci'}} = \frac{(w_c/z_{ci})^{-\theta}\Phi_i}{\sum_{i'}(w_c/z_{ci'})^{-\theta}\Phi_{i'}} = \frac{z_{ci}^{\theta}\Phi_i}{\sum_{i'} z_{ci'}^{\theta}\Phi_{i'}} = \frac{\tilde{z}_{ci}\Phi_i}{\sum_{i'} \tilde{z}_{ci'}\Phi_{i'}},$$

where $\tilde{z}_{ci} \equiv z_{ci}^{\theta}$. The Herfindahl-Hirschman Index (HHI) for country c is defined as:

$$\mathrm{HHI}_c = \sum_i s_{ci}^2.$$

For a small increase in the productivity of industry j in country c, the change in HHI is:

$$\frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\,\tilde{z}_{cj}} = \sum_i 2 s_{ci} \frac{\mathrm{d}\,s_{ci}}{\mathrm{d}\,\tilde{z}_{cj}}.$$

For $j = i$:

$$\frac{\mathrm{d}\,s_{cj}}{\mathrm{d}\,\tilde{z}_{cj}} = \frac{s_{cj}}{\tilde{z}_{cj}} - s_{cj}\frac{s_{cj}}{\tilde{z}_{cj}}.$$

For $j \neq i$:

$$\frac{\mathrm{d}\,s_{ci}}{\mathrm{d}\,\tilde{z}_{cj}} = -\frac{\tilde{z}_{ci}\Phi_i\Phi_j}{(\sum_{i'} \tilde{z}_{ci'})^2} = -s_{ci}\frac{s_{cj}}{\tilde{z}_{cj}}.$$

Hence:

$$\frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\,\tilde{z}_{cj}} = 2 s_{cj}\frac{s_{cj}}{\tilde{z}_{cj}} - 2\frac{s_{cj}}{\tilde{z}_{cj}}\sum_i s_{ci}^2 = 2\frac{s_{cj}}{\tilde{z}_{cj}}\left(s_{cj} - \mathrm{HHI}_c\right).$$

From this, the change with respect to z_{cj} can be calculated as:

$$\frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\,z_{cj}} = \frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\,\tilde{z}_{cj}}\frac{\mathrm{d}\,\tilde{z}_{cj}}{\mathrm{d}\,z_{cj}} = \theta z_{cj}^{\theta-1} 2 s_{cj} z_{cj}^{-\theta}\left(s_{cj} - \mathrm{HHI}_c\right) = 2\theta\frac{s_{cj}}{z_{cj}}\left(s_{cj} - \mathrm{HHI}_c\right).$$

With log changes, this yields:

$$\frac{\mathrm{d}\,\mathrm{HHI}_c}{\mathrm{d}\log z_{cj}} = 2\theta s_{cj}\left(s_{cj} - \mathrm{HHI}_c\right).$$

Income, on the other hand, can be obtained by solving for the wage for each country. The wage is pinned down by:

$$w_c L_c = \sum_i x_{ci} = \sum_i (w_c/z_{ci})^{-\theta}\Phi_i.$$

Solving for the wage:

$$w_c = \left[\sum_i \frac{\Phi_i}{L_c} z_{ci}^{\theta}\right]^{1/(1+\theta)}$$

Hence, with a small change in z_{cj}, the wage can be expected to change by:

$$\begin{aligned}\frac{\mathrm{d}\,w_c}{\mathrm{d}\,z_{cj}} &= \frac{1}{1+\theta}\left[\sum_i \frac{\Phi_i}{L_c} z_{ci}^{\theta}\right]^{-\theta/(1+\theta)} \frac{\Phi_j}{L_c}\theta z_{cj}^{\theta-1} \\ &= \frac{1}{1+\theta}\frac{1}{L_c z_{cj}}\Phi_j\left(w_c/z_{cj}\right)^{-\theta} \\ &= \frac{1}{1+\theta}\frac{x_{cj}}{L_c z_{cj}} = \frac{1}{1+\theta} s_{cj}\frac{w_c}{z_{cj}}\end{aligned}$$

Writing log changes in wage (income) yields:

$$\frac{\mathrm{d}\log w_c}{\mathrm{d}\log z_{cj}} = \frac{\theta}{1+\theta} s_{cj}.$$

□

References

Bahar, Dany, Ricardo Hausmann, and Cesar A. Hidalgo. 2014. "Neighbors and the Evolution of the Comparative Advantage of Nations: Evidence of International Knowledge Diffusion?" *Journal of International Economics* 92 (1): 111–23.

Bahar, Dany, and Miguel A. Santos. 2018. "One More Resource Curse: Dutch Disease and Export Concentration." *Journal of Development Economics* 132 (May): 102–14.

Broda, Christian, and David E. Weinstein. 2006. "Globalization and the Gains from Variety." *Quarterly Journal of Economics* 121 (2): 541–85.

Cadot, Olivier, Céline Carrère, and Vanessa Strauss-Kahn. 2011. "Export Diversification: What's Behind the Hump?" *Review of Economics and Statistics* 93 (January): 590–605.

Costinot, Arnaud, Dave Donaldson, and Ivana Komunjer. 2012. "What Goods Do Countries Trade? A Quantitative Exploration of Ricardo's Ideas." *Review of Economic Studies* 79 (2): 581–608.

Eaton, Jonathan, and Samuel Kortum. 2002. "Technology, Geography, and Trade." *Econometrica* 70 (5): 1741–79.

Hausmann, Ricardo, César A Hidalgo, Sebastián Bustos, Michele Coscia, Alexander Simoes, and Muhammed A. Yildirim. 2014. *The Atlas of Economic Complexity: Mapping Paths to Prosperity*. Cambridge, MA: MIT Press.

Hausmann, Ricardo, Jason Hwang, and Dani Rodrik. 2007. "What You Export Matters." *Journal of Economic Growth* 12 (1): 1–25.

Hausmann, Ricardo, and Bailey Klinger. 2007. "The Structure of the Product Space and the Evolution of Comparative Advantage." CID Working Paper 146, Center for International Development at Harvard University.

Hidalgo, César A., Bailey Klinger, A. L. Barabási, and Ricardo Hausmann. 2007. "The Product Space Conditions the Development of Nations." *Science* 317 (5837): 482–87.

Imbs, Jean, and Romain Wacziarg. 2003. "Stages of Diversification." *American Economic Review* 93 (1): 63–86.

Koren, Miklos, and Silvana Tenreyro. 2007. "Volatility and Development." *Quarterly Journal of Economics* 122 (1): 243–87.

Rauch, James E. 1999. "Networks versus Markets in International Trade." *Journal of international Economics* 48 (1): 7–35.