Scholl, Nathalie

Working Paper

An index of inter-industry wage inequality: Trends, comparisons, and robustness

Discussion Papers, No. 237

Provided in Cooperation with:
Courant Research Centre 'Poverty, Equity and Growth in Developing and Transition Countries', University of Göttingen

Suggested Citation: Scholl, Nathalie (2017) : An index of inter-industry wage inequality: Trends, comparisons, and robustness, Discussion Papers, No. 237, Georg-August-Universität Göttingen, Courant Research Centre - Poverty, Equity and Growth (CRC-PEG), Göttingen

This Version is available at:
https://hdl.handle.net/10419/168424

Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 237

An Index of Inter-Industry Wage Inequality: Trends, Comparisons, and Robustness

Nathalie Scholl

September 2017

Platz der Göttinger Sieben 5 · 37073 Goettingen · Germany
Phone: +49-(0)551-3921660 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

An Index of Inter-Industry Wage Inequality: Trends, Comparisons, and Robustness

Nathalie Scholl*

Department of Economics, University of Göttingen

Platz der Göttinger Sieben 3, 37073 Göttingen, Germany

nathalie.scholl@wiwi.uni-goettingen.de

Draft

September 13, 2017

Abstract

This paper introduces a newly constructed Theil index of between-sectoral manufacturing wage inequality and empirically tests whether the measure can serve as a basis for more general statements about the evolution of broader concepts of inequality, as argued by the authors of the University of Texas Inequality Project (UTIP) for their very similar index. Building on prior work of the UTIP, several concerns regarding the treatment of the raw data as well as important questions of internal and external validity are addressed. The index is based on sector-level data from the UNIDO Industrial Statistics for manufacturing, and I provide a detailed account of how the unbalanced raw data were treated. The newly computed index has the advantage of being consistently measured across countries and years, which makes it a valuable resource for empirical studies focusing on changes in the manufacturing structure within countries over long periods of time. However, its narrow scope also restricts the applicability of the index for other, broader uses. I argue that the latter point is one of the main drawbacks of the index and present evidence that the generalizability from between-sectoral manufacturing wage inequality to overall income inequality is severely limited. This applies not only to the extent to which the index allows conjectures about the overall level of income inequality in a society. There is reason to also question the "internal" capability of the index to accurately reflect developments in manufacturing wage inequality. I therefore do not recommend using the index as a basis for inference about the development of broader concepts of wage- or income equality.

JEL classification: J31, O15, J30.
Keywords: Inequality measurement; Wage inequality; Income inequality; Manufacturing wage inequality.

*Acknowledgements: I would like to thank Stephan Klasen, Axel Dreher, Sebastian Schneider, and the team of the University of Texas Inequality Project for helpful comments and suggestions, and Anne Körner for excellent research assistance.

1 Introduction

This paper introduces a newly constructed measure of wage inequality. I have computed a Theil index for manufacturing sectors across a large number of countries and a time period of up to 48 years. The index itself is very similar to the one developed in Galbraith et al. (1999) and Conceição and Galbraith (2000). As part of the University of Texas Inequality Project (UTIP), they constructed a Theil index based on the same type of data employed here, resulting in the UTIP-UNIDO measure of wage inequality. Building on the work of the UTIP, several concerns regarding the treatment of the raw data as well as questions of internal and external validity are addressed in this paper on the basis of the newly constructed index.

The index is based on sectoral data from the UNIDO Industrial Statistics at the 2-digit level of International Standard Industrial Classification (ISIC, Rev 3), which is a fairly crude level of aggregation. One of the main reasons why this dataset is attractive for the construction of an inequality measure is its broad time and country coverage. If the resulting, narrowly defined index of wage inequality is able to mirror overall changes in income inequality, as argued by Galbraith and Kum (2005), it can be applied in many additional contexts than just for analyses of wage inequality (or manufacturing wage inequality, for that matter). Importantly, it could serve as a proxy for developments in overall income inequality in empirical applications focusing on changes over time, such as the fixed effects model typically employed in country-level macro panel regressions. This paper tests whether the broad generalizability and applicability claimed for the UTIP index also holds for the Theil index constructed here, which is shown to be fairly similar to the UTIP-UNIDO index for many of the countries covered by both measures. A detailed comparison of the new index with the one developed by the UTIP is provided and I try to explain observed differences between the two measures. Building on prior work by Galbraith and Kum (2005), and Galbraith et al. (2015), the "external" validity of the index, that is, the extent to which the Theil index is representative of overall income inequality, is then examined. The new index is subjected to a number of comparisons with measures of income inequality to check whether it can predict developments in overall inequality. Unfortunately, the results do not lend much support to this idea. The estimates suggest that

the association between the Theil index of manufacturing and income inequality is neither very stable, nor strong enough to postulate an economically meaningful link between the two concepts.

Doubts arise not only for the "external" validity of the index, but also concerning its "internal" capability to accurately reflect developments in manufacturing wage inequality. Because the index relies on sector-level data on wages and employment which are aggregated at the 2-digit level of industrial classification, it only measures between-sectoral wage inequality and cannot give an account of inequality within sectors. Since the Theil index is decomposable into inequality between and within units, the availability of less aggregated data for subsectors at the 3- and 4-digit level allows the calculation of part of the within-component for a smaller subsample of countries and years. Despite the fact that inequality within the most detailed sectoral classification available still remains unaccounted for - which, given an average number of almost 22,000 employees in the smallest unit of record, is likely to be substantial - I find within-sectoral inequality at the 3- and 4-digit levels to make up at least 40% of overall manufacturing wage inequality. While it is obvious that manufacturing wage inequality at a single point in time is vastly underestimated, (Conceição and Galbraith, 2000) argue that the between-sectoral index is still able to trace changes in within-sectoral inequality over time. However, I find indication that in around 13% of cases, a Theil index relying on the between-sectoral component of manufacturing wages conveys an incorrect image of the direction of change in overall manufacturing wage inequality, let alone its magnitude. Given that the "true" extent of within-sectoral inequality is likely to be considerably larger because inequality between individuals within subsectors still remains unaccounted for in the more detailed data, this number might just provide a lower bound to the true discrepancy between between-sectoral and overall changes in manufacturing wage inequality.

Before moving into the analysis of the final index, the paper provides a thorough description of the challenges inherent in the raw data for creating a consistent measure of inequality over time, and the strategies employed to deal with them. The main problem with exploiting the UNIDO industrial statistics for inequality measurement is the unbalancedness of the sectoral data. For computing an index of wage inequality, data on employees and wages are used, and in order to obtain meaningful and comparable values

over time, both variables have to be present in every year in all of the sectors included in the measure. To arrive at any useful measure of inequality, some modifications of the raw data are therefore unavoidable. In the initial documentation of the UTIP-UNIDO index, it is not apparent how this was handled in the construction of the measure. With the recent release of the accompanying paper to the update of the index in 2013 (Galbraith et al. 2014), the authors themselves state that "These issues were handled on a case-by-case basis, using judgement and common sense to arrive at a set of "final revised values"." (Galbraith et al. 2014: 2). Apart from differences in the imputation of missing values, deviations between my newly computed index and the most recent version of the UTIP measure stem from a differential treatment of sectors and, to a lesser extent, from the fact that the UTIP has harmonized their index to match up with its previous versions (no longer available on the UTIP webpage). However, it is not clear how large the (respective) adjustment has been in specific cases, and which countries and/or years are affected. In a direct comparison of the index computed in this paper with that provided by the UTIP, I find several countries covered by their index for which the version of the UNIDO industrial statistics which this paper is based on does not provide data. While there is no reason to believe that the newer versions are preferable over the old ones, it is nevertheless clear that the inequality series for entire countries is based on a different version of the data. Conclusions drawn from different, earlier data versions may therefore not necessarily also hold for the newer versions, and vice versa.

Ultimately, for achieving balancedness of the sectoral data and a reasonable time coverage, the choice lies between dropping sectors with poor data coverage or the imputation of the missing values. I use both strategies in this paper, and which one I choose for a particular case depends on the number of missings as well as the number of years which would be lost due to the inclusion of a sector with limited time coverage. Testing the robustness of the measure to the discretionary decision of whether, when, and how to impute a value is not straightforward. While the UTIP data are an obvious benchmark, I also employ a few other strategies of testing the index' robustness "internally," as described in section 2.3 below. The imputation methods used are described in section 2.4, and appendix 3.B contains a more detailed version with examples from the raw data to illustrate some of the cases typically encountered. As a rule of thumb, no sector is included in the final measure

which contains more than 50% imputed data points. The imputation of data points on the raw data is regarded as relatively unproblematic since it relies on "internal" data sources originating from the UNIDO industrial statistics only. Furthermore, because imputation is performed at the sectoral level, the impact of each single imputation on the final index is very small given that the UNIDO data include up to 23 manufacturing sectors per country. While there is no way of knowing exactly the impact of imputing on the final measure, a variable is retained which contains the number of imputed data points in every year. Including it into the comparison of the Theil index to other inequality measures at least enables a judgement of whether or not it makes a difference if the index relies on imputed data.

The effect of the dropping of sectors on the accuracy of the resulting "long" inequality measure (covering fewer sectors but more years), both in terms of inequality levels and changes over time, can be assessed more easily. For those years in which data for dropped sectors are available, the Theil index is computed with and without these sectors and the resulting values are compared. As shown in section 2.3, the impact of the dropped sectors is very limited in most cases. Whenever the deviation between the two numbers is larger than 10%, an alternative version of the Theil index is computed which comprises more sectors and therefore provides more accurate numbers for inequality levels. However, the resulting "short" index covers fewer years and therefore compromises the original advantage of the index, which was to provide an account of the developments in wage inequality over long periods of time. Therefore, whenever the "long" version accurately traces developments over time, despite providing unreliable numbers for inequality levels, a recommendation is made to retain the version for dynamic analyses of inequality - that is, empirical applications which focus on changes of inequality over time. This results in a preferred version of the index with an average time coverage of 28.5 years, which I label the "dynamic" version and, unless indicated otherwise, on which the remainder of the paper is based.

Section 2 is dedicated to the construction of the index. The index and its mathematical properties are introduced in part 2.1, and part 2.2 focuses the sector-year trade-off in the construction of the index. General information and descriptives are provided in part 2.3, along with information on the properties of the dropped and retained sectors,

and an analysis of the effect of dropping sectors on the accuracy of the final measure. Part 2.4 is concerned with the sectors entering the measure and describes the imputation methods used for attaining balanced data. Section 3 provides some basic descriptives and information on the index in part 3.1 and makes a comparison to the UTIP index in part 3.2. Section 4 focuses on the role of within-sectoral inequality and section 5 relates the index to measures of income inequality. Section 6 concludes.

2 Constructing a Theil Index of Inter-Industry Wage Inequality

2.1 The Theil index

To compute the Theil index of inter-industry wage inequality, I make use of the data on total wages and employment that are provided at the sector level for a maximum of 23 manufacturing sectors, as per the ISIC 2-digit sectoral classification. The between-sector component of the Theil is defined as

$$T' = \sum_{s=1}^{S} y_s \cdot ln(\frac{y_s}{n_s})$$

with S denoting the different sectors, s=1, ..., S. y_s represents a sector's wage share, defined as a sector's wage bill divided by the sum of wages of all industries, while n_s represents the "population" (=employment) share of sector s, defined as the sector's employment over total employees (Theil 1967). This original representation of the index in shares[1] is not as common, yet it is insightful because it makes it easy to illustrate several properties of the index.[2] First, the sector's wage share can be interpreted as the weight with which each sector enters the measure. Second, if the ratio of the wage share and the population share are equal, taking their logarithm yields zero, which implies that the sector does not enter the measure. Consequently, if all income shares and population shares are equal, the between-group Theil takes its lower bound value of zero, indicating a perfectly equal distribution of income between sectors. Although the contribution of

[1] As opposed to the representation in averages, which is mathematically equivalent.

[2] For a more detailed discussion of the properties of the Theil index, consult Conceição and Ferreira (2000).

a sector to the measure will be negative whenever the population share is larger than the income share, multiplying the log value with the income share ensures that positive values have a larger weight in the final measure. This is because for every unit that has a smaller income- than population share, there must be at least one for which the opposite is true. Because positive values by construction result from income shares larger than population shares, the positive values will automatically be multiplied with a larger number than the negative ones. T' can therefore never be negative. The measure has no upper bound, which makes intuitive interpretation of a single number difficult, but comparing numbers based on the same underlying units - in this case, industrial sectors - is straightforward. Although the index is sensitive to the numer of underlying sectors S, it can be easily normalized by dividing the value by its theoretical maximum, log S. A variant of the generalized entropy class of inequality measures, the index furthermore has the adavantage of being perfectly decomposable into an infinite number of fractals, each representing within-unit inequality at a lower (i.e., more disaggregate) level. Since UNIDO also provides data at more detailed levels of sectoral aggregation (3- and 4-digit level) for some years, the use of the Theil index enables a judgement of at least part of the extent of within-sectoral inequality as compared to between-sectoral inequality. The formula including within-sectoral inequality at the 3-digit and 4-digit levels is as follows:

$$T' = \sum_{s=1}^{S} y_s \cdot (ln(\frac{y_s}{n_s}) + \sum_{s3d=1}^{S3d} y_s 3d \cdot (ln(\frac{y_{s3d}}{n_{s3d}}) + \sum_{s4d=1}^{S4d} y_s 4d \cdot ln(\frac{y_{s4d}}{n_{s4d}})))$$

y_{s3d} represents the share of each 3-digit sector's wage in their respective 2-digit sector, and y_{s4d} is the share of each 4-digit sector's wage in their respective 3-digit sector. Equivalently, n_{s3d} and n_{s4d} are the corresponding employment shares. A detailed discussion of the within-sectoral decomposition and its limitations can be found in section 4. Before moving to a discussion and analysis of the between-sectoral component of the index, the next sections describe the procedures used for achieving balanced versions of the underlying raw data, which is a prerequisite for obtaining values of the Theil index that can be meaningfully compared over time.

2.2 Between-sectoral inequality

The main challenge in exploiting the UNIDO industrial statistics for inequality measures is unbalancedness, both between sectors and over time. In order to obtain meaningful and comparable values over time, the same sectors should be included in the inequality measure every year in a given country. Hence, if data for one sector is missing in only one out of the 48 years, this means that either that year needs to be dropped, or the sector must be excluded from the index in all of the remaining 47 years. This poses great challenges given the highly unbalanced nature of the raw data. The problem is exacerbated with the inclusion of lags in empirical applications, which is typically done in macroeconomic regressions with inequality as the dependent variable due to the high degree of inertia in the measure. Already a one-year gap leads to the loss of at least 2 data points in the estimation sample, and data for single years (i.e., with missings in both the previous and subsequent year) drop out altogether. In order to obtain a workable index which can be readily used in empirical analyses, some imputation as well as interpolation is therefore indispensable. The choice between imputing missing values and dropping sectors is effectively a trade-off between two objectives. On the one hand, one wants to maximize time coverage - in particular, to fill short gaps within longer spells of data. On the other hand, the loss of information arising from the dropping of sectors should be minimized in order to ensure accuracy of the resulting inequality statistics.

It should be mentioned that the assumption is that missings in the underlying data are random across sectors. There are no patterns in the raw data suggesting otherwise, and the fact that in most instances, data is missing only in a few sectors and often in only one dimension - wages or employees - supports this view. Whether this is also true for entire years of missing data is not as clear. Because the UNIDO industrial statistics rely on surveys from establishments, the fact that no data was compiled in a certain year might have reasons which could also affect the economy as a whole, including the manufacturing industry.[3] However, there is no reason to expect that industrial sectors are affected asymmetrically and that inequality in those years in which data is missing is

[3]While the documentation of the Industrial Statistics database contains a detailed description of how non-response for individual establishments was dealt with, there is no mentioning of why entire sectors or even years are missing in some countries.

very different from that in the preceding and subsequent years. The following paragraph provides a brief overview of the sectors covered in the UNIDO Industrial Statistics, shows the impact of dropping sectors on the inequality index, and offers a solution on how to treat those cases where large differences arise between indices with and without dropped sectors. Section 2.4 will then focus on the retained sectors and describe how missing values have been dealt with.

2.3 Dropping sectors

For the construction of a time-consistent index of wage inequality, there is a trade-off between time- and sectoral coverage. Achieving a higher time coverage by excluding sectors which are available for fewer years implies a loss of accuracy in the resulting inequality measure, and vice versa. However, this trade-off between accuracy and time coverage is much less severe than one might initially expect. It turns out that in most instances, those sectors which are not well covered by the data are also the ones which are of lower economic significance for a country, and hence are also relatively small. Because the Theil index weighs the logged discrepancy between wage and employment shares by each sector's wage share, this means that the smaller sectors are also relatively less important in determining the final value of the index. Hence, omitting these sectors often changes the index very little. Before moving to a systematic analysis of the effects of dropping sectors, table 1 provides an overview of the 23 manufacturing sectors covered by the data and provides information about their average size (as measured by the wage share), the discrepancy between the wage- and employment shares, and the total number of times each sector has been included and excluded for the "long" version of the index, which aims at maximizing time coverage.[4]

Clearly, the most frequently dropped sectors are 19, 30, 32, 35, and 37. They are available only for later time periods (1990s onwards) and their inclusion would therefore mean a substantial loss in time coverage, especially when the time series is long and covers a lot of the early years.[5] Luckily, these sectors tend to be relatively small on average, with

[4]Information on the number of included sectors for individual countries can be found in appendix table A.2. All numbers presented rely on the balanced version of the data, i.e., including the imputed data points.

[5]The reason for this is the change of the ISIC classification scheme from Rev. 2 to Rev. 3 in 1989, and the accompanying re-categorization of old industries, and creation of new industrial categories such

wage shares ranging from 0.2 to a maximum of 2.9 percent of total manufacturing wages.

Apart from the wage share, the second aspect determining the importance of a sector for the Theil index of wage inequality is the discrepancy between the wage- and the employment share. Here, the omitted sectors cover a broad spectrum, with sectors 19 and 37 having a lower wage- than employment share and the rest having 25-37 percent larger wage- than employment shares. These sectors therefore do contribute to inequality, but because each contribution is weighted with a relatively small wage share, their final contribution will still be rather small.

Of course, there are other omitted sectors, and one might worry, for example, about the exclusion of sector 23 in 15 cases, which is the sector with the highest average discrepancy between the wage- and employment share, and has a wage share of 3.4 percent. Furthermore, the low average size of the frequently omitted sectors does not mean that this is also the case in an individual country, and some sectors might be of high economic significance in single economies.

As a general check of the degree to which the "long" version of the index, wherein sectoral coverage has been sacrificed for the sake of a longer time series, is representative of the overall level and development of between-sectoral wage inequality, I have therefore also computed the index for every country and year using all of the available data, including those for the dropped sectors. The resulting "full" index is not comparable over time, but it can serve as a benchmark for the comparison with the long version. The percentage difference between the two measures serves as a first indication of the degree of distortion introduced by the omission of certain sectors. Averaging over all the available countries and years,[6] the two versions seem rather similar, with the "long" version yielding 10.6% higher inequality numbers on average across all countries and years. This rather low average deviation[7] is, however, concealing large variations across, as well as within, countries. While the two versions are virtually identical in a large number of countries, others display a large difference between the indices. Moreover, in a substantial number of countries

as, e.g., sector 37 (Recycling).

[6]Only years when deviations actually occur between the two indices have been included in the computation of the different measures of convergence. In the dataset, a deviation of 0 arises if, and only if, the sectoral coverage is the same between the measure and including those years would skew the similarity indicators upwards.

[7]While this number may not appear as very small at first glance, it is driven upwards by a few "outlier" countries with very high mean deviations of above 100%

which have a low average difference, there is a lot of variation over the years. Appendix table A.1 displays the deviation between the two versions of the Theil index for all countries where the indices differ, sorted by the maximum percentage deviation.[8] In addition to the maximum, the table also reports the mean percentage deviation, and the standard deviation. These figures can provide a broad idea of the "static" resemblance of the long- with the full version of the Theil index. Researchers who care less about the level of wage inequality, but are rather interested in its development over time - which is arguably one of the main advantages of this dataset - may be more concerned about the ability of the index to trace changes in inequality. Table A.1 therefore also includes the correlation of both the level and the differences of the "long"- with the "full" index.[9] Apart from Kuwait, which has a correlation of 0.89 in differences, none of the countries with a level deviation of less than 10% has a correlation lower than 90%, neither in levels nor differences.[10] The same applies to the mean deviation, which - apart from the Philippines - is always below 6.2 percent. Interestingly, in several cases where the deviation of the level of the two indices is rather large (e.g., Botswana and the Netherlands), the correlation is still high (above 0.92). Despite starting from very different levels of wage inequality, changes over time seems to still be well captured by the long version of the index in several cases.[11]

In an attempt to address these issues and provide more accurate versions of the index, I therefore recalculated the index including more of the previously omitted sectors, with the same constraint of including the same sectors in all years. Naturally, this implies losing several years of data given that the initial motive behind excluding the sectors was to increase time coverage. In many cases, this leads to the inclusion of all sectors, but there are still sectors which are excluded also from these "short" versions. I have calculated short versions for all countries with a deviation of more than 10% in any year

[8] The mean deviation is based on the absolute value of the negative deviations, i.e., cases in which the long version is larger than the full one.

[9] Only looking at the correlations gives a slightly more optimistic, but qualitatively similar picture. Those countries showing lower percentage deviations of inequality levels between the long and the full version of the index generally have higher correlations as well, but not necessarily vice versa. Senegal, for example, has a correlation of 0.998 between the two indices over the 28 years.

[10] The correlation is based on only those years with non-zero deviations in order to not artificially drive the correlations upwards.

[11] While the reverse case can also be true, there are only two countries - Brazil and Algeria - which have a high similarity of the inequality levels (less than 10% deviation on average), but a low correlation of the indices over time.

(as indicated by the maximum)[12] and then repeated the above exercise (results are shown in panel 2 of table A.1).[13]

Of course, nothing can be said about the counterfactual deviation in the years for which data is missing on the sectors which have been dropped from the index, but there is no reason to suspect that the deviation would be larger than in the years covered. Specifically, I checked whether there is a discernible time trend in the deviation over the years, and it does not seem to be the case that the contributions of omitted sectors is growing or decreasing over time. The contributions are also not varying in any other systematic manner which would allow inference about their development outside the sample range. Overall, given the possibility to combine the two versions of the index provided for countries with a deviation of more than 10% between the long and the full version, the Theil index is able to provide an accurate picture of the extent of between-sectoral wage inequality in manufacturing. It then depends on the purpose of the research which version is preferable: those applications of the index for which the development over time is of interest may still benefit from the long version despite larger differences in the levels, and vice versa. The last column of table A.1 provides a recommendation of which index to use in dynamic applications. Since the main purpose of constructing the Theil index was its ability to trace inequality changes over a longer time horizon, I decided to keep this "dynamic" version of the index as the preferred version for the remainder of the paper. In particular, part 3.1, describing the final index, and section 5 which analyzes its similarity to other inequality indices, are based on the "dynamic" version.[14]

[12] Apart from it being the strictest criterion, I focus on the maximum percentage deviation for another reason: Because not all omitted sectors are always present at the same time, if there is a deviation between the two indices, this is not necessarily the "full" deviation. For example, of, say, 5 sectors which are not covered by the "long" index, only one might be included in a given year in the "full" index. If, for that reason, the deviation is lower in years where fewer sectors are present in the "full" version as well, taking the maximum deviation will provide a more accurate indication of the potential bias arising from the omittance of sectors.

[13] In countries with remaining deviations, i.e., where some sectors are still being dropped in the short version, the differences between the full- and the short version are now well below the 10% cut-off. There are a few exceptions where the long version is retained despite larger deviations. They are marked with an asterisk in appendix table A.1 and the reasons for keeping them are explained in detail for every case below the table.

[14] Because the "dynamic" version ensures accuracy in capturing changes over time and the comparison with other inequality measures is based on fixed effects models, using the "dynamic" version of the index is considered as unproblematic.

2.4 Retaining sectors: Imputation

Even after dropping sectors with low data coverage, the remaining dataset is far from balanced. There are a lot of observations where only one of the two variables, wages and employees, necessary for the index is provided. In other years, both variables are missing in certain sectors. The remaining missings are therefore imputed in order to attain a workable inequality index. It should be noted that due to the extremely heterogeneous data coverage across variables, countries, and years, it is impossible to apply the same imputation procedure to all countries, let alone sectors. There are different ways to impute missing values, with varying degrees of sophistication, and which one is most suitable has to be decided on a case-by-case basis.

The preferred method here is a regression-based approach. I prefer this approach over other imputation methods because it allows exploiting other information from the UNIDO industrial statistics to predict a missing value. Especially in years where there are large changes in wage- or employment shares, simply interpolating values without consulting other information provided in the dataset may lead to suboptimal outcomes and erratic movements in inequality numbers due to large changes in relative sector shares.[15] Two more variables, output and the number of establishments, are provided at the sectoral level and are positively associated with both the number of employees and their (total) wages in a given sector. Their development can be indicative of changes in those variables for which information is missing, and indeed, the relationship between these variables is very strong in many instances. Additionally, often only one of the two variables needed for the computation of the index is missing, in which case the available variable is included as a regressor as well (e.g., if a value exists for employees but not for wages, the "employees" variable enters as one of the predictors of wages). Finally, a time trend in the development of wages or employee numbers is sometimes discernible and is also included in the set of potential regressors. The fitted value from a simple OLS of the following exemplary form is used to fill the missing value (in this case for wages):

[15] For example, if a sector's employment numbers drop drastically in one year and the information on wages is missing, simply linearly interpolating the value for wages based on the previous and next year's value would lead to a large change in the relative ratio of the sector shares, whereas taking into account the information on employment and adjusting the wage value downwards leads to a smoother series.

$$\text{Wages}_t = \alpha + \rho\text{Employees}_t + \beta\text{Establishments}_t + \gamma\text{Output}_t + \delta_t + \epsilon_t$$

Again, the main obstacle to the use of this imputation method is data availability. It is not possible to always use the same regressors across countries or sectors, with available variables differing even within the same sector between years. The above example therefore only represents the most general specification while many of the actual regressions only contain a subset of the variables.

Sometimes there is no further information available at all for a missing observation, or predicting fitted values is not feasible for other reasons (e.g., due to a too-short time period which leaves no degrees of freedom for estimation). In this case, alternative imputation methods have to be explored. Second-best solutions employed in this paper are a simplified hot-deck type approach,[16] where an observation similar to the missing is used, or linear interpolation based on the surrounding values. All methods are described in detail in appendix 3.B, starting with the regression approach.

[16] See Andridge and Little 2010 for a review of the method.

Table 1: Overview of manufacturing sectors

ISIC code	Manufacturing sector (ISIC Rev. 3)	Number of times... included	excluded	total	Wage share	Average ratio of wage- to employment share	Technology level
15	Food products and beverages	113	0	113	22.7	102	Low
16	Tobacco products	96	9	105	2.4	176	Low
17	Textiles	111	2	113	10.7	85	Low
18	Wearing apparel; dressing and dyeing of fur	105	5	110	8.3	73	Low
19	**Tanning and dressing of leather; luggage, handbags, saddlery, harness & footwear**	**12**	**62**	**74**	**23**	**72**	Low
20	Wood and of products of wood and cork, excl. furniture; articles of straw and plaiting materials	109	3	112	36	80	Low
21	Paper and paper products	109	3	112	25	118	Low
22	Publishing, printing and reproduction of recorded media	103	9	112	39	124	Low
23	Coke, refined petroleum products and nuclear fuel	89	15	104	34	271	Medium-Low
24	Chemicals and chemical products	109	3	112	76	144	Medium-High
25	Rubber and plastics products	101	8	109	35	104	Medium-Low
26	Other non-metallic mineral products	109	3	112	64	110	Medium-Low
27	Basic metals	102	8	110	44	149	Medium-Low
28	Fabricated metal products, except machinery and equipment	106	7	113	52	103	Medium-Low
29	Machinery and equipment not elsewhere classified	100	9	109	42	110	Medium-High
30	**Office, accounting and computing machinery**	**9**	**52**	**61**	**11**	**137**	Medium-High
31	Electrical machinery and apparatus not elsewhere classified	98	10	108	38	116	Medium-High
32	**Radio, television and communication equipment and apparatus**	**9**	**54**	**63**	**29**	**128**	Medium-High
33	Medical, precision and optical instruments, watches and clocks	91	13	104	08	111	Medium-High
34	Motor vehicles, trailers and semi-trailers	96	11	107	41	122	Medium-High
35	**Other transport equipment**	**10**	**57**	**67**	**27**	**125**	Medium-High
36	Furniture; manufacturing not elsewhere classified	107	5	112	36	83	Low
37	**Recycling**	**8**	**50**	**58**	**02**	**84**	Low

3 Inter-Industry Wage Inequality: Trends and Comparisons

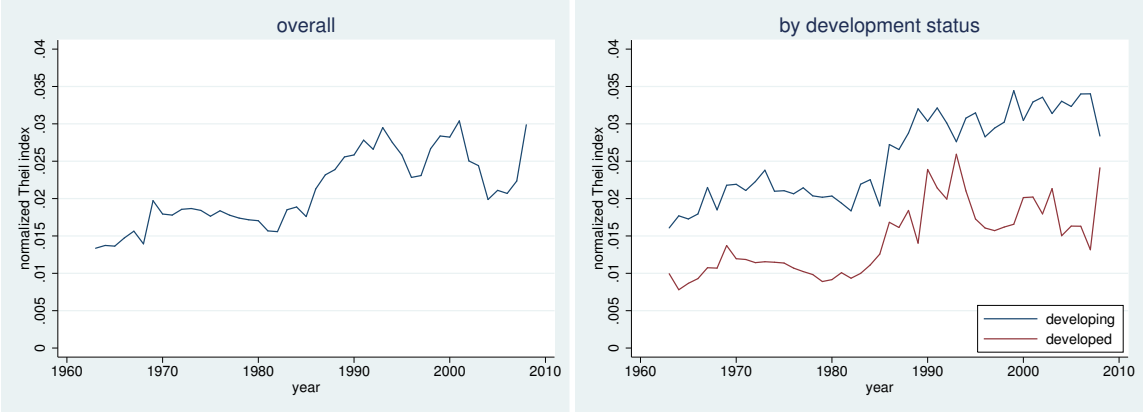
3.1 Trends in between-sectoral wage inequality

Before moving to a comparison of the newly computed index with that constructed by the UTIP, a few facts and figures of the index constructed and discussed so far are presented below. All graphs and figures are based on the "dynamic" version of the index as developed in section 2.3, which is based on the balanced sectoral dataset after imputation. The overall development of wage inequality is depicted in figure 1 below. The first graph shows the overall evolution of the index over a time period of almost 50 years and the second graph breaks it down into developing and developed countries (as per the World Bank GNI threshold definition). Note that in order to make inequality numbers comparable between countries which differ in the number of sectors underlying the measure, the graphs rely on the normalized version of the Theil index. It is clear that between-sectoral manufacturing wage inequality has been increasing over the sample period, but the largest increase occurs in the second half of the 1980s and the early 1990s. Inequality is higher in developing countries throughout the entire time period and the two series develop rather similarly. This is in line with Galbraith and Kum (2005), who find the same patterns for the first version of the UTIP dataset with data until 1999. Breaking the data down by region, as shown in figure 2, is more informative in terms of differential developments across country groups.[17]

It becomes apparent that although a small spike around 1990 appears in several country groups, the large increase from 1980 to 1990 seems to be driven to a large extent by the Middle East and North African (MENA) region (comprising both developed and developing countries). Within this group of countries, it is Tunisia and Kuwait which show very large increases in the late 1980s (shown in appendix figure A.1). The country means of the normalized (Theil(n)) and non-normalized Theil index are compiled in appendix table A.2 along with the main outcomes of the robustness exercises from sections 2.3 and 2.4, i.e., the number of sectors included the measure in each country and the number of imputed data points. Besides the basic information, which is provided for the preferred,

[17]The regional grouping relies on the World Bank classification, but Europe and North America have been pooled together into one category due to the small number of countries in the former.

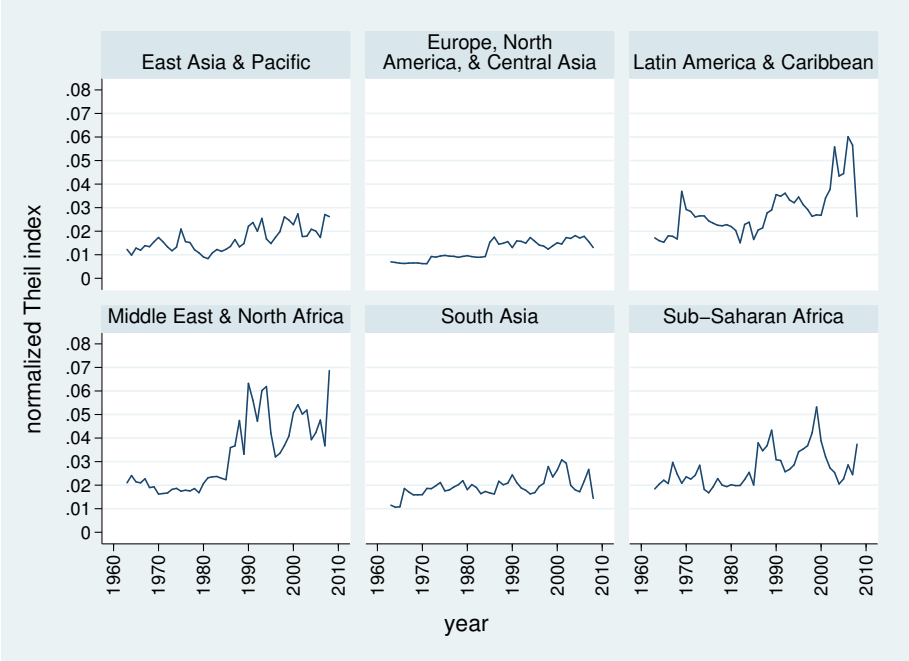
Figure 1: Evolution of the Theil index over the sample period



Notes. The first graph is based on a (relatively) constant sample of the 56 countries with a minimum time coverage of 30 years to reduce fluctuations in the time series caused by countries entering or exiting the sample in certain years. Similarly, years with fewer than 30 data points are not shown (affecting the years after 2008). The same years are omitted in the second graph for the same reasons and to ensure comparability of the two graphs.

"dynamic" version for all countries covered by the index, it contains the standard deviation of the (non-normalized) version in the last column to give an idea of the variation of the index within a country. The overall impression from the country averages is in line with the usual country rankings in terms of inequality: the lowest numbers are found in Europe, especially the Scandinavian countries, and high numbers are most prevalent in developing countries and countries from the Middle East. There are a few surprising cases though, such as The Gambia, Nicaragua, and Afghanistan scoring very low on manufacturing wage inequality and Romania, a former communist country, scoring very high. This already provides some indication that manufacturing wage inequality is not always very closely related to a country's overall income inequality, and can sometimes generate a misleading image if such generalizations are drawn. The relationship between manufacturing wage inequality as constructed here and overall income inequality is examined more systematically in section 5. Before moving to the question of internal and external validity of the newly constructed index, I compare it to the UTIP index. The more similar the two measures turn out to be, the more will the results from the subsequent validity analysis also apply to the UTIP index. Furthermore, the comparison with the UTIP index can yield some indication as to whether the extent of imputation systematically distorts the resulting measure.

Figure 2: Development of the Theil index by region



Notes. The graphs are subject to composition effects due to the entering and/or exiting of countries throughout the sample period. The regional subsamples have not been restricted any further to improve time consistency in country coverage because of the lower number of countries per group as compared to the graphs in figure 1

3.2 Comparison to the UTIP-index

The first noticeable difference between the newly constructed index and the Theil index calculated by the UTIP is country and time coverage. The average time coverage of the new Theil index is 28.5 years,[18] vs. 26.2 years for the UTIP one, and has information for 137 countries, whereas the UTIP index covers a total of 154 countries. There are 2 countries (Liberia and Serbia) with a total of 15 observations which are in the new index but not in the UTIP one, and 19 countries with a total of 210 observations for which the UTIP provides information but which are not covered by the new index. 10 out of these 19 countries are not part of the version of the UNIDO industrial statistics used in this paper,[19] and another five countries (Armenia, the Bahamas, Rwanda, Sudan, and Zimbabwe) have not been included in the new index due to the lack of useable raw data.[20] This implies that

[18]For the "long" version, average time coverage is 31.3 years.

[19]These are Bahrain, Bhutan, Cap Verde, Czechoslovakia, the German Democratic Republic, West Germany, Equatorial Guinea, Myanmar, the Seychelles, and Togo.

[20]That is, although some data is provided for these countries in the UNIDO industrial statistics, the data never covers both wages and employees at the same time.

the UTIP indices are based on older versions of the UNIDO data for those countries.[21] Because the older UNIDO data rely on different industrial classification schemes, the index values do not necessarily compare easily from the new to the old versions. In particular, it appears that the previous version of the UTIP index was based on a more detailed, 3-digit level of classification, which makes it more accurate in capturing manufacturing wage inequality. Since it is not clear how exactly the data were harmonized with the previous version of the index, I could not empirically determine how much of the difference between the new index and the UTIP one arises from the differential treatment of sectors, and how much is due to varying data sources (or harmonization efforts thereof).

The overall correlation between the UTIP and the newly constructed Theil index across all 135 countries covered by both indices is 0.83 for inequality levels and 0.79 for changes in inequality. Although the two indices appear to develop rather similarly on average, the correlations by country reveal large differences and range from a perfect correlation of 1.0 in 17 countries to negative correlations in Bulgaria, Germany, Estonia, Jamaica, and Uganda.[22] Appendix table A.3 displays the number of imputed data points, the year coverage for the UTIP and the "dynamic" version of the Theil index, as well as the correlations and relative deviations of the two measures in levels and differences. The degree of divergence between the two measures is weakly, but significantly correlated with the extent of imputation[23] in the new index. This makes perfect sense given that the reason for imputing values was that the raw data was not utilizable and hence the construction of any sort of index requires a choice of whether to impute or not, and, if applicable, of the imputation method. Obviously, if no imputation is carried out, differences in the measures are implied due to the resulting fluctuation in sectoral coverage. But even if the data have been modified in some way, the outcome is not necessarily the same and the resulting

[21] The other four countries covered by the UTIP but not by the new index have been excluded due to insufficient time coverage. Angola, the United Arab Emirates, Bosnia and Herzegovina, and Cambodia have a maximum time coverage of four years in the UNIDO industrial statistics, of which a maximum of two years are consecutive. The resulting inequality measure would therefore be of little use for comparisons over time, which is the main selling point and the reason for constructing the index in the first place.

[22] The correlation is equal to one in Burundi, Benin, Burkina Faso, Belize, Congo, Cuba, the Dominican Republic, Gabon, Iraq, Kazakhstan, Kuwait, Nigeria, Puerto Rico, El Salvador, and Tanzania. Five more countries display negative correlations in differences: Australia, Belgium, Moldova, the Netherlands, and Puerto Rico.

[23] As measured by the total number of imputed values over all years and sectors for a given country. Appendix table A.3 provides an overview of the correlation between the two measures and the extent of imputation.

indices are likely to still differ to some - smaller - extent. On average, the dynamic version of the Theil index is 3.8 percent higher than the UTIP measure whereas the long version is 1.2 percent lower. This makes perfect sense given that the dynamic version contains more sectors, whereas time coverage has been maximized at the cost of sectoral coverage in the long version. While these averages again differ substantially across countries, neither the dynamic, nor the long version display significant correlations between the average sign of the deviation and the number of imputed data points at the country level.

Looking at those countries which display low correlations or very high deviations from the UTIP in more detail, a few peculiarities are noticeable. First, the association with the number of imputations is not stronger in the countries displaying negative correlations with the UTIP index than for the rest of the sample. This supports the stance that the imputation of missings in the underlying sectoral wage and employment data does not lead to systematically different numbers in the resulting inequality index. Second, in many cases with low correlations, the deviations between the UTIP and the new index are equally high across all versions of the index - that is the long, short, and full ones - which is again in line with the fact that the data used for the UTIP index for these countries stem, at least partly, from other versions of the UNDIO industrial statistics.[24] If anything, correlations are lower with the short version of the index, which is an indication that the UTIP in some cases also use only a subset of sectors for the calculation of their index. As confirmed by the UTIP, this is a reflection of efforts to keep the measure time-consistent.[25] The lower correlation with the short- as compared to the long version of the index occurs in several instances where the short version was kept due to the inaccurate representation of inequality levels and/or dynamics of the long version, as explained in

[24] E.g., in Puerto Rico, Estonia, Bulgaria, Jamaica, and Uganda, among others.

[25] This is more prevalent for the levels than for the differences, which is in line with the previous finding that even when inequality levels are different, a slimmer version of the index is still able to trace changes over time quite well. By construction, countries where this was the case have been included in the "dynamic" version of the index and hence the higher deviation of the "short" version for the levels as compared to the differences is implied. The lower similarity with the UTIP also shows up in the average correlation across all countries, which drops to 0.6. To name a few country cases, lower correlations for the short version are found in the Netherlands, Great Britain, Bolivia, and Romania, among many others. It should not go unmentioned that the opposite case also occurs in the data a few times, e.g., in Botswana, where the correlations jump from 0.23 from the long/dynamic to 0.95 for the short version in differences, or Madagascar, where they rise from 0.66 to 0.96 for the levels. It should be noted, however, that these cases also display almost equally high (and sometimes even higher, as, e.g., in Ireland,) correlations with the "full" (time-inconsistent) version of the index and the short version may merely be a reflection of sectoral coverage in the full version, especially if there are few years with missing sectors.

section 2.3. This finding may stem from grouping sectors together in the UTIP index, which effectively reduces the information content of the data and automatically leads to lower inequality numbers. Lastly, in several instances, the deviations from the UTIP are substantially lower (but never zero) with the "full" (time-inconsistent) version of the Theil index.[26]

In order to get an idea of the drivers of the divergence between the UTIP measure and the new index, a simple panel regression[27] is employed with the percentage difference between the UTIP- and the new index[28] (in its "long" version) as the dependent variable. The number of dropped sectors and the number of imputed data points in the underlying sectors in each year are the main explanatory variables, and year dummies are added to the model to check whether the difference between the indices is growing over time. Table 2 contains the results.

Table 2: Explaining differences to the UTIP index: imputation vs. sectoral coverage

	(1)	(2) r	(3)	(4) r
Imputations	3.030***	3.030**	3.111***	3.111***
	(0.466)	(1.225)	(0.460)	(1.154)
Dropped sectors	-5.282***	-5.282	-2.396***	-2.396
	(0.815)	(3.444)	(0.607)	(1.817)
Constant	0.0131	0.0131	0.105	0.105
	(10.30)	(5.216)	(4.026)	(4.081)
Observations	3.627	3.627	3.627	3.627
Year dummies	YES	YES	NO	NO
# of countries	135	135	135	135
R^2	0.036	0.036	0.016	0.016

Notes. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the percentage deviation between the long version of the newly constructed Theil index and the UTIP index. All estimations are employing random effects. The "r" in the top column indicates that standard errors are robust. R^2 refers to the within R^2.

Clearly, the number of imputations is related to the divergence of the two measures,

[26]Most notably, this is the case for Madagascar, New Zealand, Moldova, Great Britain, and Austria for inequality levels. The problem is less prevalent for differences, where the correlation is often higher with the short version than the "full" one.

[27]The initial idea was to estimate the model in fixed effects to account for the fact that the UTIP relies on data sources other than the UNIDO industrial statistics in some countries. However, a Hausman test indicates that the estimates do not differ from the more efficient random effects model (chi2(47) =49.61, p = 0.3697), which is therefore retained.

[28]the "long" version of the index is used here because it has a lot more variation in the "dropped sectors" variable compared to the "dynamic" version, making comparisons to the UTIP index more meaningful. The main results hold when using the "dynamic" version, but the coefficient on the "dropped sectors" variable is only around half the size.

with one additional imputed data point implying a 3 percentage point higher deviation. Interestingly, the number of dropped sectors has a negative coefficient, meaning that for every dropped sector, the two indices are on average 5 percentage points more similar. However, the use of robust standard errors, as warranted by a maximum likelihood ratio test, renders the coefficient insignificant. The last two columns do not contain the year fixed effects, which clearly reduces the size of the coefficient on the dropped sectors. Looking at the values of the year dummies (displayed in the full version of the table in the appendix, table A.4, it becomes clear why this is the case: from 1990 onwards, the year dummies become positive and keep increasing over the 1990s and 2000s. This can be explained by the fact that, as mentioned in section 2.3, data for five sectors (19, 30, 32, 35, and 37) are only available from 1990 onwards and often only start in the mid-1990s. As explained earlier, they are therefore frequently dropped for the long version of the Theil index. Additionally, harmonization efforts of the new and old UTIP index mainly focus on the time periods in which the indices overlap, and data from older classifications schemes can only be employed before the transition from the old to the new classification took place. The year dummies pick up this effect which is similar across all countries and allow the coefficient on the sectoral coverage to capture the remaining variation in sectoral coverage. Also note that the constant is not significantly different from zero, which means there does not seem to be an inherent difference between the two indices.

Overall, while I can replicate a major part of the UTIP inequality statistics with the newly computed index, there are large differences in quite a few cases. This is in line with the low explanatory power of the model analyzing the differences between the new index and the UTIP one. These results suggest that other factors not contained in the model - one of them most likely being the re-grouping of the underlying sectoral data in the UTIP index - are more relevant for causing the difference between the two indices. For the remainder of this paper, this implies that all conclusions drawn only apply to the inequality numbers based on the sectoral information from the UNIDO industrial statistics using the ISIC Rev. 3, and not necessarily to those stemming from other, possibly more detailed earlier versions of the data or differently grouped industries. However, given that future values will be in the new classification scheme, the relevance of my results will be growing as the time coverage of the index is extended to more recent years.

4 On the role of within-sectoral inequality

Although the UNIDO industrial statistics do not contain individual-level data, one can still compute part of the within-sectoral inequality by exploiting the more refined sectoral classifications up to the 4-digit level, as provided by the Industrial Statistics Database (INDSTAT4). The share in total wage inequality of the within-component at the 3- and 4-digit level can give at least a rough idea of the lower bound of overall manufacturing wage inequality.[29] Unfortunately, the time coverage is much lower than for the 2-digit level data and spans only for the years from 1990 onwards. It should be noted that the raw data at the 3- and 4-digit level suffer from the same problems of unbalancedness as the 2-digit ones, but have not been modified in any way to address the resulting problems of comparability.[30] Inequality numbers - both between- and within-sectors - are therefore not generally comparable over time and shall merely provide an indication of the potential magnitude of within-sectoral inequality.

Within-sectoral inequality is created at three levels, 4d representing the most detailed (4 digit) level. The formula, introduced in section 2.1, is, in its expanded version, easily separable into different components:

$$T' = \sum_{s=1}^{S} y_s \cdot ln(\frac{y_s}{n_s}) + \sum_{s=1}^{S} y_s \sum_{s3d=1}^{S3d} y_s 3d \cdot ln(\frac{y_{s3d}}{n_{s3d}}) + \sum_{s=1}^{S} y_s \sum_{s3d=1}^{S3d} y_s 3d \sum_{s4d=1}^{S4d} y_s 4d \cdot ln(\frac{y_{s4d}}{n_{s4d}})$$

The different parts are calculated separately in order to enable statements about the contribution of 3-digit "between sector"-inequality as the within-sectoral component at the 2-digit level, without adding the 4-digit level contribution as well. The following terms

[29] Note that the sum of the between-component and the within-components at the 3- and 4-digit levels is in the following referred to as "total" or "overall" inequality for the sake of simplicity, although it is technically not total or overall inequality given that within-sectoral inequality at the 4-digit level remains unaccounted for.

[30] Another problem of the multi-level data for the calculation of a decomposable Theil index is that subgroups must be exhaustive and mutually exclusive. I.e., all lower-level (4- and 3-digit) numbers must add up to the total value provided for next level. Since this is a necessary requirement, the raw data had to be adjusted in a way such that the numbers add up at the different levels. If the higher-level value was higher or lower than the sum of the lower-level values, the difference has been added to or subtracted from the higher-level figure. While a desirable alternative would have been to create an extra category at the lower level containing the missing amounts in the case of too-low sublevel sum, this would have meant that in some cases, positive numbers for one variable (wages or employees) are matched up with zeros for the other one, and including this "residual" sector in the calculation of the Theil index is impossible due to the logarithmic transformation of the ratios.

are retained separately:

Between sectoral inequality at the 2 digit level: $$BE2 = \sum_{s=1}^{S} y_s \cdot ln(\frac{y_s}{n_s})$$

Between sectoral inequality at the 3 digit level: $$BE3 = \sum_{s3d=1}^{S3d} y_s 3d \cdot ln(\frac{y_{s3d}}{n_{s3d}})$$

Between sectoral inequality at the 4 digit level: $$BE4 = \sum_{s4d=1}^{S4d} y_s 4d \cdot ln(\frac{y_{s4d}}{n_{s4d}})$$

Within sectoral inequality at the 3 digit level: $$WI3 = \sum_{s3d=1}^{S3d} y_s 3d \cdot BE4$$

Within sectoral inequality at the 2 digit level (without the 4 digit level contribution): $$WI2 = \sum_{s=1}^{S} y_s \cdot BE3$$

"Total" within sectoral inequality amounts to: $$WI2 = \sum_{s=1}^{S} y_s \cdot WI3$$

The final index is then computed as BE2 + WI. The average contribution across all sectors, countries, and years of within-sectoral inequality at the 3- and 4-digit levels (WI) is 33.7%, which would indicate that between-sectoral inequality (BE2) still explains around two thirds of overall inequality in manufacturing. In terms of contributions to the within-component of the 3- vs. the 4-digit level, interestingly, the one-third/two-third ratio found previously for the between- versus within 2-digit level is reversed. On average, little over one third of within-sectoral inequality stems from inequality at the more aggregate 3-digit level (BE3) while two thirds can be attributed to inequality between 4-digit level sectors (BE4). Of course, true total within-sectoral inequality will be larger given that inequality within the 4-digit level sectors remains unaccounted for here.

Especially the result for the overall contribution of the within-component should, however, be taken with caution given the unbalancedness of the raw data.[31] The actual 3- and 4-digit within-sectoral inequality is certain to be higher in years with larger gaps and more missing data at the lower levels, and a first, crude correlation analysis indeed confirms a small positive correlation of 0.2 between the number of subsectors per 2-digit category

[31] This is less of a problem for subsectors at the 3- and 4-digit level, given that a missing 2-digit sector implies that all of its subsectors are missing as well, whereas a missing 3-digit sector "only" leads to missings at the 4-digit level, which is the smallest available bracket already.

and the share of the within-component. Moreover, the variation in the importance of the within-component across countries and years is very large and there are cases where within-sectoral inequality explains as much as 87% (Moldova in 2002) of overall inequality. Country averages also show a lot of variation and range from 66.6% in Lebanon to 5.4% in Kuwait. There are no clear trends in the development over time, either - in some countries, the within component seems to be growing, in other it is decreasing, and in several cases it is relatively constant over the years. Again, it is important to keep in mind that at least part of the variation in the within component stems from the unbalancedness of sectoral coverage over the years. Appendix figure A.2 displays, for every country, the development of both the contribution of the within component - that is, the percent of total inequality which stems from the within-component - and the total number of subsectors (both 3- and 4-digit) with non-missing values in all 2-digit categories per year. As can be seen in the graphs, there are several countries with consistently high (or low) data coverage, which potentially mask the importance of the balancedness issue for teasing out the true within-component. Indeed, the standard deviation of the variation in sub-sector coverage is a mediating variable in the association of the contribution of the within component and sectoral coverage.[32] Once the countries with a low variation in sectoral coverage are discarded, the correlation between the share of within-sectoral inequality and sectoral coverage rises to 0.25 (countries above the mean variation). Only using countries with one standard deviation above the mean variation, it is 0.52, and for countries with variation higher than two standard deviations above the mean, it is 0.71. This indicates that very large changes in sectoral coverage are accompanied by increases in the importance of the within-component as well. Nevertheless, balancedness does not seem to be the only driver of the importance of the within component across the entire sample, and in particular it is not very relevant for those countries with good data coverage throughout.[33]

The second major factor for the extent of within-sectoral inequality is the sectoral composition of the manufacturing industry of a country. Some sectors by construction

[32]The correlation between the standard deviation in the total number of subsectors (across years within a country) and the correlation of the same with the share of the within component is 0.44.

[33]Another explanation for the low average correlation is the very crude measure of data coverage provided by the total number of subsectors. It could still very well be - in fact, it is likely to be the case fairly often - that some sectors are included in some years while others are not. This variability is very likely to substantially affect the within-component. In order words, it does not only matter how many sectors are covered, but also *which* ones are included (and which ones are not).

have more subsectors than others. In the extreme case of only one subcategory per 2- and 3-digit category, there is no within-group inequality by construction. This is the case for sectors 16 (Tobacco products) and 30 (Office, accounting and computing machinery), and consequently, countries whose manufacturing industry is concentrated in those sectors are likely to have a lower share of within sectoral inequality. Averages across sectors indeed reveal large differences in the importance of the within-component, and the ranking of 2-digit sectors in terms of the size of their within-component (taking both 3- and 4-digit sectoral inequality into account) is clearly correlated with the number of subsectors into which each category is divided.[34] Appendix table A.5 provides more detailed information on the association between the number of subsectors and the size of the within component for every sector.[35]

In order to work out the importance of the sectoral composition, a simple country fixed effects regression is conducted, where the share of within-sectoral inequality is regressed on the number of subsectors and a set of year dummies.[36] The results are displayed in column 1 of table 3. For an assessment of the importance of sectoral coverage versus sectoral composition, the wage shares of the different 2-digit sectors are then added to the regression in column 2.[37]

Clearly, the sectoral composition takes away from the sectoral coverage effect, which decreases by more than 40%. While the fixed effects estimator does remove all time-invariant

[34]The correlation is 0.75 and the number of subsectors refers to the mean number of 4-digit sectors per 2-digit category. The correlation with the total number of subsectors (3- and 4-digit sectors) is very similar (0.77). Only cases which have a non-zero within-component have been considered in these calculations.

[35]Clearly, those sectors ranking high on within-sectoral inequality (that is, the logged ratio of the wage- over the employment share) also tend to have a higher number of subsectors. This is still true for the weighted component shown in panel 3 of table A.5, although the association is slightly weaker due to the weighting with the sector's wage share shown in panel 2.

[36]The results presented use the number of 3-digit categories per 2-digit sectors because, as previously show, the 3-digit level accounts for two thirds of the within-component. Results are very similar when the number of 4-digit categories, or the number of total subcategories is used instead (results available upon request).

[37]Note that sectors 16 and 30 have been omitted from the regressions. If all sectors (including 16 and 30, which have no subsectors and can therefore never positively contribute to the share of the within-component) are included in the regression containing the sector shares, all sectoral coefficients have positive signs and interpretation of the results is not straightforward. This is because the shares of all other sectors are implicitly evaluated against the shares of sectors 16 and 30, which by definition (due to the lack of subsectors) never contribute to within-sectoral inequality. Hence, the larger the shares of the other sectors, the smaller will be by construction the share of sectors 16 and 30, which ceteris paribus implies a larger within-component. Moreover, because both sectors 16 and 30 have on average larger wage- than employment shares (the ratios being 176 and 137, see table 1) , whenever the wage share of those sectors rises, the between component of the Theil index will rise as well, implying by definition a smaller within-component.

country-specific factors which potentially affect the size of the within-component, it still estimates a common slope parameter for the sectoral coverage variable for both high- and low-variability countries. Random effects estimation confirms that the coefficient on the sectoral coverage variable ("subsectors") is hardly affected by the removal of the country fixed effects (as shown in appendix table 3.B.6). As argued above, sectoral coverage is likely to be a relevant factor skewing the size of the within-component only for those countries where sectoral coverage varies substantially over the years. The fixed effects regression is therefore repeated for two different high-variation subsamples: one with above-average variation in sectoral coverage, and one with one standard deviation above the average variation. Table 3 displays the results.

According to the point estimate of the number of subsectors, an additional sector is associated with a 0.026 percentage point higher wage share. Although it is highly significant and robust across specifications, this is a rather small number. Relating it to the standard deviation of the "subsectors" variable, an increase of one standard deviation (43) would imply a mere 1.12 percent higher within-sectoral wage share. For the high-variation subsamples, the point estimates rise to 0.03 and 0.05, implying higher within-sectoral wage shares of 1.4 and 2.1 percent, respectively, for a one-standard deviation increase in subsectoral coverage. If one considers the average maximum distance between the highest and the lowest sectoral coverage within a country, numbers for between-sectoral inequality would be between 3 and 5.75 percent higher on average. More noticeable than the increase in the coefficient is the substantial rise in the R^2 for the high-variation subsamples. It does indeed seem that sectoral coverage explains the lion's share of the variation in the within-component in those countries displaying major changes in sectoral coverage.

In order to get at least a rough idea of what the within-component would be if sectoral coverage had been larger in those countries displaying large variations over the years, the coefficient estimates obtained from the above regressions are used to obtain counterfactual values when sectoral coverage is raised to a level found in other years in the same country.[38]

[38]It is not obvious what this level should be and setting it is somewhat arbitrary. In order to not overestimate the potential within-sectoral inequality numbers due to outliers at the top, the sectoral coverage numbers are split into quintiles and the lowest value of the highest quintile is used as the counterfactual coverage value for all years with lower sectoral coverage. When numbers of sectoral coverage are identical at the upper end of the distribution, or when there are too few data points for a country, the highest value of the 4th quintile is used instead. In those cases where both values are available, the difference between the two is very small (8 on average), with a maximum of 33 for Ireland, where the values are rather high

Table 3: Within-component: sectoral coverage vs. sectoral composition, FE results

	(1)	(2)	(3) High variation sample1	(4) High variation sample2
Subsectors	0.0451***	0.0261***	0.0301***	0.0505***
	(0.016)	(0.009)	(0.009)	(0.012)
Share_15		0.799	1.805**	-0.205
		(0.970)	(0.798)	(2.772)
Share_17		-0.0140	0.728	-0.797
		(0.950)	(0.838)	(2.224)
Share_18		-0.616	-0.360	-5.747**
		(0.951)	(0.915)	(2.361)
Share_19		-0.654	1.798	8.376**
		(1.306)	(1.732)	(3.015)
Share_20		2.989**	4.801***	6.038*
		(1.416)	(1.102)	(2.934)
Share_21		0.370	1.637	1.527
		(1.401)	(1.186)	(3.452)
Share_22		-0.452	-0.466	-6.338**
		(0.967)	(1.015)	(2.013)
Share_23		-1.629	-0.826	-3.250
		(1.508)	(0.580)	(2.187)
Share_24		-0.542	-0.409	-2.067
		(0.976)	(0.879)	(2.292)
Share_25		0.604	0.848	-3.434
		(1.037)	(0.992)	(1.965)
Share_26		1.135	1.419	-0.253
		(1.040)	(1.069)	(2.192)
Share_27		-1.132	-0.494	-3.148
		(0.817)	(0.725)	(2.401)
Share_28		0.658	1.135	-0.683
		(0.866)	(0.836)	(2.440)
Share_29		-0.181	0.358	-0.784
		(0.900)	(0.833)	(3.053)
Share_31		-0.489	0.119	-2.919
		(1.231)	(0.852)	(3.596)
Share_32		-0.578	-0.00135	-4.494*
		(0.902)	(0.947)	(2.454)
Share_33		-0.108	-0.685	-6.274
		(1.516)	(2.246)	(4.080)
Share_34		-0.446	2.191*	-8.611
		(1.028)	(1.109)	(5.839)
Share_35		-0.567	0.124	-1.805
		(0.877)	(0.838)	(2.448)
Share_36		-2.478*	-1.338	-4.480
		(1.297)	(1.311)	(2.993)
Share_37		2.189	1.967	2.196
		(2.656)	(2.070)	(6.554)
Constant	15.14**	26.44	-46.73	201.4
	(6.63)	(85.10)	(73.66)	(223.0)
Year FE	YES	YES	YES	YES
Observations	429	429	221	74
R^2	0,465	0,465	0,654	0,898
# of countries	53	53	27	11

Notes. Robust standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the share of within-sectoral inequality in %. Numbers 15 to 37 refer to the 2-digit sector's wage share in total manufacturing wages in %. High variation samples 1 and to refer to subsamples of countries with above-average variation in sectoral coverage (1), and countries with one standard deviation above the average variation (2).

The adjustment is done for all three samples.[39] The first adjustment on the full sample yields a within-component of 34.5, which is less than 1 percentage point higher than the unadjusted value of 33.7. The adjustment for the first subsample, displaying an above-average standard deviation of sectoral coverage, results in a very similar value of 34.4 for the full sample. However, adjusting the high-variation subsample leads to a substantial increase in the average within-component to around 40.5%. Given that the rest of the sample remains unadjusted, the true extent of within-sectoral inequality at the 3- and 4-digit level is likely to be above 40%. Furthermore, sectoral composition has not yet been accounted for, either. As shown in table 3, including the sector shares takes away from the positive effect of sectoral coverage and hence sectoral composition is likely to increase the within-component even further. This also applies to those countries with a low variation in sectoral coverage, where the sectoral coverage adjustment would make very little difference.

The full extent of within-sectoral inequality in manufacturing is of course not covered by including sectoral data at lower levels. There is certainly a substantial amount of inequality within 4-digit sectors, which on average still have almost 22,000 employees. In a country like Great Britain, where the within-component accounts for as much as 85% of overall inequality in 2009, average employee numbers in that year at the 4-digit levels are 44,600, leaving room for a substantial amount of unequal pay among these workers. If inequality within the 4-digit sectors was added to the within-component, it is fairly certain that between-sectoral inequality would explain very little of the overall inequality in manufacturing.

Nevertheless, it is possible that changes in inequality over time in between-sectoral inequality can reflect the overall trends in inequality. Conceição and Galbraith (2000: 67) argue that this is likely to be the case, given that "while within-group inequalities are likely to be large relative to differences between group averages, the internal rigidity of

at 415 and 448.

[39]Because sectoral composition is by construction skewed in those years with missing data because sectoral shares for non-missing sectors are larger than they would be if the missing sectors were present, simply using fitted values from the previous regression model would distort the results substantially for precisely those years where sectoral coverage is lower. Therefore, only the coefficient estimate for sectoral coverage is used and is multiplied with the yearly difference between the counterfactual high data coverage and the actual number of 3-digit level subsectors. This value is then added onto the observed within-component. Where the counterfactual high data coverage is lower than the actual number of 3-digit level subsectors, the original value is retained and consequently, the within-component is not modified.

industrial structure tends to assure that changes in within group inequalities in an industrial classification will be small relative to changes between groups." However, building on the argument that "industries [...] mean something, and if they mean anything at all, the effect must be to impose a measure of homogeneity on entities classified together, and a measure of distinctiveness to entities classified as being in different groups," it is much easier for a worker to switch between subsectors within a broad industrial category than to switch between industries. Or, in other words, there is likely to be more "movement among the more fine-grained subsectors. In fact, it is possible that there is no change in inequality at the broader 2-digit sector level, with employee and wage bills remaining unchanged, but there can be a substantial amount of re-shuffling within industries which remains unaccounted for entirely. It is true that for a large number of sectors (wherein 22, being the number of manufacturing sectors covered by the data, can be considered large), the overall effect of large within-changes is mitigated due to the presence of 21 other sectors. However, these 22 sectors are far from equally sized, and it is precisely the large sectors which are divided into more subcategories and display larger amounts of within-sectoral inequality to begin with. Most notably, sector 15 (food and beverages) makes up 20% of the wage share on average. Looking at developing countries, where this sector is of higher economic significance than in the developed world, it accounts for over one fourth of overall manufacturing wages. It also has the highest within-sectoral inequality, and, consequently, more than forty percent of within-sectoral inequality can be attributed to sector 15 on average in developing countries. The average contribution of 14% of the sector to between-sectoral inequality at the 2-digit level is also large. It is, however, to a large extent attributable to the sector's large wage share - in fact, sector 15's discrepancy between wage- and employment shares is among the lowest of all sectors, at least for developing countries. Hypothetically, if one assumed an increase in wages in sector 15, this would decrease between-sectoral inequality because the sector's contribution to the Theil index is negative, i.e., it has a lower wage- than employment share. Nevertheless, the within-sector component would be assigned a higher weight due to the sector's increased wage share, even assuming that the increase in wages is distributed within the sector in such a way that does not lead to higher within-sectoral inequality itself. As a result, between-sectoral inequality would decrease but within-sectoral inequality would increase.

Unfortunately, the unbalancedness of the 3- and 4-digit level data, which is even more severe for developing countries, makes it difficult to empirically test whether this scenario is occurring in practice.

What is feasible, however, is a check of whether the data given in any single year would theoretically allow for this case to happen. That is, the change in the wage share which would lead to a zero within-component for a given sector is multiplied with the within-component (which, for a conservative scenario, is assumed to remain unchanged). This increase in the within-component is then compared to the maximum possible inequality decrease in the opposite direction for the between-sectoral component. Mathematically, this is equivalent to comparing the following two elements in a simplified two-sector scenario:

$$\frac{\Delta Y_j}{Y} T_j = \frac{Y_j}{Y} ln(\frac{\frac{Y_j}{Y}}{\frac{N_j}{N}}) + (1 - \frac{Y_j}{Y}) ln(\frac{1 - \frac{Y_j}{Y}}{1 - \frac{N_j}{N}})$$

The first element is the difference of the new and the old wage share of the sector of interest j, which constitutes the weight for the (constant) within-sectoral inequality T_j. The second element, on the right-hand side, is the between-component. One can think of this example as a hypothetical 2-sector scenario in which all sectors (which, for simplification purposes, are assumed to not display any within-sectoral wage inequality) apart from the sector of interest are aggregated into one large sector. Although the interest here is in comparing changes in the two elements, because the minimum value for between sectoral inequality is zero, the maximum possible decrease is equal to the entire between-sectoral component at the 2-digit level - arguably, a rather unrealistic scenario, but nevertheless one which serves in proving the point that within-sectoral inequality trends can outweigh between-sectoral movements in inequality. What is not so unlikely is an increase in wages to a level that leads to a contribution of a large sector close to zero, given that the employment and wage shares are already relatively equal for sector 15 in many cases.

It turns out that there is only one single case in the data where in the above scenario it is theoretically possible that the first component outweighs the second after a change in the sectoral wage shares.[40] It is worth noting, however, that in a number of cases, the two effects - the inequality-decreasing effect of the between-component and the

[40]This is sector 15 in Rwanda in 1999. Note, however, that because there is only a single year of data, Rwanda is not included in the final dataset.

inequality-increasing effect of the within-component - almost cancel themselves out. The true decrease in overall wage inequality is therefore substantially lower than what it seems if only the change in the between-component is considered and the opposing effects of the within-component are ignored.

There are more reasons to believe that the between-component is a poor indicator of overall movements of between-sectoral inequality. First, it seems implausible that overall wage inequality drops to zero as a result of an increase in wages in one sector. The true decrease in the between-component in the above scenario is therefore likely to be much smaller, leaving more room for the (weighted) within-component to counteract this effect. Second, the within-component is still vastly underestimated even at the 4-digit level in many countries and years due to the unbalancedness in the raw data. This may be one of the reasons for why the above counterfactual exercise only yields a single case in which the within-component could outweigh the between-sectoral effect if the latter drops to zero. Third, adding to this underestimation, the within component is in all cases missing a further element due to the lack of individual-level data. Fourth, the assumption of a zero change in within-sectoral wage inequality was made to demonstrate the most conservative (and mathematically most simple) case of changes in the two components, where the change in the within-component was constructed to be minimal and the change in the between-component to be maximal. If the assumption of a zero change of the within-component is dropped as well and replaced with an increase in within-sectoral inequality - which is, after all, the scenario we are truly interested in - it is very likely that more cases can be identified in the dataset which have the potential to display divergent trends in between- and within-sectoral wage inequality. Going through different scenarios of changes in the within- and between-components is a tedious exercise, which can be circumvented by directly performing comparisons of the two components on the raw data. Despite the previously discussed limitations of comparing changes over time due to the unbalancedness of the 3- and 4-digit level data, conclusions can still be drawn from comparisons of the direction of changes of the within- and the between component given the following considerations. While it is clear that the size (and hence the share in overall inequality) of the within-component is affected by the availability and composition of subsectoral data, this does not affect the change observed in the between component. Assuming that the

pattern of missings is random across subsectors and it is not the case that sectors with higher within sectoral inequality are missing more (or less) often than those with lower sectoral inequality, looking at changes between years with similar subsectoral coverage can provide some indication of whether changes in the between- and the within-component go into different directions. Indeed, out of the 968 observations with changes of less than 1% in subsectoral coverage from one year to the next, around 13% (stemming from both developing and developed countries from all regions) show opposing trends of changes in overall and changes in between-sectoral coverage. The change in the within-component goes in the opposite direction than that of the between-component, and is large enough to outweigh its effect. This picture changes very little if only those cases are considered where sectoral coverage is entirely unchanged: again, around 14% of the 294 cases convey a wrong picture of the direction of change of inequality if only the between-component is considered due to opposing trends in within-sectoral wage inequality.[41]

In sum, there is strong indication that the Theil index relying on the between-sectoral component of manufacturing wages computed here and by the UTIP may provide a wrong image of overall changes in manufacturing wage inequality in around 13% of cases. Given that the "true" extent of within-sectoral inequality (taking into account individual-level data) is likely to be substantially larger, this number might just constitute a lower bound to the true discrepancy between between-sectoral and overall changes in manufacturing wage inequality.[42]

Of course, many more things can be done to assess the plausibility and extent of error of only looking at between-sectoral changes in inequality. Besides the counterfactual exercises on the UNIDO data discussed above, one could look into different country cases with better data for manufacturing wages, as done by Conceição et al. (2000) for the United States. This would allow, at least in some cases, the calculation of inequality up to the individual level and provide some indication of the remaining extent of inequality not captured by the sector-level data, no matter how detailed. However, doubts also arise

[41]The result also holds for different threshold of change in subsectoral coverage of between 5% and 50%. In fact, the share of 13% remains remarkably stable across all chosen thresholds.

[42]There are more cases in which within-sectoral inequality is increasing (587) compared to cases in which it is decreasing (453), which means that the importance of the within-component is likely to increase over time. One should also mention that not only does the between-component indicate the wrong direction of change in 13% of cases, but it very likely also understates changes by 100% or more in approximately the same number of cases (although here, the number is only around 7%).

about "external validity" of the index in the remainder of this paper, which, if taken seriously, limits its relevance to a narrowly defined set of applications focusing specifically at manufacturing. I therefore leave it up to those who have such a confined focus and need to take into account changes within sectors to assess this last component of within-sectoral inequality which remains unaccounted for here.

5 The relationship to overall income inequality

5.1 Comparisons with income inequality statistics

In an effort to validate the capacity of the Theil index to serve as a proxy for, and basis of, developments in overall monetary inequality, Galbraith and Kum (2005) (henceforth GK2005) relate it to the Gini coefficients compiled by Deininger and Squire (1996) (DS1996). Adding the share of employment in manufacturing as a control variable, and dummies for the different income concepts and accounting units underlying the DS1996 Gini coefficients, they find an elasticity of between 6 and 8.5 %. They provide three explanations for the rather small magnitude of this number: firstly, the inherent high sensitivity of the between-groups measure when the number of groups (and hence the absolute size of the measure) is small,[43] which makes the Theil very small in absolute terms and highly susceptible with respect to even small changes in the wage-/population share ratio. Secondly, they point to "the much greater volatility of the Theil measure [due to the varying number of manufacturing industries per year and country](GK2005: 128). And, thirdly, they mention the greater volatility of manufacturing pay compared with household income "[because it includes income from other sources such as non-labor wage, land and capital]" (GK2005: 128). The predicted relationship between their Theil index and the Gini coefficients is then used to scale up the Theil index and obtain a broad measure of income inequality, the "Estimated Household Income Inequality" (EHII) dataset. In a more recent update of their estimates, they confirm the relationship with the DS1996 data (Galbraith et al. 2014, 2015).[44] It should be noted that the estimates used for deriving

[43]This is potentially exacerbated by grouping sectors together to handle the emergence of new industrial categories.

[44]Note that in the 2014 update paper, it is not clear whether the estimates, presented in table 1, rely on a fixed or a random effects model. Also note that they do not use the WIID data, which is the successor to the DS1996 dataset, but continue to rely on the DS1996 dataset. However, DS1996 only covers years

the EHII from the Theil index stem from a simple OLS (without any control variables apart from the DS1996 dummies for the underlying income and accounting concepts), which implies that the underlying model does not only exploit changes within a given (country-)series, but that a major part of the variation stems from the variation in *levels* between countries. This can be considered problematic from two aspects: Firstly, because the number of covered sectors varies between countries (a problem which can be easily addressed by normalizing the measure by log n), this adds noise[45] to the estimates. Secondly, on a more conceptual level, the use of cross-country level variation is at odds with the initial acknowledgement (Conceição and Galbraith 2000: 64) of the limitation of the (between-)index to yield a complete picture of industrial wage inequality at each point in time, and that the interesting property of the index is in its capability to measure changes in total inequality - the "time evolution of inequality".

Notwithstanding these conceptual concerns, this paper also expands upon the initial approach of directly regressing measures of income inequality on the Theil index, which served as the basis for deriving the EHII. It tries to also explore the factors associated with the two measures being more (dis-)similar by regressing the difference between the Theil index and overall income inequality on a number of control variables. Instead of the Deininger and Squire (1996) dataset, which only contains data until 1996, the Gini coefficients from the World Income Inequality Database (WIID) provided by UNU-WIDER are used which comprise and extend the DS1996 data. All the controls proposed by GK05 are included and several other potentially important determinants of the association between manufacturing wage inequality and overall inequality are added to the model. To eliminate volatility stemming from differential sectoral coverage of the Theil index, the normalized version of the Theil index is used so that numbers are comparable in the random effects estimations between countries with differential sectoral coverage. To tackle another potential source of volatility in the income inequality measure, a few more control variables are added, and other, arguably more consistent, measures of income inequality

until 1996, and will therefore naturally lead to an estimate similar to the original one given that the new, updated years are not in the estimation sample of overlapping observations between the new UTIP index and the Gini coefficients. However, as shown in appendix table A.10, moving from the DS1996 to the WIID does not change the (FE) estimate much.

[45] That is, assuming that sectoral coverage is not systematically related to income inequality levels

are used in addition to the WIID.[46]

Most notably, the Luxembourg Income Study (LIS) Gini coefficients are derived from harmonized primary micro data, which is currently considered the "gold standard" in terms of consistency and accuracy of the resulting inequality measures. Unfortunately, time and country coverage of the LIS data is still limited and the resulting sample size is correspondingly small. Nevertheless, the comparison with the LIS can be used to validate the results from the WIID. The second alternative measure of income inequality used are the SWIID Gini coefficients provided by Solt (2015). He addresses the inconsistencies in the WIID, arising from the previously discussed heterogeneity of data sources by providing multiple solutions to mitigate the same and combining them into a single workable dataset.[47] The result is a balanced multiply imputed dataset of broad country and time coverage. Although the underlying method has been criticized (Jenkins 2015), the SWIID certainly provides a more sophisticated, prudent, and explicit way of making the WIID data comparable, especially when compared to the much more crude alternative of merely introducing dummy variables for the numerous categories of income and other underlying concepts on which the WIID relies. That the latter approach imposes constant differences between concepts across countries and over time is just one of its problems[48] and has been argued to be invalid (Atkinson and Brandolini 2009, Galbraith and Kum 2003). Two further data sources are added: the EU SILC data and the Gini coefficients from the World Development Indicators (WDI). They are employed because Galbraith et al. 2015 use them (in addition to the LIS figures and Ginis from the OECD) to validate their EHII data by comparing the final values from the upscaled Theil index to the Gini coefficients from these datasets, without testing the underlying relationship.

[46] It should not go unmentioned that the creators of the EHII have put their resulting estimates through a number of validity checks and comparisons with other data on income inequality, including the LIS (Galbraith et al. 2014, and 2015). They have not, however, repeated the initial exercise of relating the different data sources directly to the UTIP-UNIDO index of wage inequality.

[47] Importantly, it should be noted that the SWIID uses other data sources to cross-check its values, among them the UTIP-UNIDO Theil index. One might therefore suspect a built-in association between the index calculated in this paper and the SWIID Gini coefficients which is closer than for the other data sources. As shown in table 4, this is clearly not the case and suggests that the use of the SWIID is unproblematic for the present purpose.

[48] Another problem is how to deal with multiple observations per country and year of the same quality. Here, the researcher faces a trade-off between various dimension, e.g., sacrificing demographic for geographic coverage. Approaches which directly adjust the WIID Ginis by adding or subtracting the average differences between the underlying categories, as e.g. in Gruen and Klasen (2012) and Easterly (2007) do not circumvent the problem, either, since differences remain also for the adjusted Ginis.

The three control variables used in GK2005 are the ratio of manufacturing employment to population, the share of urban population, and the population growth rate.[49] Apart from measuring the importance of manufacturing for overall incomes, the share of manufacturing is also supposed to capture the part of the labor market which tends to be more unionized, and is therefore expected to be associated with lower inequality. Urbanization is expected to be associated with more inequality because "wealthy people live in cities," and population growth serves as a proxy for the age structure of a country and the composition of households, and is expected to be associated with higher inequality because a high population growth rate should, on average, imply a younger population.

Instead of the ratio of manufacturing employment, the share of value-added in GDP from manufacturing is used here, which features a good coverage of the countries in the sample.[50] The variable, along with the share of urban population (urban) and population growth is taken from the WDI (2016). In addition to these variables, three more controls are added. The first one is the price level of investment, taken from the Penn World Tables (PWT, V8.1, Feenstra et al. 2015). The variable is a proxy for the rate of returns of capital, and since capital is a component of overall income, higher returns to capital might increase the divergence of the two measures.

GK2005 argue that one of the reasons why changes in manufacturing wage inequality will likely not counteract developments in overall wage inequality is that low-skilled workers, forming the lower end of the distribution in manufacturing wage inequality, are substitutes for low-skilled workers in other sectors such as agriculture and services. It is therefore unlikely that wages at the lower end of manufacturing pay decrease or increase without an equivalent shift in the wage levels of other sectors of the economy. That being said, the same logic does not apply to the upper end of the wage spectrum, where workers are skilled in a specific profession and are much less likely to easily switch between manufacturing and other sectors. To also account for changes at the upper end of the wage distribution, a measure of total factor productivity (tfp) is included which is constructed to reflect cross-country differences in aggregate technology (Feenstra et al. 2015). As the

[49]Note, however, that for arriving at the EHII from the UTIP-UNIDO Theil index, estimates from a regression controlling only for the WIID category dummies and the population growth rate were used.

[50]The variable leads to very similar coefficient estimates as found in Galbraith and Kum (2005) when their index is used on the WIID data. Results can be found in appendix table A.10.

technological frontier of a country shifts outwards, this is likely to encompass all sectors of the economy and affect skill premia everywhere. Technological change can therefore be assumed to sway manufacturing wage inequality and overall income inequality in the same (upward) direction.

Naturally, both of these mechanisms are weakened by the extent of openness of an economy. While "cheaper" foreign workers may - indirectly through trade - be substitutes for some low-skilled manufactures, the same cannot be said for non-tradable services or some segments of agriculture. Similarly, countries can gain access to technology through trade, which may again affect the tradable sectors more than the non-tradable ones. Interactions between the tfp and the openness variable are supposed to capture these relationships. I also include an interaction between openness and the price level of investment variable given that in more open economies, capital markets also tend to be more liberalized and returns to capital can also stem from abroad, thereby weakening the wedge it might drive between manufacturing and income inequality. In addition to these effects, trade openness (defined as import and export value over GDP and taken from the WDI) is also a proxy of the extent to which a country is vulnerable to external shocks which are likely to affect overall (income) inequality much more and cause divergences from the wage inequality measure. To generally account for shocks which potentially affect all countries, year dummies are added to the model as well. Lastly, since many of the above control variables are correlated with GDP per capita, it is included to make sure that its effect gets picked up separately.

In addition to these "external" variables, the number of imputations is added to the regression to account for the fact that in the case of linear interpolation, the idea was to be as conservative as possible in mapping observed changes in employment and wages in the underlying sectors over the missing years. Consequently, actual changes which may show up in the overall income inequality statistics are less likely to be captured in those cases where imputations were necessary. The impact of the second "internal" variable, sectoral coverage of the wage inequality measure (# of ISIC), is time-invariant and cannot be estimated with the fixed effects approach. Therefore, random effects estimations are employed additionally to get an idea of the role of this variable as well as of the impact of controlling for all other country-specific time-invariant factors (results are shown in the

appendix).

Before moving to the analysis of the deviations between the two measures, the specification by GK2005 is replicated using my newly constructed index and the WIID instead of the DS1996 data. On the one hand, this serves as a check as to whether the new Theil index yields results similar to theirs. On the other hand, it tests whether their results also hold with the extensions discussed above which will be used in the analysis of the deviations of the measures, and with the WIID. Having reduced the two sources of volatility identified by GK2005, the relationship between the inequality data and the Theil index should be stronger in general, and in particular with the newly added, more consistent income inequality measures. Table 4 contains the fixed effects results for the expanded specification of GK2005 and features the estimates from their paper in the first column for better comparison.[51] As in their model, the Theil index enters in logs to simplify interpretation and to account for its log-normality (as shown in appendix figure A.3).[52] The random effects results are displayed in appendix table A.8 and do not show major changes in the results. Note, however, and in line with the conceptual concerns expressed at the beginning of this section, that the Theil index should not be employed as a representation of *overall* manufacturing wage inequality in random effects models in general, given that within-sectoral is not accounted for and the measure hence massively understates overall manufacturing wage inequality *levels*. Fixed effects models on the other hand only consider mean-deviations over time, which, if one follows the argumentation of Conceição and Galbraith 2000, is appropriate for the index' capability to trace *changes* in overall manufacturing wage inequality - although the accuracy thereof is questionable as well (see section 2).[53]

First, comparing columns 1 and 2 of table 4, the coefficients on the Theil index are

[51] Note that these are not based on estimations done in this paper, but they are literally the numbers published in table 5 of their paper.

[52] In addition to the changes in the model described above, it also contains dummy variables for the underlying categories in the WIID data. Note that the results do not change much when the full set of control variables as described above is included (the fixed effects results can be found in appendix table A.7). The largest change in coefficients is triggered by the inclusion of the GDP per capita variable, which affects the estimates of the control variables, but not that of the Theil index (results available upon request).

[53] Although the fixed effects model is clearly preferable due to the removal of time-invariant country specific factors, the random effects model is estimated to be able to compare also the random effects estimates by GK2005, and to get a benchmark estimate of the effects of the time-invariant factors on the income inequality measures to be able to better interpret the results from the next specification trying to explain the differences between the Theil index and the income inequality measures.

substantially lower than those found by GK2005 for the UTIP measure. This is not due to the use of the DS1996 instead of the WIID data in column 1: very similar coefficient estimates are obtained when the UTIP index is regressed on the WIID instead of the DS1996 data (results for the reduced model can be found in appendix table A.10.) Despite the larger sample size, significance disappears in the fixed effects model, and drops to the 5 percent level in the random effects specification, as shown in column 2 of appendix table A.8. The coefficient remains small and insignificant for all other inequality measures. Ironically, the only variable displaying similar effects as in the GK2005 estimations is the importance of the manufacturing sector, which is based on a measure different from theirs. The negative coefficients are more in line with the interpretation of the manufacturing sector as a proxy for the extent of unionization than as a mediating variable capturing the role of manufacturing wage inequality for overall income inequality, but are very small throughout.

While the results look rather similar for most measures, many coefficients change drastically when the LIS data are used. Most notably, the sign on the Theil index becomes negative (although the standard error is very large). The other variables population growth and urbanization also change substantially and are significant in some cases, despite the small sample size. When using only the LIS countries in the other specifications, it becomes clear that this is entirely due to the sample composition.[54] The fact that these differences arise between different samples despite the fact that country fixed effects are contained in the model also puts into question the universality of the relationship between the two measures for all countries.

In terms of the "internal variables," the effect of the time-invariant variable sectoral coverage can be seen in the random effects results (table A.8). The coefficient is small and only significant for the LIS data. This is reassuring given that the Theil index has been normalized with the underlying number of sectors for the random effects estimations already. It does not seem to be the case that countries with better sectoral coverage systematically differ from those with worse coverage, even when many time-invariant factors are not controlled for. This, again, supports the stance that the missings in the underlying

[54]Results available upon request.

Table 4: Relationship between Theil and income inequality: FE results, reduced model

	(1) GK05	(2) wiid	(3) swiid	(4) lis	(5) silc	(6) wb
ln(Theil)	0.079***	0.00909	0.0115	0.00468	0.0315	0.00354
	(6.60)	(0.0112)	(0.0117)	(0.0312)	(0.0187)	(0.0146)
Population growth	-0.578	-0.00217	0.0134**	-0.0252	-0.0287**	0.00701
	(-0.81)	(0.0205)	(0.00652)	(0.0247)	(0.0133)	(0.0163)
Share urban	0.001	-0.0134***	0.00270	0.00753	-0.00758*	0.00802*
	(-1.57)	(0.00403)	(0.00294)	(0.00737)	(0.00427)	(0.00424)
Manuf. value add.	-0.001***	-0.00522	-0.00307*	0.00177	0.00189	-0.00120
	(4.50)	(0.00377)	(0.00176)	(0.00574)	(0.00598)	(0.00335)
Constant	3.893***	4.421***	3.566***	-1.779***	3.992***	3.489***
	(51.38)	(0.301)	(0.0943)	(0.520)	(0.200)	(0.201)
Observations	481	633	1,765	121	256	538
Year dummmies	NO	YES	YES	YES	YES	YES
WIID dummmies	YES	YES	-	-	-	-
R^2	unreported	0.799	0.057	0.481	0.181	0.106
# of countries	81	71	100	36	28	87

Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logged Gini coefficient from the data source indicated in the top row. Swiid refers to net inequality from the SWIID database. Silc denotes the EU SILC data and wb the WDI Gini coefficients. The dummies for the underlying WIID categories are included in column 2 but are not shown to save space (available upon request).

sectoral data are random.[55]

Overall, these results do not lend much support for a continuing stable association between the Theil index of manufacturing and overall inequality, such as that found in GK2005 for their UTIP-UNIDO index. While the findings for the WIID data are qualitatively still relatively similar to their estimates, other, and arguably more consistent, measures of income inequality yield rather different results. Not only does the association with the Theil index become insignificant, but the coefficients are also too small to postulate any economically meaningful link between the two variables. This is true even for the WIID specification, where according to the (insignificant) WIID fixed effects point estimate, doubling the Theil index would lead to an increase in the Gini coefficient of little under one percent. Given the high R^2 of 0.8 in the specification, the low association does not seem to be the result of an incomplete or widely misspecified model, either. One reason for the weak association of the between-sectoral Theil index and overall measures

[55] As for the newly added control variables, the model containing these additional variables is shown in appendix table A.7, which also displays the year dummies. Adding the control variables causes some of the coefficients of the Theil index to increase, but does not render them significant. The overall explanatory power of the model has increased in all cases, and more than doubled for those using the SWIID, the SILC, and the World Bank data.

of income inequality could be the neglect of the within-component. It might be worthwhile to repeat the exercise with manufacturing wage inequality measures using more detailed sector- or individual-level data. That the UTIP index displays a stronger link with the income inequality measures could also be owing to the fact that it partly relies on earlier industrial classification schemes with higher levels of detail.

Given that there are good theoretical reasons to expect a robust relationship between manufacturing wage inequality and overall income inequality, an explicit analysis of the factors which might cause the two measures to differ stands to reason. All of the theoretically motivated variables discussed above are included in the model, along with the full set of year dummies and, for the WIID data, the underlying categories. The dependent variable is the logged percentage difference between the (normalized) Theil index and the respective Gini coefficient, as indicated in the top row of table 5. The logarithmic transformation is used, on the one hand, to make interpretation easier, and on the other hand, because the differences are approximately log-normally distributed (see appendix figure A.4). The GDP per capita, trade openness, total factor productivity, and price level of investment variables also enter in logs, for the same reasons.

Most of the results of the control variables match the theoretical predictions derived for the variables above, although not all coefficients are significant. To begin with, a higher share of manufacturing value-added is associated with a higher discrepancy of the Theil index and income inequality for all data sources except the WIID, but is insignificant. In line with the interpretation that the variable is capturing the extent of unionization, one way of reading this result is that with a larger manufacturing sector, a higher share of the economy is isolated from other (dis-)equalizing forces which drive up overall income inequality, but not wage inequality.

The population growth and urban population variables are positive except in the LIS and SILC specifications, and population growth is significant with the WIID and the World Bank samples. This is consistent with the UTIP story for including the variables into their models.

Trade openness is associated with a higher similarity between income inequality and the Theil index for all measures, and is significant for most, with relatively stable point estimates across the six models. What is more, the absolute effect is sizeable: A ten

Table 5: Determinants of the difference between wage and income inequality, FE results

	(1) wiid	(2) swiid	(3) lis	(4) silc	(5) wb
GDP per capita	0.551*	0.228	0.695	1.007**	0.646
	(0.319)	(0.267)	(0.678)	(0.429)	(0.544)
Population growth	0.137**	0.0702	-0.00705	-0.0790	0.136**
	(0.0675)	(0.0467)	(0.134)	(0.0514)	(0.0656)
Share urban	0.0272	0.0117	-0.0106	-0.0225	0.0186
	(0.0170)	(0.0154)	(0.0706)	(0.0183)	(0.0193)
Manuf. value-added	0.00290	0.0108	0.0358	0.000163	0.0128
	(0.0144)	(0.00768)	(0.0472)	(0.0261)	(0.0121)
Trade openness	-0.655**	-0.440*	-0.521	-0.353	-0.528**
	(0.258)	(0.231)	(0.354)	(0.285)	(0.237)
Price level inv.	0.570	0.258	0.379	2.134**	1.351
	(0.730)	(0.609)	(1.437)	(0.850)	(1.029)
Tfp	-9.597**	-2.936*	-1.475	-22.95***	-5.942**
	(4.233)	(1.482)	(5.061)	(5.200)	(2.283)
Open.*p.l. inv.	-0.237	-0.00689	-0.122	-0.539**	-0.378
	(0.187)	(0.143)	(0.382)	(0.201)	(0.254)
Open.*tfp	2.376**	0.651	0.372	5.095***	1.419**
	(0.954)	(0.399)	(1.221)	(1.146)	(0.572)
# imputed	-0.0337**	-0.0599***	-0.000321	-0.0294***	-0.0573***
	(0.0162)	(0.0149)	(0.0139)	(0.00717)	(0.0178)
Constant	2.260	7.483***	-0.528	1.512	4.126
	(3.002)	(1.863)	(7.138)	(4.883)	(5.114)
Year dummies	YES	YES	YES	YES	YES
WIID dummmies	YES	-	-	-	-
Observations	618	1.521	120	256	483
R^2	0.419	0.353	0.562	0.374	0.405
# of countries	66	82	35	28	73

Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logged percentage difference between the (normalized) Theil index and the Gini coefficient from the data source indicated in the top row. Swiid refers to net inequality from the SWIID database. Silc denotes the EU SILC data and wb the WDI Gini coefficients.

percent increase in the openness ratio implies a 3.5-6.5% lower dissimilarity of the Theil index of inter-industry wage inequality and the Gini coefficient of income inequality.

The tfp variable, capturing the level of technology, is associated with a smaller gap between wage and income inequality in all specifications and is significant in most. Its effect is large[56] compared to that of the other variables, and it would appear that technological change is affecting both wage and income inequality in the same way. However, as an economy becomes more open, the gap grows larger again with higher tfp values - although

[56]A ten percent increase in tfp would imply a lower difference between the Theil index and the indices of income inequality of between 1.5 and 23%. The very high value in the SILC sample might have to do with the fact that technological progress has a much lower variation in this rather homogeneous sample of developed European countries, where technological progress has stabilized at rather high levels. Also note that the counteracting effect of the openness variable is much stronger here and since most European countries are open, the large effects partly cancel out.

at a smaller rate.[57]

In line with expectations, the price level of investment - proxying for returns to capital - is associated with a larger difference between the Theil index and the Ginis. However, the variable is only significant with the SILC measure. This association is weaker the more open the economy, and again, this effect is individually significant only for the SILC sample. However, when tested together, the price level of investment variable and the openness interaction are jointly significant in all models except the LIS one.

A surprising finding is that the variable measuring the extent of imputation in the underlying data points is negative and significant in the SWIID and the World Bank Gini specifications. One interpretation of this finding is that, because of the attempt to tamper with the data as little as possible, the Theil index tends to show small changes in inequality in years with more imputations in the sectoral data. Because income inequality is more sluggish than wage inequality, this could mean that the smoother series has a closer resemblance to the developments in income inequality than the more erratic one. Another explanation is that countries where manufacturing wage inequality is closer to overall inequality have more missing data points, which is somewhat puzzling. As for the second "internal" variable, sectoral coverage (shown in the random effects results in table A.9), it is positive, but significant only for the WIID. Considering the fact that it effectively ranges from 0 to 18, the effect is rather large. According to the point estimates, the inclusion of one additional sector in the Theil index is associated with an 8 percent higher difference between wage and income inequality. One would expect that a better sectoral coverage would lead to more accurate numbers of between-sectoral wage inequality and, because wage inequality is a constitutive part of income inequality, this would lead to lower average differences between the two measures.

It is also interesting to note the value of the constant. Focusing on the fixed effects specifications, there appears to be a "baseline" difference between wage and income inequality only for the SWIID specification of around 7.5 percent. For the other measures, the constant is rather small, and insignificant - for the LIS, it is even negative. Hence, there does not seem to be a universal, inherent difference between wage- and income inequal-

[57] In model 1, for example, the effect at the openness value of 100 is 1.33 (significant at the 5% level). It becomes significant both statistically and economically at values above 100, e.g., at an openness value of 125, the coefficient is 1.4 and significant at the 10% level.

ity. Apart from Europe and Central Asia, which display consistently larger differences between the two measures across all specifications, no new insights are obtained about the baseline difference between wage and income inequality from the regional dummies shown in the random effects results (see table A.9, the reference category being East Asia and the Pacific). The year dummies do, however, indicate that the differences between the wage inequality index and the income inequality measures are decreasing over time for all measures except the WIID Ginis.

The last thing worth mentioning is that although almost nothing is significant in the LIS specifications, the included variables explain over 50% of the variation in the differences between the two measures. Given that the LIS data are of high quality and the most consistently measured data source of the four measures included, this suggests that the other measures still suffer from a substantial amount of measurement error. Apart from the abovementioned change in the industrial classification scheme to a new and more crude version, this might be the main reason for the lack of a robust association between the Theil index and other measures of income inequality in the previous estimations.

6 Conclusion

The core of this paper is the construction of a Theil index of between-sectoral wage inequality for the manufacturing industry, based on the UNIDO industrial statistics. A very similar index has been built by the University of Texas Inequality Project (UTIP) for the years 1970 to 2008; however, their UTIP-UNIDO index does not include within-sectoral inequality, it is not clear which sectors are included in the index every year, and there is no publicly available information on which cases (countries and years) rely on previous versions of the UNIDO industrial statistics directly, or through smoothing out differences with previous versions of the UTIP index. I have therefore recalculated the Theil index for the 48-year time period from 1963 to 2010 for which data was available at the time of writing of this paper. The index relies exclusively on the UNIDO industrial statistics (Rev. 3), which contain data at the 2-digit sector level. I provide detailed information with respect to the sectors covered in each country, as well as the imputation methods and further variables from the UNIDO incorporated into the same. I then make a rec-

ommendation as to which version can be used best in the context of dynamic empirical applications of wage inequality based on an analysis of different versions of the Theil index, reflecting the trade-off between time and sectoral coverage.

The narrow scope of the wage inequality measure, on the one hand, has the advantage of being consistently defined across countries and years, but, on the other hand, restricts the applicability of the index. This paper argues that the latter point is one of the main drawbacks of the index, and presents evidence that its generalizability is severely limited. This applies not only to the extent to which the index allows conjectures about the overall level of income inequality in a society. There is reason to also question the "internal" capability of the index to accurately reflect developments in manufacturing wage inequality. Because it relies on sector-level data on wages and employment which is aggregated at the 2-digit level of industrial classification, the index only measures between-sectoral wage inequality and cannot give account of inequality within sectors. Using data provided at the more disaggregated 3- and 4-digit level, the potential magnitude of within-sectoral inequality is estimated to be at least 40 percent of overall manufacturing wage inequality. Moreover, I find that the between-sectoral index is not generally able to trace changes in between-sectoral inequality over time very well, as opposed to what has been argued by Conceição and Galbraith (2000) and Conceição et al. (2000). Using more detailed sector-level data, I find that looking only at changes in between-sectoral inequality leads to an erroneous image of the direction of change in overall manufacturing wage inequality in around 13% of cases. Given that the sector-level data still do not account for individual inequality within sectors, the true error might be larger and remains open to further exploration.

The analysis of the "external" validity of the index, that is, the extent to which the Theil index is representative of overall income inequality, builds on prior work by Galbraith and Kum (2005), and Galbraith et al. (2015). The authors argue in favor of a stable relationship between the narrowly defined Theil index of wage inequality and the Gini indices of income inequality provided in the WIID and other data sources, which comprise other components besides labor market income. Their finding of a stable relationship between the two concepts cannot be confirmed for my index in a broader setting which employs several additional, arguably more consistent, measures of income inequality. Going one

step further, this paper tries to find out what causes the measures to show such a weak association with the Theil index, given that there are good theoretical reasons to expect a strong link between them. The deviations between the Theil index and the other measures of income inequality are regressed on a number of potential explanatory variables. The explanatory power of this model is under 50% in most cases, and measurement error as well as the shortcomings of the Theil index itself in capturing overall manufacturing wage inequality rather than just its between-sectoral component are likely candidates for the weak association, both in the model looking at levels and the one examining the difference between the two inequality concepts.

In sum, while a measure of between-sectoral wage inequality certainly has its merits and is a valuable resource for empirical analyses with a focus on manufacturing and/or the development of industrial sectors, the general applicability of the index appears to be much more limited than suggested elsewhere. Although all conclusions drawn only apply to the inequality numbers based on the most recent industrial classification scheme used in the UNIDO industrial statistics, given that newly added years after 2010 will - at least for now - be in the new classification scheme as well, the relevance of these results will be growing as the time coverage of the index is extended to more recent years.

References

Andridge, R. R. and Little, R. J. (2010). A review of hot deck imputation for survey non-response. *International Statistical Review*, 78(1):40–64.

Atkinson, A. B. and Brandolini, A. (2009). On data: a case study of the evolution of income inequality across time and across countries. *Cambridge Journal of Economics*, 33(3):381–404.

Conceição, P. and Ferreira, P. (2000). The young person's guide to the theil index: Suggesting intuitive interpretations and exploring analytical applications. *UTIP working paper No.14*.

Conceição, P. and Galbraith, J. K. (2000). Constructing Long and Dense Time-Series of Inequality Using the Theil Index. *Eastern Economic Journal*, 26(1):61–74.

Conceição, P., Galbraith, J. K., and Bradford, P. (2000). The theil index in sequences of nested and hierarchic grouping structures. *Eastern Economic Journal*, 26(1):61–74.

Deininger, K. and Squire, L. (1996). A new data set measuring income inequality. *The World Bank Economic Review*, 10(3):565–591.

Easterly, W. (2007). Inequality does cause underdevelopment: Insights from a new instrument. *Journal of Development Economics*, 84(2):755–776.

Feenstra, R. C., Inklaar, R., and Timmer, M. P. (2015). The next generation of the Penn World Table. *The American Economic Review*, 105(10):3150–3182.

Galbraith, J. K., Choi, J., Halbach, B., Malinowska, A., and Zhang, W. (2015). A comparison of major world inequality data sets. *UTIP Working Paper No.69*.

Galbraith, J. K., Halbach, B., Malinowska, A., Shams, A., and Zhang, W. (2014). UTIP global inequality data sets 1963-2008: updates, revisions and quality checks. *UTIP Working Paper No.68*.

Galbraith, J. K., Jiaqing, L., and Darity Jr, W. A. (1999). Measuring the evolution of inequality in the global economy. *UTIP Working Paper No.7*.

Galbraith, J. K. and Kum, H. (2003). Inequality and Economic Growth: A Global View Based on Measures of Pay. *CESifo Economic Studies*, 49(4):527–556.

Galbraith, J. K. and Kum, H. (2005). Estimating the inequality of household incomes: a statistical approach to the creation of a dense and consistent global data set. *Review of Income and Wealth*, 51(1):115–143.

Gruen, C. and Klasen, S. (2012). Has transition improved well-being ? *Economic Systems*, 36(1):11–30.

Jenkins, S. P. (2015). World income inequality databases: an assessment of WIID and SWIID. *The Journal of Economic Inequality*, 13(4):629–671.

Solt, F. (2015). On the assessment and use of cross-national income inequality datasets. *The Journal of Economic Inequality*, (4):1569–1721.

Theil, H. (1967). *Economics and Information Theory*. North-Holland, Amsterdam.

3.A Appendix

Table A.1: Deviations between the long and short versions of the Theil index by country

Country	LONG VERSION							SHORT VERSION							# of sectors dropped in long version	Recommended version for dynamic analysis
	Max. absolute deviation	Mean absolute deviation	Std. dev.	Corr. with full version	Corr. of diff. with full version	# of years with dev.	Mean Theil index	Max. absolute deviation	Mean absolute deviation	Std. dev.	Corr. with full version	Corr. of diff. with full version	Mean Theil index	# of years with dev.		
CAF	4628.98	1142.2	1489.72	0.05	0.14	8	0.0296	0					0.0604		1	Short
MDA	4546.43	483.7	1005.23	0.22	0.27	25	0.0164	0					0.0919		11	Short
NLD	691.79	81.8	181.68	0.35	0.03	14	0.0103	0					0.0254		5	Short
NZL	649.09	128.9	133.09	0.93	0.96	24	0.0415	0					0.1268		4	Long
BWA	520.4	161.8	130.33	0.96	0.92	13	0.0964	3.54	0.73	1.75	1	1	0.2419	8	11	Long
ROU	402.39	75.6	90.66	0.53	0.92	21	0.0538	0					0.0472		6	Long
HUN	359	27.2	85.61	0.09	0.51	18	0.0246	0					0.0568		5	Short
MUS	235.51	24.4	52.81	0.98	0.96	43	0.0586	0					0.0683		4	Long
MLT	225.96	60.3	62.46	0.68	0.47	14	0.0128	0					0.029		7	Short
MDG	218.44	81.7	68.62	0.96	0.92	29	0.0463	0					0.0427		5	Long
JAM	192.6	56.5	65.51	0.77	0.64	44	0.15	11.89	4.18	4.85	0.99	0.99	0.2936	23	7	Short
SWE	102.18	16.2	23.8	0.91	0.41	20	0.0054	0					0.0083		5	Short
FIN	97.21	21.9	31.3	0.76	0.8	19	0.0107	0.21	0.03	0.11	1	1	0.0125	15	5	Short
MEX	96.72	17.2	28.82	0.96	0.87	22	0.0517	0.04	0.04				0.0641	1	5	Long
MNG	76.36	13.4	21.2	0.91	0.9	17	0.0842	2.72	-0.36	0.82	1	1	0.09	16	5	Long
DZA	60.09	20.1	25.27	0.83	0.78	14	0.0165	0					0.008		10	Short
LSO	59.87	47.8	9.72	0.81	0.49	9	0.2062	0					0.1371		5	Short
MWI	57.98	15.2	16.39	1	0.98	32	0.0948	6.09	0.59	1.76	1	1	0.0812	25	4	Long
AZE	51.08	18	14.36	0.92	0.86	16	0.1113	0					0.1608		5	Long
AUS	49.72	14.1	19.94	1	1	14	0.0526	0					0.1456		4	Long
FRA	46.83	24	10.49	0.86	0.8	20	0.0285	0					0.0177		5	Short
SWZ	43.11	9.4	14.72	0.97	0.94	11	0.1115	0					0.1003		7	Long
BRA	40.89	8	9.4	0.7	0.39	15	0.124	0					0.122		5	Short
GBR	40.48	9.6	9.82	0.86	0.83	17	0.0144	0					0.0182		5	Short
SGP	40.35	32.8	4.08	0.95	0.66	20	0.0598	2.7	-1.53	0.68	1	1	0.0374	20	5	Short
AUT	37.14	14.5	11.25	0.77	0.77	20	0.0166	0.94	-0.07	0.4	1	1	0.0197	15	5	Short
ALB	29.94	7.5	12.23	0.98	0.98	11	0.0689	0.65	0.06	0.61	1	1	0.1045	3	9	Long
SVK	28.02	11.7	8.05	0.98	0.84	17	0.0249	0					0.0277		1	Short
HRV	27.2	8.6	6.92	0.95	0.89	14	0.0296	0					0.0406		5	Long
MYS	26.59	19.8	3.27	0.89	0.77	11	0.0332	0					0.0342		5	Short
POL	26.44	6.2	8.71	0.96	0.87	18	0.0154	0					0.0282		5	Long
CHN	24.94	13.6	9.95	0.84	0.99	8	0.0785	0					0.0292		5	Long
HTI*	24.35	10	11.88	0.94	0.99	10	0.104	3.48	2.04	0.6	1	1	0.1096	10	2	Long
ESP	23.75	11	4.91	0.98	0.92	17	0.0276	0					0.0277		5	Long

UGA*	22.83	6.4	8.48	1	1	23	0.109						0.1092		4	Long
KOR	20.83	11	5.43	0.98	0.87	17	0.024	0					0.0212		5	Long
SUR*	19.11	3.5	5.97	1	0.98	19	0.0488						0.0514		7	Long
SEN*	18.46	7.7	7.26	0.92	0.97	5	0.041						0.0418		5	Long
ISR	17.23	4	5.2	0.99	0.92	46	0.0485	0					0.0713		4	Long
TTO*	17.14	2.7	5.86	1	0.99	8	0.1525						0.1569		5	Long
TON*	16.33	4.9	6.66	0.99	0.98	9	0.0626						0.0627		5	Long
MAC	16.27	6.1	5.31	1	0.99	11	0.014	0					0.0258		2	Long
PRT	15.37	10.3	2.67	0.99	0.98	15	0.0489	0					0.0508		5	Long
GRC	14.14	6.1	3.46	0.98	0.93	15	0.0284	0.07	-0.03	0.05	1	1	0.031	6	5	Long
PHL	14.08	9.1	5.21	0.99	0.98	13	0.055	0					0.0503		5	Long
SVN	13.63	2	3.59	1	0.98	24	0.0246	2.41	-1.08	0.79	1	1	0.0275	11	6	Long
ITA	13.6	6.2	4.2	0.93	0.98	19	0.0179	0					0.0186		5	Long
NOR	13.55	6	7.27	0.99	0.99	18	0.0107	0					0.014		5	Long
ECU	13.06	5	4.36	0.99	1	14	0.0408	0					0.0547		5	Long
LUX	12.45	4	4.98	1	1	20	0.03	0					0.039		4	Long
IDN	11.35	4	3.94	1	1	20	0.0854	0					0.0968		6	Long
BOL	11.3	3.8	5.55	0.92	0.94	7	0.0543	0					0.0789		5	Long
TUN	11.14	4.9	4	1	1	33	0.1527	0					0.1276		4	Long
IRL*	10.61	5	4.6	0.98	0.91	19	0.0166	10.48	-4.02	3.37	0.99	0.97	0.0154	19	5	Long
COUNTRIES WITH ABSOLUTE DEVIATIONS OF LESS THAN 10%																
USA	9.11	5.3	4.97	0.98	0.95	11	0.0251								5	
JPN	8.76	3.6	3.15	1	0.99	26	0.0418								6	
TUR	8.6	3.2	2.42	1	1	18	0.0505								5	
MKD	8.37	2.8	3.73	0.99	0.97	11	0.0634								5	
JOR	8.1	3	3.21	0.99	1	37	0.0839								6	
UKR	8.08	5.8	2	1	0.99	11	0.0465								5	
DNK	7.97	2.9	3.64	0.99	0.99	18	0.007								5	
EGY	7.82	2.1	2.91	1	0.99	14	0.0523								5	
LBY	7.6	7.6				1	0.0376								1	
THA	7	4.9	2.8	1	0.98	11	0.0578								5	
PNG	6.97	2.8	2.19	1	1	15	0.0806								2	
MAR	6.89	3.6	2.25	1	0.98	12	0.0911								6	
COL	6.71	4.2	1.35	1	0.99	11	0.0346								4	
LKA	6.44	4.2	4.12	1	1	5	0.062								5	
PSE	6.36	4	3.65	1	0.99	14	0.0377								6	
ZAF	6.34	3.5	1.57	1	0.99	20	0.0566								5	
ARG	6.24	2.9	1.47	1	1	10	0.0524								4	
PER	6.18	4	1.75	1	0.99	12	0.2091								5	
BRB	5.57	2.3	2.52	0.99	0.99	11	0.055								5	
CZE	5.05	1.1	1.28	1	0.99	20	0.0099								6	
URY	5.02	3.1	1.4	1	1	11	0.0452								5	
CAN	4.97	2.1	1.88	0.99	0.99	21	0.0183								4	
PAN	4.9	1.3	1.82	1	1	14	0.0589								2	
SYR	4.87	2	2.28	1	1	15	0.1205								4	
TZA	4.81	3.5	0.98	1	1	5	0.0803								2	

BGD	4.76	3.5	3.42	1	1	4	0.0299	5
BGR	4.68	1.6	1.97	1	1	16	0.0841	6
KGZ	4.62	1.3	1.73	1	1	17	0.1636	5
KWT	4.42	3.6	0.74	1	0.89	3	0.2844	5
IND	4.33	1.4	1.94	0.98	0.95	12	0.0778	5
NPL	4.04	1	1.89	1	1	6	0.0635	9
CYP	3.97	1.5	2	1	1	12	0.0286	5
IRN	3.54	1.6	0.86	1	1	12	0.035	5
BEL	3.28	1.1	1.38	1	1	15	0.0566	5
FJI	2.87	0.9	1.06	1	1	32	0.5819	5
SOM	3.16	3.2				1	0.044	2
CHE	2.83	1.7	1.11	1	1	3	0.0229	1
CIV	2.58	1.9	0.7	1	1	4	0.052	2
QAT	2.06	0.9	0.62	1	1	11	0.3886	7
GEO	1.94	0.6	0.85	1	1	5	0.061	2
PAK	1.64	1.6				1	0.0556	4
CHL	1.55	1.2	0.71	1	1	8	0.0606	4
RUS	1.44	0.6	0.45	1	1	9	0.0494	4
PRY	0.72	0.7				1	0.0277	3

Notes.

HTI: Sectors 16, 18, 28, 29, and 34 have imputed values for years 1988-1997. A short version of the index (including sectors 19 and 32) would therefore mean that those other sectors should be dropped, and consistency would still not be established. One alternative would be to have two different, non-comparable short versions for Haiti: one from 1988-1997, and one from 1968-1987. The long version of the index displays larger deviations of around 20% only in the last 4 years (1994-1997), despite the fact that the same 2 sectors are omitted throughout - sectoral composition is therefore not driving the differences in the deviations between the long- and full versions. Because of the high correlation between the two measures over time in both levels and differences, retaining the long version seems justifiable.

UGA: For Uganda, the long version is retained despite deviations of up to 22% in first few years since including the sector causing this deviation (27) would effectively mean a time coverage of only four years from 1963-1966. Additionally, the contribution of the sector causing the deviation is vanishing over time and in the second time spell where data for the sector is present, the deviations are very small (between 2 and 7.5%), in line with the decrease in absolute size of the sector. Additionally, Uganda has a low average deviation of around 6% and almost perfect co-movement of the long- with the full version over time, as indicated by correlation coefficients which round up to one at 2 digits.

SUR: In Suriname, the index would decrease from 20 to only four years of data coverage in the short version. I have therefore decided to keep the long version given that the maximum deviation of 19% only arises in the first year of data (1974) and keeps decreasing thereafter to around 13% in 1975 and 1976 and 10% in 1977. Assuming that the downward trend continues, sacrificing 16 years of data for achieving higher accuracy of supposedly less than 10% seems unreasonable. The very high correlation of the long- with the full version in both levels and differences also supports the long version.

SEN: In Senegal, keeping the sectors causing the deviation of around 18.5% would leave only 5 years of data (1998-2002). Given that only a single year has such a high deviation (again, this is not because more sectors are omitted in that year) and the correlations over time are fairly high, the long version is retained.

TTO: In Trinidad and Tobago, only a single year (1998) is causing the deviation of around 17%. Upon closer inspection of the data, this deviation can be traced back to what is likely to be a glitch in the data, with employee numbers in sector 35 suddenly dropping to 16 (160 being a much more reasonable number) before rising again to 176 in 1999. This drop is also not warranted by changes in any other variables, or by a similar drop in other sectors in that year.

TON: In Tonga, sector 35 is responsible for the one-year deviation of around 16% in 1991. The contribution of the sectors is decreasing thereafter and the deviations are very small. While this does not point towards a lower contribution of the sector in the years preceding 1991, keeping only the years 1991-2004 for which data are provided in sector 35 would lead to another problem: many other sectors have 0s for wages and employees in the later years, making the short index not very informative for the overall level and development of inequality in the country. Given the low average deviation of less than 5%, and the high correlation of the long- and full indices, retaining the long version therefore seems like the better option.

IRL: In Ireland, changing to the short version requires the dropping of two sectors (23 and 36) which have a lot of imputed data in the years covered by the "short" version. Omitting the dropped sectors would result in a deviation of approximately the same magnitude (10.5%), but fewer years (18 instead of 46).

Table A.2: Overview of the Theil index by country

country code	country	years	sectors	imputed	dynamic version	region	dev. status	Theil(normalized)	Theil	standard deviation Theil
AFG	Afghanistan	9	9	6	long	SA	developing	0.0023	0.00497	0.00519
ALB	Albania	18	12	40	long	ECANA	developing	0.0292	0.06891	0.05679
ARG	Argentina	19	18	72	long	LAC	developed	0.0182	0.05239	0.01377
AUS	Australia	44	18	184	long	EAP	developed	0.0181	0.05263	0.08744
AUT	Austria	20	18	80	short	ECANA	developed	0.0067	0.01973	0.00277
AZE	Azerbaijan	21	18	49	long	ECANA	developing	0.0386	0.11126	0.0596
BDI	Burundi	23	18	166	long	SSA	developing	0.0235	0.06206	0.0324
BEL	Belgium	47	18	130	long	ECANA	developed	0.0181	0.05659	0.06289
BEN	Benin	8	18	9	long	SSA	developing	0.0256	0.07825	0.01784
BFA	Burkina Faso	10	18	0	long	SSA	developing	0.0115	0.03322	0.02159
BGD	Bangladesh	32	18	144	long	SA	developing	0.0096	0.02987	0.02131
BGR	Bulgaria	48	17	55	long	ECANA	developing	0.0293	0.08411	0.04892
BLZ	Belize	4	16	24	long	LAC	developing	0.035	0.1149	0.07713
BOL	Bolivia	32	18	22	long	LAC	developing	0.0188	0.05431	0.02979
BRA	Brazil	15	18	636	short	LAC	developing	0.0394	0.12196	0.01522
BRB	Barbados	28	12	0	long	LAC	developed	0.0221	0.055	0.01579
BWA	Botswana	30	8	73	long	SSA	developing	0.038	0.0964	0.08864
CAF	Central African Republic	8	16	64	short	SSA	developing	0.0211	0.06039	0.01242
CAN	Canada	48	18	24	long	ECANA	developed	0.0063	0.01831	0.0037
CHE	Switzerland	11	19	158	long	ECANA	developed	0.0034	0.02285	0.01407
CHL	Chile	46	18	65	long	LAC	developed	0.021	0.06057	0.02375
CHN	China	34	18	294	long	EAP	developing	0.0272	0.07853	0.09696
CIV	Cte d'Ivoire	22	15	9	long	SSA	developing	0.0194	0.05199	0.01735
CMR	Cameroon	33	18	205	long	SSA	developing	0.0382	0.10795	0.06666
COG	Congo	21	14	142	long	SSA	developing	0.0222	0.06684	0.02836
COL	Colombia	48	18	78	long	LAC	developing	0.0119	0.03459	0.00721
CRI	Costa Rica	41	18	364	long	LAC	developing	0.0118	0.05152	0.02943
CUB	Cuba	15	14	42	long	LAC	developing	0.0015	0.00477	0.00293
CYP	Cyprus	48	18	30	long	ECANA	developed	0.0099	0.02861	0.00953
CZE	Czech Republic	21	17	38	long	ECANA	developed	0.0035	0.00988	0.00389
DEU	Germany	27	18	72	long	ECANA	developed	0.0015	0.00438	0.00563
DNK	Denmark	47	18	152	long	ECANA	developed	0.0022	0.00699	0.00289
DOM	Dominican Rep.	23	18	0	long	LAC	developing	0.0219	0.06321	0.02334
DZA	Algeria	14	8	48	short	MENA	developing	0.0028	0.00802	0.00273
ECU	Ecuador	46	18	2	long	LAC	developing	0.0141	0.04084	0.01833
EGY	Egypt	47	18	256	long	MENA	developing	0.0147	0.05229	0.04363
ERI	Eritrea	19	23	9	long	SSA	developing	0.0134	0.04191	0.02442
ESP	Spain	47	18	6	long	ECANA	developed	0.0095	0.02757	0.00788
EST	Estonia	19	22	337	long	ECANA	developed	0.0411	0.13342	0.16341

country code	country	years	sectors	imputed	dynamic version	region	dev. status	Theil(normalized)	Theil	standard deviation Theil
ETH	Ethiopia	20	23	7	long	SSA	developing	0.0117	0.0365	0.02189
FIN	Finland	19	18	24	short	ECANA	developed	0.0041	0.01247	0.00482
FJI	Fiji	42	15	276	long	EAP	developing	0.0203	0.05819	0.04232
FRA	France	20	18	339	short	ECANA	developed	0.006	0.01766	0.00208
GAB	Gabon	16	18	181	long	SSA	developing	0.0326	0.10511	0.04791
GBR	United Kingdom	17	18	171	short	ECANA	developed	0.0059	0.01825	0.00366
GEO	Georgia	13	21	20	long	ECANA	developing	0.0202	0.06097	0.0227
GHA	Ghana	25	18	0	long	SSA	developing	0.0328	0.09483	0.03213
GMB	Gambia	8	18	0	long	SSA	developing	0.0049	0.01419	0.00389
GRC	Greece	45	18	146	long	ECANA	developed	0.0098	0.02839	0.00401
GTM	Guatemala	31	18	180	long	LAC	developing	0.0284	0.07959	0.06715
HKG	Hong Kong	35	18	29	long	EAP	developed	0.0092	0.02642	0.03769
HND	Honduras	34	18	233	long	LAC	developing	0.0239	0.06183	0.03547
HRV	Croatia	25	18	12	long	ECANA	developed	0.0103	0.02959	0.01218
HTI	Haiti	30	17	121	long	LAC	developing	0.0415	0.10405	0.09331
HUN	Hungary	18	18	112	short	ECANA	developed	0.0168	0.05677	0.02448
IDN	Indonesia	40	17	57	long	EAP	developing	0.0305	0.08543	0.03494
IND	India	47	18	8	long	SA	developing	0.027	0.07777	0.01954
IRL	Ireland	47	18	77	long	ECANA	developed	0.0058	0.01658	0.00319
IRN	Iran	43	18	36	long	MENA	developing	0.012	0.03503	0.0227
IRQ	Iraq	30	18	108	long	MENA	developing	0.0082	0.02301	0.01413
ISL	Iceland	29	16	39	long	ECANA	developed	0.0086	0.02371	0.01183
ISR	Israel	47	16	36	long	MENA	developed	0.0173	0.04853	0.02216
ITA	Italy	43	18	18	long	ECANA	developed	0.0062	0.01793	0.00581
JAM	Jamaica	34	11	233	short	LAC	developing	0.1169	0.29362	0.15091
JOR	Jordan	48	17	87	long	MENA	developing	0.0303	0.08387	0.02922
JPN	Japan	48	17	12	long	EAP	developed	0.0148	0.04183	0.02265
KAZ	Kazakhstan	10	23	0	long	ECANA	developing	0.021	0.0657	0.02997
KEN	Kenya	48	18	400	long	SSA	developing	0.0234	0.0787	0.02738
KGZ	Kyrgyzstan	21	18	58	long	ECANA	developing	0.0569	0.16359	0.18953
KOR	Korea	44	18	0	long	EAP	developed	0.0083	0.02396	0.00564
KWT	Kuwait	44	17	210	long	MENA	developed	0.0912	0.28442	0.14956
LBR	Liberia	3	18	0	long	SSA	developing	0.0192	0.0554	0.01436
LBY	Libya	17	16	32	long	MENA	developing	0.0141	0.03762	0.03062
LKA	Sri Lanka	41	18	244	long	SA	developing	0.0215	0.06201	0.02747
LSO	Lesotho	9	7	27	short	SSA	developing	0.0566	0.13708	0.04618
LTU	Lithuania	19	23	66	long	ECANA	developed	0.014	0.04312	0.01508
LUX	Luxembourg	25	13	52	long	ECANA	developed	0.0141	0.02996	0.0307
LVA	Latvia	19	21	101	long	ECANA	developed	0.0139	0.04154	0.04273
MAC	Macao	30	18	94	long	EAP	developed	0.0056	0.01396	0.01243
MAR	Morocco	35	17	64	long	MENA	developing	0.0325	0.09112	0.03225

country code	country	years	sectors	imputed	dynamic version	region	dev. status	Theil(normalized)	Theil	standard deviation Theil
MDA	Moldova	10	3	18	short	ECANA	developing	0.036	0.09187	0.05026
MDG	Madagascar	29	15	75	long	SSA	developing	0.012	0.04632	0.05341
MEX	Mexico	27	18	224	long	LAC	developing	0.0146	0.05173	0.03196
MKD	Macedonia	21	18	97	long	ECANA	developing	0.022	0.06341	0.03355
MLT	Malta	14	16	54	short	MENA	developed	0.0094	0.029	0.02796
MNG	Mongolia	19	18	102	long	EAP	developing	0.0312	0.08422	0.03902
MOZ	Mozambique	26	18	252	long	SSA	developing	0.0507	0.1965	0.19345
MUS	Mauritius	43	15	20	long	SSA	developing	0.0218	0.05858	0.0268
MWI	Malawi	40	11	58	long	SSA	developing	0.0403	0.09476	0.06529
MYS	Malaysia	11	18	46	short	EAP	developing	0.011	0.03418	0.00308
NGA	Nigeria	34	18	216	long	SSA	developing	0.0106	0.02885	0.01468
NIC	Nicaragua	21	18	0	long	LAC	developing	0.005	0.01453	0.00502
NLD	Netherlands	14	18	66	short	ECANA	developed	0.0087	0.02539	0.02359
NOR	Norway	46	18	78	long	ECANA	developed	0.0037	0.01067	0.006
NPL	Nepal	13	14	58	long	SA	developing	0.0254	0.06354	0.03242
NZL	New Zealand	47	18	295	long	EAP	developed	0.0165	0.0415	0.08545
OMN	Oman	18	22	22	long	MENA	developed	0.0357	0.10944	0.03435
PAK	Pakistan	44	18	432	long	SA	developing	0.0159	0.05563	0.02346
PAN	Panama	43	18	228	long	LAC	developing	0.0217	0.05888	0.02269
PER	Peru	28	18	177	long	LAC	developing	0.0676	0.20908	0.16057
PHL	Philippines	46	18	152	long	EAP	developing	0.0188	0.05496	0.01344
PNG	Papua New Guinea	25	16	0	long	EAP	developing	0.0291	0.08062	0.02301
POL	Poland	40	18	130	long	ECANA	developed	0.005	0.01541	0.01157
PRI	Puerto Rico	20	18	144	long	LAC	developed	0.0319	0.1185	0.08301
PRT	Portugal	20	18	142	long	ECANA	developed	0.0172	0.04886	0.01008
PRY	Paraguay	2	17	0	long	LAC	developing	0.0098	0.02766	0.00007
QAT	Qatar	25	13	189	long	MENA	developed	0.0134	0.38861	0.05925
ROU	Romania	33	17	124	long	ECANA	developing	0.1512	0.0538	0.07169
RUS	Russia	15	18	0	long	ECANA	developed	0.0191	0.04944	0.01314
SEN	Senegal	29	18	120	long	SSA	developing	0.0171	0.04101	0.02397
SGP	Singapore	20	18	40	short	EAP	developed	0.0158	0.03737	0.00509
SLV	El Salvador	36	18	216	long	LAC	developing	0.0124	0.0407	0.02331
SOM	Somalia	14	16	6	long	SSA	developing	0.0141	0.04401	0.02047
SRB	Serbia and Montenegro	12	18	90	long	ECANA	developing	0.016	0.12232	0.11691
SUR	Suriname	20	11	0	long	LAC	developing	0.0427	0.04876	0.02528
SVK	Slovakia	17	21	18	short	ECANA	developed	0.0203	0.02772	0.00936
SVN	Slovenia	24	17	66	long	ECANA	developed	0.0091	0.02459	0.00939
SWE	Sweden	20	18	38	short	ECANA	developed	0.009	0.00829	0.00549
SWZ	Swaziland	24	5	30	long	SSA	developing	0.0027	0.11148	0.04479
SYR	Syria	48	18	267	long	MENA	developing	0.069	0.12055	0.05836
THA	Thailand	39	18	588	long	EAP	developing	0.0458	0.05783	0.02888

country code	country	years	sectors	imputed	dynamic version	region	dev. status	Theil(normalized)	Theil	standard deviation Theil
TON	Tonga	30	18	175	long	EAP	developing	0.0225	0.06259	0.06513
TTO	Trinidad and Tobago	42	17	257	long	LAC	developed	0.0223	0.15252	0.08948
TUN	Tunisia	37	14	116	long	MENA	developing	0.0505	0.15268	0.13522
TUR	Turkey	47	18	42	long	ECANA	developing	0.0369	0.05052	0.03092
TWN	Taiwan	29	18	72	long	EAP	developed	0.0173	0.01477	0.00341
TZA	Tanzania	43	18	349	long	SSA	developing	0.0051	0.08027	0.03656
UGA	Uganda	23	10	26	long	SSA	developing	0.0279	0.10898	0.0796
UKR	Ukraine	19	18	6	long	ECANA	developing	0.0488	0.04651	0.01613
URY	Uruguay	41	18	216	long	LAC	developed	0.0161	0.04516	0.01753
USA	United States	45	18	54	long	ECANA	developed	0.0158	0.02505	0.00448
VEN	Venezuela	35	18	72	long	LAC	developed	0.0086	0.04673	0.01972
YEM	Yemen	9	17	18	long	MENA	developing	0.0156	0.08202	0.02211
ZAF	South Africa	48	18	223	long	SSA	developing	0.0289	0.05658	0.00942
ZMB	Zambia	22	17	146	long	SSA	developing	0.0046	0.0518	0.01617

Notes. The column "years" is the number of total (not necessarily consecutive) years covered for each country. "Sectors" refers to the number of ISIC 2-digit level sectors on which the Theil index is based. "Imputed" contains the total number of imputed data points across all sectors and years. It should be noted that this number tends rise with higher time coverage. "dynv" is short for "dynamic version" and contains the recommendation as to which version in the case of deviations between the two version of more than 10% in any year, with the exceptions discussed in appendix table A.1. "Region" refers to the geographic region and relies on the World Bank classification. SSH=Sub-Saharan Africa, SA=South Asia, LAC=Latin America and Caribbean, ECANA=Europe, Central Asia, and North America, MENA=Middle East and North Africa, and EAP=East Asia and Pacific. "Devstat" refers to the classification of countries as developed or developing and relies on the World Bank categorization, which is based on GNI. Theil(n) is the normalized version of the Theil index. The standard deviation in the last column is for the non-normalized version of the Theil.

Table A.3: Correlation with the UTIP index and extent of imputation

country code	# of imputed data points	# of years	# of years in UTIP	correlation with UTIP, levels	correlation with UTIP, differences	mean %-devations from UTIP, levels	mean %-devations from UTIP, differences
AFG	6	9	22	0.982	0.979	59.3	59.3
ALB	40	18	19	0.973	0.974	25.9	25.9
ARG	72	19	17	0.999	0.997	0.9	0.9
AUS	184	44	40	0.568	-0.393	21.1	21.1
AUT	80	20	44	0.673	0.584	12.5	12.5
AZE	49	21	17	0.973	0.749	11.3	11.3
BDI	166	23	17	1	1	0	0
BEL	130	47	42	0.136	-0.054	22.3	22.3
BEN	9	8	7	1	1	0	0
BFA	0	10	10	1	1	0	0
BGD	144	32	28	0.997	0.97	0.6	0.6
BGR	55	48	45	-0.169	-0.152	47.5	47.5
BLZ	24	4	2	1	n/a	0	0
BOL	22	32	32	0.882	0.366	24.4	24.4
BRA	636	15	17	0.879	0.919	13.1	13.1
BRB	0	28	28	0.984	0.988	3.4	3.4
BWA	73	30	27	0.048	0.234	62.5	62.5
CAF	64	8	19	0.984	0.997	28.9	28.9
CAN	24	48	45	0.972	0.906	2	2
CHE	158	11	5	0.872	0.954	3.5	3.5
CHL	65	46	44	0.994	0.988	5.7	5.7
CHN	294	34	16	0.997	0.61	2.4	2.4
CIV	9	22	22	1	0.999	0.3	0.3
CMR	205	33	28	1	0.998	1.4	1.4
COG	142	21	14	1	1	0	0
COL	78	48	43	0.997	0.996	0.6	0.6
CRI	364	41	22	1	1	0.1	0.1
CUB	42	15	13	1	1	0	0
CYP	30	48	46	0.948	0.825	19.8	19.8
CZE	38	21	20	0.991	0.831	10	10
DEU	72	27	30	-0.454	-0.025	344.6	344.6
DNK	152	47	42	0.998	0.998	3	3
DOM	0	23	23	1	1	0	0
DZA	48	14	27	0.997	0.998	2.7	2.7
ECU	2	46	45	0.997	0.997	1.3	1.3
EGY	256	47	39	1	0.997	0.2	0.2
ERI	9	19	42	0.734	0.966	26.3	26.3
ESP	6	47	45	0.992	0.994	2.7	2.7
EST	337	19	9	-0.027	-0.599	29.3	29.3
ETH	7	20	44	0.996	0.995	0.4	0.4
FIN	24	19	45	0.931	0.751	7.5	7.5
FJI	276	42	32	0.717	0.861	13.7	13.7
FRA	339	20	30	0.902	0.934	11.8	11.8
GAB	181	16	8	1	1	0	0
GBR	171	17	41	0.647	0.508	12.7	12.7
GEO	20	13	11	0.998	0.999	1.6	1.6
GHA	0	25	28	0.191	0.619	35.6	35.6
GMB	0	8	8	0.994	1	4.2	4.2

country code	# of imputed data points	# of years	# of years in UTIP	correlation with UTIP, levels	correlation with UTIP, differences	mean %-devations from UTIP, levels	mean %-devations from UTIP, differences
GRC	146	45	41	0.987	0.979	1.1	1.1
GTM	180	31	26	0.996	0.831	24.1	24.1
HKG	29	35	36	0.847	0.692	12.9	12.9
HND	233	34	26	0.178	0.1	41.1	41.1
HRV	12	25	23	0.996	0.984	3.6	3.6
HTI	121	30	21	0.288	0.423	23.8	23.8
HUN	112	18	43	0.954	0.694	8.2	8.2
IDN	57	40	36	0.903	0.922	3.8	3.8
IND	8	47	45	0.999	0.988	0.3	0.3
IRL	77	47	45	0.858	0.921	12.3	12.3
IRN	36	43	42	0.999	0.998	1.4	1.4
IRQ	108	30	27	1	1	0	0
ISL	39	29	20	0.985	0.941	3.4	3.4
ISR	36	47	44	0.996	0.944	2.1	2.1
ITA	18	43	40	0.997	0.989	1.7	1.7
JAM	233	34	34	-0.136	-0.032	57.5	57.5
JOR	87	48	42	0.743	0.471	8.1	8.1
JPN	12	48	45	0.999	0.998	1.8	1.8
KAZ	0	10	10	1	1	0	0
KEN	400	48	40	0.602	0.879	6.1	6.1
KGZ	58	21	13	0.856	0.961	24.2	24.2
KOR	0	44	44	0.995	0.989	2.1	2.1
KWT	210	44	35	1	1	0	0
LBR	0	3	n/a	n/a	n/a	n/a	n/a
LBY	32	17	17	0.984	0.94	6.7	6.7
LKA	244	41	26	0.999	1	0.5	0.5
LSO	27	9	14	0.973	0.964	4	4
LTU	66	19	16	0.984	0.954	1.5	1.5
LUX	52	25	45	0.105	0.288	20.8	20.8
LVA	101	19	16	0.686	0.143	22.8	22.8
MAC	94	30	26	0.971	0.952	13.7	13.7
MAR	64	35	33	0.99	0.961	4.7	4.7
MDA	18	10	17	0.556	-1	20.8	20.8
MDG	75	29	26	0.661	0.784	64.3	64.3
MEX	224	27	31	0.887	0.249	9.6	9.6
MKD	97	21	20	0.986	0.959	9.9	9.9
MLT	54	14	44	0.979	0.991	19.3	19.3
MNG	102	19	17	0.928	0.907	8.5	8.5
MOZ	252	26	13	0.966	1	2.1	2.1
MUS	20	43	40	0.992	0.987	9.6	9.6
MWI	58	40	35	0.971	0.958	15.1	15.1
MYS	46	11	39	0.994	0.996	2.5	2.5
NGA	216	34	28	1	1	0	0
NIC	0	21	21	1	1	0	0
NLD	66	14	43	0.473	-0.129	43.4	43.4
NOR	78	46	44	0.304	0.057	6.4	6.4
NPL	58	13	10	1	1	0.3	0.3
NZL	295	47	41	0.467	0.457	24.6	24.6
OMN	22	18	15	1	1	0	0
PAK	432	44	32	1	1	0.1	0.1

country code	# of imputed data points	# of years	# of years in UTIP	correlation with UTIP, levels	correlation with UTIP, differences	mean %-devations from UTIP, levels	mean %-devations from UTIP, differences
PAN	228	43	40	0.864	0.848	36	36
PER	177	28	21	1	0.996	1.6	1.6
PHL	152	46	41	0.964	0.997	2.4	2.4
PNG	0	25	27	0.997	0.998	1.7	1.7
POL	130	40	37	0.993	0.902	2.7	2.7
PRI	144	20	12	1	1	0	0
PRT	142	20	27	0.089	-0.686	7.3	7.3
PRY	0	2	3	-1		13.7	13.7
PSE	9	14	15	0.987	0.981	8.3	8.3
QAT	189	25	15	1	1	0.4	0.4
ROU	124	33	26	0.942	0.699	14.6	14.6
RUS	0	15	44	0.999	0.999	0.8	0.8
SEN	120	29	29	0.979	0.968	17.2	17.2
SGP	40	20	46	0.986	0.968	6.1	6.1
SLV	216	36	28	1	1	0	0
SOM	6	14	12	0.978	0.966	2	2
SRB	90	12	n/a	n/a	n/a	n/a	n/a
SUR	0	20	24	0.997	0.985	3.3	3.3
SVK	18	17	17	0.918	0.823	14.6	14.6
SVN	66	24	22	0.933	0.86	13.6	13.6
SWE	38	20	38	0.903	0.67	21.1	21.1
SWZ	30	24	26	0.985	0.968	4.9	4.9
SYR	267	48	28	0.94	0.719	43.8	43.8
THA	588	39	23	0.885	0.622	7	7
TON	175	30	23	0.998	0.996	2.5	2.5
TTO	257	42	26	0.997	0.992	0.7	0.7
TUN	116	37	29	1	0.999	6.2	6.2
TUR	42	47	43	1	0.998	0.5	0.5
TWN	72	29	25	1	1	0	0
TZA	349	43	34	0.875	0.878	9.1	9.1
UGA	26	23	21	-0.016	-0.01	56.1	56.1
UKR	6	19	19	0.992	0.996	4.1	4.1
URY	216	41	32	0.985	0.859	2.9	2.9
USA	54	45	42	0.714	0.424	3	3
VEN	72	35	34	0.938	0.837	38.1	38.1
YEM	18	9	10	0.051	0.431	33.5	33.5
YUG	0	27	35	1	1	0	0
ZAF	223	48	41	0.969	0.953	1.8	1.8
ZMB	146	22	18	0.983	0.977	7.4	7.4

Table A.4: Imputed values vs. dropping of sectors: RE and FE results

	(1)		(2)		(3)		(4)	
	RE	se	FE	se, r	RE	se	RE	se, r
Imputations	3.030***	(0.466)	3.030**	(1.225)	3.111***	(0.460)	3.111***	(1.154)
Dropped sectors	-5.282***	(0.815)	-5.282	(3.444)	-2.396***	(0.607)	-2.396	(1.817)
1964	1.090	(13.44)	1.090	(1.367)				
1965	0.915	(13.25)	0.915	(1.908)				
1966	-0.912	(13.15)	-0.912	(2.741)				
1967	2.733	(12.86)	2.733	(3.271)				
1968	-0.793	(12.36)	-0.793	(3.941)				
1969	-1.377	(12.61)	-1.377	(4.602)				
1970	-0.557	(12.48)	-0.557	(4.956)				
1971	-2.374	(12.34)	-2.374	(5.034)				
1972	0.193	(12.41)	0.193	(5.308)				
1973	3.219	(12.34)	3.219	(6.548)				
1974	3.266	(12.22)	3.266	(6.888)				
1975	2.218	(12.20)	2.218	(7.074)				
1976	3.129	(12.21)	3.129	(7.159)				
1977	0.184	(12.13)	0.184	(7.882)				
1978	4.355	(12.13)	4.355	(7.570)				
1979	-0.618	(12.13)	-0.618	(6.960)				
1980	4.184	(12.07)	4.184	(8.647)				
1981	2.055	(12.07)	2.055	(9.078)				
1982	0.204	(12.10)	0.204	(9.514)				
1983	-4.433	(12.10)	-4.433	(9.099)				
1984	-5.342	(12.10)	-5.342	(10.06)				
1985	-41.32***	(12.09)	-41.32	(37.52)				
1986	-6.042	(12.08)	-6.042	(13.27)				
1987	-7.133	(12.05)	-7.133	(13.97)				
1988	-13.52	(12.20)	-13.52	(16.13)				
1989	-8.212	(12.17)	-8.212	(14.38)				
1990	1.174	(12.05)	1.174	(7.983)				
1991	2.955	(11.99)	2.955	(8.181)				
1992	2.617	(12.13)	2.617	(8.634)				
1993	5.108	(12.10)	5.108	(9.354)				
1994	-4.776	(11.97)	-4.776	(9.583)				
1995	-6.343	(12.20)	-6.343	(11.63)				
1996	11.93	(12.19)	11.93	(13.04)				
1997	12.82	(12.26)	12.82	(14.22)				
1998	18.33	(12.29)	18.33	(15.03)				
1999	17.17	(12.58)	17.17	(15.78)				
2000	19.42	(12.39)	19.42	(15.90)				
2001	18.17	(12.49)	18.17	(15.88)				
2002	22.11*	(12.69)	22.11	(16.88)				
2003	17.89	(12.79)	17.89	(17.33)				
2004	22.07*	(12.80)	22.07	(16.96)				
2005	18.88	(12.84)	18.88	(17.25)				
2006	21.41*	(12.95)	21.41	(17.46)				
2007	23.76*	(13.45)	23.76	(18.16)				
2008	5.556	(17.10)	5.556	(25.21)				
Constant	0.0131	(10.30)	0.0131	(5.216)	0.105	(4.026)	0.105	(4.081)
Observations	3,627		3,627		3,627		3,627	
# of countries	135		135		135		135	
R^2 (within)	0.036		0.036		0.016		0.016	

Notes. Standard errors in parentheses as indicated in the top column; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the percentage deviation between the dynamic version of the newly constructed Theil index and the UTIP index. FE refers to fixed effects estimation, RE refers to random effects estimation, se refers to the standard error, and r indicates that standard errors are robust.

Table A.5: Contribution of the 2-digit sectors to the within-component of inequality

	(1)		(2)	(3)		(4)	
	Within-sectoral inequality		Weight (=sectoral wage share)	Weighted within-sectoral inequality		Subcategories per sector (average)	
sector	mean	rank	mean	mean	rank	3 digit	4 digit
15	0.0575	1	0.209	0.01477	1	4.8	12.9
16	0	22	0.0218	0	22	1	1
17	0.0182	11	0.0934	0.00094	11	2.8	5.6
18	0.0019	21	0.0739	0.00008	20	1.6	1.6
19	0.0165	15	0.0195	0.0003	18	1.9	2.5
20	0.0175	13	0.0355	0.00065	13	1.9	4.1
21	0.0115	19	0.0294	0.00037	17	1	2.7
22	0.0232	5	0.0451	0.0012	7	2.5	5.3
23	0.0136	17	0.0315	0.00057	15	1.5	1.5
24	0.0254	4	0.0743	0.00284	3	2.5	7
25	0.0138	16	0.0376	0.0006	14	1.9	2.5
26	0.0486	2	0.0615	0.00474	2	1.9	6.1
27	0.0178	12	0.057	0.00107	10	2.6	3
28	0.0188	10	0.0611	0.00114	9	1.9	6
29	0.0224	6	0.0605	0.00122	6	2.9	11.9
30	0	23	0.0148	0	23	1	1
31	0.0193	9	0.0468	0.00056	16	5.2	5.2
32	0.0202	8	0.039	0.00066	12	2.7	2.7
33	0.0128	18	0.0138	0.00022	19	2.4	3.9
34	0.0217	7	0.0544	0.00117	8	2.5	2.5
35	0.027	3	0.0333	0.00129	5	3.3	4.8
36	0.0166	14	0.0367	0.00176	4	1.9	4.3
37	0.0105	20	0.0026	0.00004	21	1.7	1.7

Notes. Columns (1) and (2) contain the unweighted and weighted within-components and the ranking of every 2-digit sector for each of these categories. Column (2) contains the weight and links the numbers in columns (1) and (2). Columns (4) display the average number of sectors covered by the data at the 3- and 4-digit level.

Table A.6: Sectoral composition vs. sectoral coverage: FE and RE results

	(1)		(2)	
	Fixed effects	Standard error	Random effects	Standard error
Subsectors	0.0261***	(0.009)	0.0245***	(0.006)
Share_15	0.799	(0.970)	0.219	(0.692)
Share_17	-0.0140	(0.950)	-0.209	(0.633)
Share_18	-0.616	(0.951)	-0.986	(0.653)
Share_19	-0.654	(1.306)	-0.520	(1.032)
Share_20	2.989**	(1.416)	1.057	(0.865)
Share_21	0.370	(1.401)	-1.124	(0.843)
Share_22	-0.452	(0.967)	-0.257	(0.790)
Share_23	-1.629	(1.508)	-1.089	(1.061)
Share_24	-0.542	(0.976)	-0.725	(0.695)
Share_25	0.604	(1.037)	0.640	(0.911)
Share_26	1.135	(1.040)	0.735	(0.649)
Share_27	-1.132	(0.817)	-1.301**	(0.636)
Share_28	0.658	(0.866)	-0.00429	(0.736)
Share_29	-0.181	(0.900)	-0.476	(0.754)
Share_31	-0.489	(1.231)	-0.761	(0.904)
Share_32	-0.578	(0.902)	-0.788	(0.764)
Share_33	-0.108	(1.516)	1.158	(1.376)
Share_34	-0.446	(1.028)	-1.049*	(0.618)
Share_35	-0.567	(0.877)	-0.451	(0.657)
Share_36	-2.478*	(1.297)	-2.267*	(1.232)
Share_37	2.189	(2.656)	-0.343	(1.824)
Constant	26.44	(85.10)	57.99	(62.20)
Year FE	YES		YES	
Observations	429		429	
R^2	0.465			
# of countries	53		53	

Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the share of within-sectoral inequality in percent. Numbers 15 to 37 refer to the 2-digit sector's wage share in total manufacturing wages and are also in percent. High variation samples 1 and to refer to subsamples of countries with above-average variation in sectoral coverage (1), and countries with one standard deviation above the average variation (2). Note that a Hausman test clearly rejects the random effects model at the <1% significance level.

Table A.7: Relationship between Theil and income inequality: FE results, extended model

	(1) wiid	(2) swiid	(3) lis	(4) silc	(5) wb
ln(Theil)	0.0104	0.00593	-0.0227	0.0243	0.00209
	(0.0125)	(0.0122)	(0.0320)	(0.0185)	(0.0168)
GDP per capita	-5.13e-06	2.79e-06	-9.38e-06*	-1.13e-06	-1.50e-07
	(3.25e-06)	(1.89e-06)	(4.93e-06)	(2.82e-06)	(5.56e-06)
Population growth	-0.00443	0.00658	0.0191	-0.0175	-0.00469
	(0.0150)	(0.00974)	(0.0316)	(0.0129)	(0.0169)
Share urban	-0.0106***	0.00147	0.0212**	-0.00305	0.00578**
	(0.00396)	(0.00286)	(0.0101)	(0.00483)	(0.00248)
Manufacturing value-added	-0.00694*	-0.00317*	-0.00742	-0.00176	0.000603
	(0.00371)	(0.00185)	(0.00741)	(0.00583)	(0.00288)
Trade openness	0.000250	0.00244*	0.00361	0.00306	0.00486***
	(0.00188)	(0.00135)	(0.00283)	(0.00229)	(0.00180)
Price level inv.	-0.0362	0.0985**	-0.278	0.168	0.0531
	(0.0670)	(0.0427)	(0.166)	(0.116)	(0.0954)
Tfp	0.293	0.141	0.728**	0.484*	0.516***
	(0.180)	(0.0883)	(0.348)	(0.244)	(0.0997)
Open.*price level inv.	0.00148***	-0.000834*	0.00193*	-0.000582	0.000785
	(0.000555)	(0.000463)	(0.00104)	(0.000641)	(0.000882)
Openness*tfp	-0.000524	-0.00179	-0.00432**	-0.00199	-0.00507***
	(0.00161)	(0.00114)	(0.00201)	(0.00202)	(0.00160)
# imputed	-0.00201*	0.000727	-0.00487	0.00127	-0.00176
	(0.00106)	(0.00126)	(0.00352)	(0.00152)	(0.00183)
1964		0.0392			
		(0.0328)			
1965		0.0234			
		(0.0380)			
1966		-0.00793			
		(0.0199)			
1967		0.0239			
		(0.0362)			
1968		-0.00778			
		(0.0181)			
1969		0.00744			
		(0.0294)			
1970	0.457***	0.0215			
	(0.108)	(0.0279)			
1971		0.00176			
		(0.0299)			
1972	0.366***	-0.0152			
	(0.118)	(0.0262)			
1973	0.598***	-0.00603			
	(0.105)	(0.0252)			
1974		0.00277			
		(0.0248)			
1975	-0.0189	-0.00384			
	(0.0363)	(0.0261)			
1976	0.467***	-0.00543			
	(0.118)	(0.0232)			
1977		0.00246			
		(0.0274)			
1978		0.00120			
		(0.0347)			
1979	0.257**	0.00408			
	(0.104)	(0.0366)			
1980	0.327**	-0.00799			
	(0.126)	(0.0440)			
1981		-0.0187			
		(0.0388)			
1982	0.277**	-0.0307			
	(0.111)	(0.0377)			
1983	0.0639	-0.0234			-0.0411
	(0.142)	(0.0331)			(0.0585)
1984	0.335***	-0.0390	-0.00308		-0.105*
	(0.120)	(0.0323)	(0.0609)		(0.0531)

1985	0.226	-0.0401			-0.115*
	(0.154)	(0.0317)			(0.0618)
1986	0.259***	-0.0458	0.00446		-0.116
	(0.0813)	(0.0307)	(0.0301)		(0.0892)
1987	0.323***	-0.0558			-0.0467
	(0.103)	(0.0340)			(0.0453)
1988	0.357***	-0.0560			-0.104
	(0.0951)	(0.0358)			(0.0652)
1989	0.341***	-0.0614	0.0915*		-0.0618
	(0.0848)	(0.0384)	(0.0459)		(0.0633)
1990	0.363***	-0.0684			-0.106
	(0.0875)	(0.0433)			(0.0661)
1991	0.257**	-0.0497	0.00565		-0.0706
	(0.0995)	(0.0437)	(0.0495)		(0.0680)
1992	0.350***	-0.0467	0.106**		-0.115
	(0.0986)	(0.0451)	(0.0467)		(0.0698)
1993	0.308***	-0.0368	0.104**		-0.0801
	(0.0976)	(0.0466)	(0.0500)		(0.0754)
1994	0.372***	-0.0370	0.0915*		-0.125**
	(0.0932)	(0.0475)	(0.0514)		(0.0625)
1995	0.392***	-0.0304	0.0888		-0.0887
	(0.0898)	(0.0486)	(0.0605)		(0.0645)
1996	0.376***	-0.0282	0.0879	-0.0194	-0.0936
	(0.0916)	(0.0495)	(0.0590)	(0.0184)	(0.0638)
1997	0.392***	-0.0294	0.0871	-0.0510**	-0.103
	(0.0901)	(0.0509)	(0.0633)	(0.0218)	(0.0654)
1998	0.394***	-0.0331	0.161**	-0.0405	-0.0992
	(0.0884)	(0.0515)	(0.0665)	(0.0296)	(0.0652)
1999	0.398***	-0.0350	0.110	-0.0412	-0.0828
	(0.0888)	(0.0508)	(0.0708)	(0.0244)	(0.0667)
2000	0.414***	-0.0318	0.124	-0.0272	-0.0929
	(0.0899)	(0.0523)	(0.0824)	(0.0370)	(0.0677)
2001	0.396***	-0.0335	-0.00168	-0.0287	-0.0825
	(0.0877)	(0.0529)	(0.104)	(0.0392)	(0.0693)
2002	0.419***	-0.0369	0.0907	-0.0369	-0.0792
	(0.0946)	(0.0536)	(0.0628)	(0.0419)	(0.0721)
2003	0.378***	-0.0402	0.0853	-0.0173	-0.0795
	(0.0932)	(0.0546)	(0.0999)	(0.0445)	(0.0691)
2004	0.397***	-0.0420	0.118	-0.0487	-0.103
	(0.0970)	(0.0550)	(0.0811)	(0.0461)	(0.0685)
2005	0.397***	-0.0432	0.116	-0.0291	-0.113
	(0.100)	(0.0552)	(0.0985)	(0.0469)	(0.0693)
2006	0.385***	-0.0497	-0.113	-0.0358	-0.123*
	(0.105)	(0.0564)	(0.154)	(0.0503)	(0.0739)
2007	0.378***	-0.0437	0.0879	-0.0678	-0.131*
	(0.105)	(0.0584)	(0.0979)	(0.0567)	(0.0756)
2008	0.384***	-0.0615	0.251*	-0.0773	-0.159**
	(0.102)	(0.0587)	(0.141)	(0.0567)	(0.0770)
2009	0.427***	-0.0591	-0.0159	-0.0458	-0.139*
	(0.0935)	(0.0583)	(0.135)	(0.0577)	(0.0732)
2010	0.379***	-0.0326	0.141	-0.0543	-0.189**
	(0.0994)	(0.0603)	(0.105)	(0.0510)	(0.0797)
Constant	4.051***	3.373***	-3.124***	3.141***	2.818***
	(0.369)	(0.160)	(0.890)	(0.531)	(0.263)
WIID dummmies	YES	-	-	-	
Observations	619	1,521	120	256	483
R-squared	0.810	0.142	0.577	0.259	0.223
# of countries	66	82	35	28	73

Notes. Standard errors in parenthese; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logged Gini coefficient from the data source indicated in the top row. Swiid refers to net inequality from the SWIID database. Silc denotes the EU SILC data and wb the WDI Gini coefficients. The dummies for the underlying WIID categories are included in column 1 but are not shown to save space (available upon request).

Table A.8: Relationship between Theil and income inequality: FE results, extended model

	(1) wiid	(2) swiid	(3) lis	(4) silc	(5) wb
ln(Theil)	0.0532***	0.0154	0.103***	0.0309	0.0275
	(0.0156)	(0.0135)	(0.0291)	(0.0193)	(0.0181)
Population growth	0.0343**	0.0188***	0.0173	-0.0280***	0.0307**
	(0.0163)	(0.00611)	(0.0308)	(0.0108)	(0.0125)
Share urban	-0.00408***	0.000202	-0.00295	-0.00432*	0.000873
	(0.00145)	(0.00168)	(0.00246)	(0.00224)	(0.00124)
Manufacturing value-added	-0.00588*	-0.00166	-0.000276	-0.00170	0.000755
	(0.00338)	(0.00156)	(0.00443)	(0.00438)	(0.00237)
# imputed	-0.00230	0.000585	-0.000266	-0.000831	-0.00379
	(0.00149)	(0.00177)	(0.00279)	(0.00143)	(0.00336)
# of ISIC	0.00820	-0.00102	0.00658	-0.00593	-0.00240
	(0.00567)	(0.00405)	(0.0133)	(0.00486)	(0.00651)
1964		0.0213			
		(0.0226)			
1965		-0.0110			
		(0.0252)			
1966		-0.0315*			
		(0.0188)			
1967		-0.0217			
		(0.0298)			
1968		0.00628			
		(0.0338)			
1969		-0.0118			
		(0.0288)			
1970	0.562***	0.00199			
	(0.0792)	(0.0296)			
1971		-0.0220			
		(0.0274)			
1972	0.589***	-0.0272			
	(0.0624)	(0.0310)			
1973	0.707***	-0.0267			
	(0.0802)	(0.0248)			
1974		-0.0110			
		(0.0279)			
1975	0.0149	-0.0187			
	(0.0119)	(0.0250)			
1976	0.633***	-0.0150			
	(0.0631)	(0.0223)			
1977		0.00457			
		(0.0228)			
1978		2.44e-06			
		(0.0345)			
1979	0.355***	0.00629			
	(0.0739)	(0.0333)			
1980	0.469***	0.00258			
	(0.150)	(0.0384)			
1981		-0.0178			
		(0.0308)			
1982	0.395***	-0.0269			
	(0.0880)	(0.0303)			
1983	0.489***	-0.00937			-0.0985
	(0.0604)	(0.0289)			(0.156)
1984	0.497***	-0.0289	0.106***		-0.0745
	(0.0926)	(0.0266)	(0.0395)		(0.120)
1985	0.381***	-0.0235			-0.0975
	(0.107)	(0.0271)			(0.157)
1986	0.372***	-0.0247	0.0560***		-0.0731
	(0.0544)	(0.0278)	(0.0191)		(0.169)
1987	0.449***	-0.0295			-0.0185
	(0.0675)	(0.0303)			(0.127)
1988	0.458***	-0.0248			-0.0207
	(0.0835)	(0.0306)			(0.159)
1989	0.438***	-0.0300	0.0672*		0.0226
	(0.0702)	(0.0319)	(0.0389)		(0.157)

1990	0.461***	-0.0398			-0.0356
	(0.0715)	(0.0332)			(0.148)
1991	0.349***	-0.0179	-0.0432		-0.00455
	(0.0839)	(0.0333)	(0.0521)		(0.160)
1992	0.449***	-0.0111	0.118**		-0.0213
	(0.0756)	(0.0348)	(0.0508)		(0.160)
1993	0.382***	7.52e-05	0.117***		0.0151
	(0.0694)	(0.0352)	(0.0434)		(0.165)
1994	0.472***	0.00284	0.126**		-0.0306
	(0.0628)	(0.0354)	(0.0516)		(0.155)
1995	0.482***	0.00951	0.0834***		0.0697
	(0.0637)	(0.0350)	(0.0267)		(0.160)
1996	0.462***	0.00858	0.0915**	-0.0154	0.00432
	(0.0649)	(0.0358)	(0.0458)	(0.0179)	(0.155)
1997	0.468***	0.00633	0.0237	-0.0489***	0.0290
	(0.0670)	(0.0371)	(0.0350)	(0.0186)	(0.156)
1998	0.473***	0.00422	0.143***	-0.0322	0.00612
	(0.0588)	(0.0381)	(0.0452)	(0.0248)	(0.158)
1999	0.464***	-0.000158	0.0176	-0.0329**	0.00600
	(0.0643)	(0.0378)	(0.0518)	(0.0156)	(0.159)
2000	0.482***	0.00403	0.0669*	-0.0288	0.0174
	(0.0644)	(0.0380)	(0.0397)	(0.0252)	(0.154)
2001	0.459***	-0.00196	-0.110	-0.0318	0.0472
	(0.0627)	(0.0385)	(0.0910)	(0.0256)	(0.155)
2002	0.483***	0.00144	0.106	-0.0376	0.0243
	(0.0696)	(0.0389)	(0.0757)	(0.0298)	(0.158)
2003	0.451***	0.000926	0.0259	0.0115	0.0263
	(0.0655)	(0.0392)	(0.0543)	(0.0279)	(0.154)
2004	0.490***	0.000619	0.0874**	-0.00795	0.00797
	(0.0708)	(0.0396)	(0.0436)	(0.0293)	(0.155)
2005	0.484***	0.00266	0.0149	0.0109	0.00584
	(0.0702)	(0.0391)	(0.0527)	(0.0278)	(0.157)
2006	0.473***	0.000768	-0.123	0.0138	0.0106
	(0.0736)	(0.0390)	(0.145)	(0.0284)	(0.156)
2007	0.478***	0.0135	0.0547	-0.00381	0.0142
	(0.0711)	(0.0400)	(0.0489)	(0.0313)	(0.155)
2008	0.491***	-0.00554	0.159	-0.0136	0.00767
	(0.0704)	(0.0404)	(0.145)	(0.0303)	(0.158)
2009	0.519***	-0.00485	0.00977	-0.0200	0.00609
	(0.0732)	(0.0426)	(0.0710)	(0.0388)	(0.156)
2010	0.499***	0.0354	0.107*	-0.0187	-0.0127
	(0.0781)	(0.0423)	(0.0626)	(0.0351)	(0.163)
Constant	3.717***	3.665***	-0.635*	4.003***	3.685***
	(0.169)	(0.0965)	(0.337)	(0.204)	(0.204)
WIID dummmies	YES	-	-	-	-
Observations	632	1,765	121	256	542
R^2	0.397	0.192	0.478	0.208	0.279
# of countries	71	100	36	28	88

Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logged Gini coefficient from the data source indicated in the top row. Swiid refers to net inequality from the SWIID database. Silc denotes the EU SILC data and wb the WDI Gini coefficients. The dummies for the underlying WIID categories are included in column 1 but are not shown to save space (available upon request).

Table A.9: Relationship between Theil and income inequality: FE results, extended model

	(1) wiid	(2) swiid	(3) lis	(4) silc	(5) wb
GDP per capita	0.529***	0.277*	0.200	-0.0238	0.634***
	(0.138)	(0.165)	(0.280)	(0.298)	(0.205)
Population growth	0.119**	0.0800*	0.0446	-0.00800	0.138***
	(0.0557)	(0.0441)	(0.126)	(0.115)	(0.0523)
Share urban	-0.00624	0.00489	0.0108	-0.00104	-0.000284
	(0.00507)	(0.00613)	(0.00944)	(0.00752)	(0.00467)
Manuf. value-added	0.00365	0.0146**	0.0372**	0.00855	0.0138
	(0.0101)	(0.00670)	(0.0177)	(0.0166)	(0.0105)
Trade openness	-0.0114	-0.165	-0.331*	-0.374**	-0.305**
	(0.123)	(0.173)	(0.176)	(0.150)	(0.152)
Price level of investment	0.375	0.0891	-1.111	4.026***	1.402*
	(0.764)	(0.613)	(2.015)	(1.428)	(0.850)
Tfp	-7.185**	-3.352**	0.0805	-16.07**	-5.765***
	(3.289)	(1.538)	(4.635)	(8.164)	(1.659)
Open.*price lev. inv.	-0.121	0.0450	0.447	-0.568*	-0.379*
	(0.192)	(0.143)	(0.507)	(0.303)	(0.210)
Open.*tfp	1.651**	0.738*	-0.0919	3.621*	1.370***
	(0.756)	(0.404)	(1.143)	(1.885)	(0.451)
# imputed	-0.0224	-0.0547***	0.00959	-0.0171	-0.0471**
	(0.0143)	(0.0166)	(0.0268)	(0.0129)	(0.0208)
# of ISIC	0.0819***	0.0292	0.0164	-0.00729	0.0465
	(0.0280)	(0.0277)	(0.0335)	(0.0289)	(0.0297)
ECA	0.500*	0.282	1.513***		0.281
	(0.282)	(0.202)	(0.451)		(0.271)
LAC	0.351	-0.206	0.622		0.0284
	(0.306)	(0.224)	(0.594)		(0.279)
MENA	0.0180	-0.0559		-0.143	-0.0211
	(0.270)	(0.269)		(0.342)	(0.438)
NA	0.269	0.0488	1.021**		-0.157
	(0.398)	(0.382)	(0.486)		(0.420)
SA	0.175	0.100	1.580*		0.130
	(0.339)	(0.371)	(0.867)		(0.336)
SSH	0.569	0.161	1.941***		0.477
	(0.431)	(0.237)	(0.625)		(0.371)
1964		-0.165			
		(0.163)			
1965		-0.0962			
		(0.226)			
1966		-0.0670			
		(0.206)			
1967		-0.0756			
		(0.234)			
1968		-0.0877			
		(0.216)			
1969		-0.337			
		(0.259)			
1970	1.527***	-0.0261			
	(0.469)	(0.289)			
1971		-0.214			
		(0.280)			
1972	0.998***	-0.257			
	(0.226)	(0.274)			
1973	1.747***	-0.146			
	(0.470)	(0.330)			
1974		-0.260			
		(0.355)			
1975	0.193	-0.251			
	(0.210)	(0.368)			
1976	1.415***	-0.385			
	(0.443)	(0.367)			
1977		-0.296			
		(0.401)			
1978		-0.140			
		(0.372)			
1979	1.088**	-0.232			
	(0.520)	(0.396)			

1980	1.458***	-0.252			
	(0.449)	(0.418)			
1981		-0.207			
		(0.393)			
1982	1.534***	-0.300			
	(0.466)	(0.411)			
1983	1.518**	-0.268			-1.497***
	(0.724)	(0.402)			(0.369)
1984	1.844***	-0.261	0.550		-0.536
	(0.463)	(0.403)	(0.562)		(0.338)
1985	1.698***	-0.217			-1.102**
	(0.435)	(0.388)			(0.436)
1986	1.330***	-0.355	0.183**		-1.151**
	(0.441)	(0.397)	(0.0751)		(0.538)
1987	1.379***	-0.545			-1.197***
	(0.423)	(0.422)			(0.386)
1988	1.185**	-0.535			-1.494***
	(0.467)	(0.429)			(0.408)
1989	1.190***	-0.660	-0.154		-1.273***
	(0.452)	(0.429)	(0.430)		(0.434)
1990	1.409***	-0.708			-1.356***
	(0.458)	(0.456)			(0.521)
1991	1.031**	-0.708	-0.907***		-1.509***
	(0.496)	(0.457)	(0.250)		(0.415)
1992	1.282***	-0.632	0.136		-1.319***
	(0.461)	(0.449)	(0.221)		(0.391)
1993	1.131***	-0.668	-0.621***		-1.215***
	(0.423)	(0.446)	(0.138)		(0.464)
1994	1.217***	-0.664	-0.392		-1.404***
	(0.468)	(0.445)	(0.281)		(0.411)
1995	1.282***	-0.650	-0.565***		-1.266***
	(0.466)	(0.449)	(0.151)		(0.381)
1996	1.131**	-0.722	-0.194	-0.0884	-1.409***
	(0.494)	(0.452)	(0.186)	(0.0591)	(0.390)
1997	1.096**	-0.762*	-1.029***	0.0271	-1.290***
	(0.487)	(0.450)	(0.156)	(0.106)	(0.410)
1998	1.043**	-0.817*	-0.405	-0.101	-1.249***
	(0.488)	(0.442)	(0.276)	(0.0975)	(0.409)
1999	0.954**	-0.857**	-1.016***	-0.00742	-1.496***
	(0.481)	(0.432)	(0.157)	(0.135)	(0.425)
2000	0.954*	-0.784*	-0.732***	0.132	-1.455***
	(0.490)	(0.429)	(0.212)	(0.149)	(0.410)
2001	0.964*	-0.810*	-2.009***	0.212	-1.608***
	(0.495)	(0.416)	(0.477)	(0.162)	(0.441)
2002	0.911*	-0.859**	-0.212	-0.346	-1.557***
	(0.473)	(0.425)	(0.253)	(0.248)	(0.420)
2003	0.972**	-0.901**	-0.846**	-0.270*	-1.512***
	(0.485)	(0.435)	(0.416)	(0.151)	(0.425)
2004	1.001**	-0.911**	-0.917***	-0.332**	-1.562***
	(0.508)	(0.453)	(0.246)	(0.155)	(0.428)
2005	0.984*	-0.920**	-1.355***	-0.269*	-1.563***
	(0.512)	(0.455)	(0.394)	(0.159)	(0.417)
2006	0.845	-0.942**	-1.616***	-0.400**	-1.589***
	(0.515)	(0.462)	(0.438)	(0.167)	(0.427)
2007	0.946*	-1.021**	-1.133***	-0.476***	-1.581***
	(0.525)	(0.468)	(0.314)	(0.183)	(0.422)
2008	0.912*	-0.949**	-1.169***	-0.767***	-1.455***
	(0.549)	(0.481)	(0.290)	(0.233)	(0.411)
2009	0.976*	-0.929*	-1.556***	-0.668***	-1.450***
	(0.540)	(0.480)	(0.341)	(0.177)	(0.417)
2010	1.026*	-0.748	-0.590	-0.204	-1.386***
	(0.572)	(0.482)	(0.419)	(0.229)	(0.426)
Constant	1.397	5.693***	0.855	10.76***	3.408
	(1.257)	(1.568)	(2.712)	(2.943)	(2.078)
Observations	619	1,521	120	256	483
WIID dummies	YES	-	-	-	-
R^2	0.509	0.371	0.705	0.454	0.375
# of countries	66	82	35	28	73

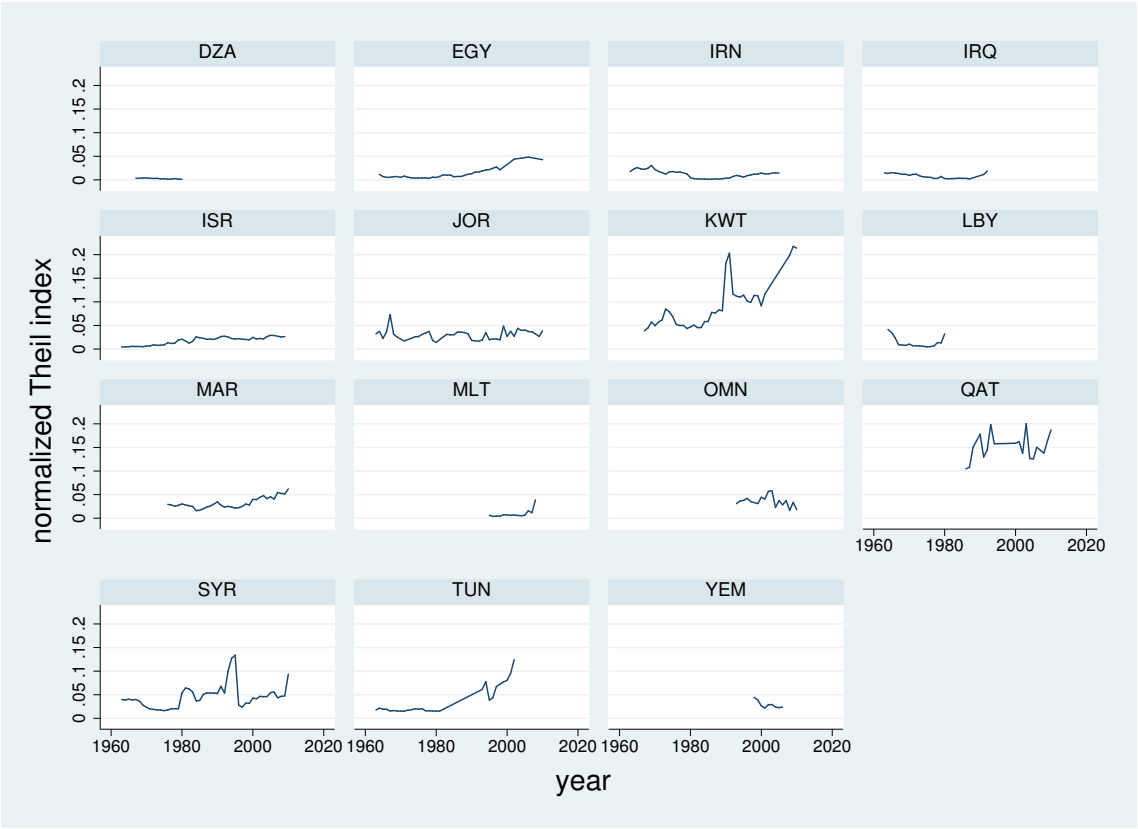
Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. For more information, see table A.8 notes

Table A.10: Relationship between Theil and Income Inequality: FE results, using the UTIP index

	(1) wiid	(2) swiid	(3) lis	(4) silc	(5) wb
ln(Theil_utip)	0.0903***	0.0116	0.0731**	0.0335	0.0218
	(0.0195)	(0.0178)	(0.0273)	(0.0254)	(0.0171)
Pop. growth	-0.0179	0.0163***	-0.0158	-0.0266	0.00745
	(0.0206)	(0.00581)	(0.0255)	(0.0181)	(0.0173)
Share urban	-0.0135***	0.00168	0.00304	-0.00546	0.00650*
	(0.00433)	(0.00261)	(0.00425)	(0.00606)	(0.00361)
Manuf. v.add.	-0.000860	-0.00295*	0.00489	0.00390	-0.00442
	(0.00311)	(0.00156)	(0.00465)	(0.00767)	(0.00307)
Constant	4.724***	3.589***	-1.154***	3.853***	3.459***
	(0.289)	(0.0980)	(0.298)	(0.521)	(0.218)
WIID dummies	YES	-	-	-	-
Year FE	YES	YES	YES	YES	YES
Observations	598	1,827	120	205	486
R^2	0.790	0.067	0.748	0.186	0.106
# of countries	72	110	33	27	91

Notes. Standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.1. The dependent variable is the logged Gini coefficient from the data source indicated in the top row. Swiid refers to net inequality from the SWIID database. Silc denotes the EU SILC data and wb the WDI Gini coefficients. The dummies for the underlying WIID categories are included in column 2 but are not shown to save space (available upon request).

Figure A.1: Development of the Theil index in the countries of the MENA region



Notes. The years 1982-1988 in Tunisia rely on linearly imputed values. This means that the increase in inequality from the low level in the early years until 1981 to the peak in 1989 can, theoretically, occur less continuously - and not necessarily in a monotonous manner in any of the imputed years. Within Tunisia, the spike in 1989 is attributable to huge increases in the wage bills in several sectors, most notably, 15 and 18.

Figure A.2: Size of the within-component and sectoral coverage by country

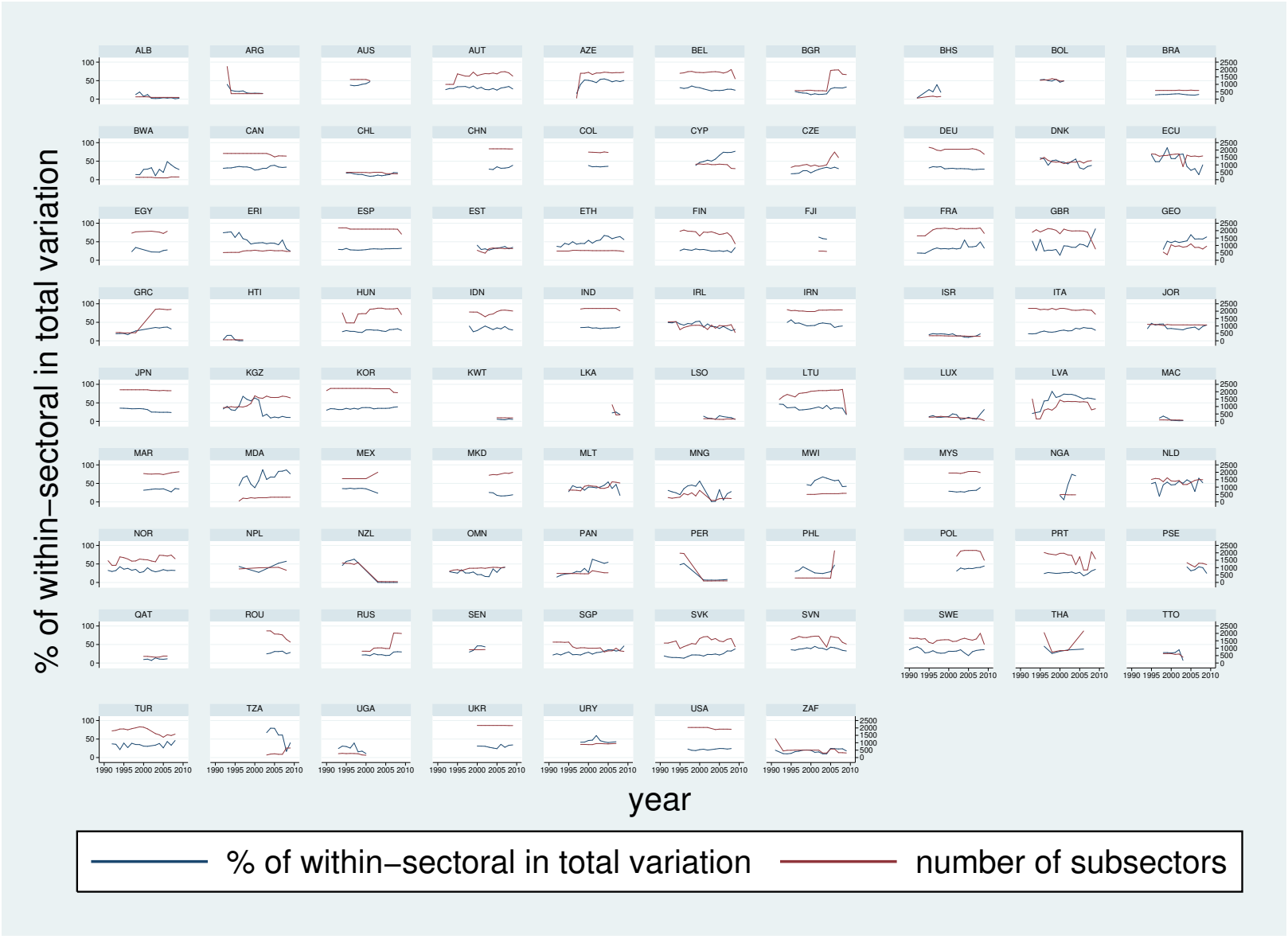


Figure A.3: Log-normality of the (normalized) Theil index

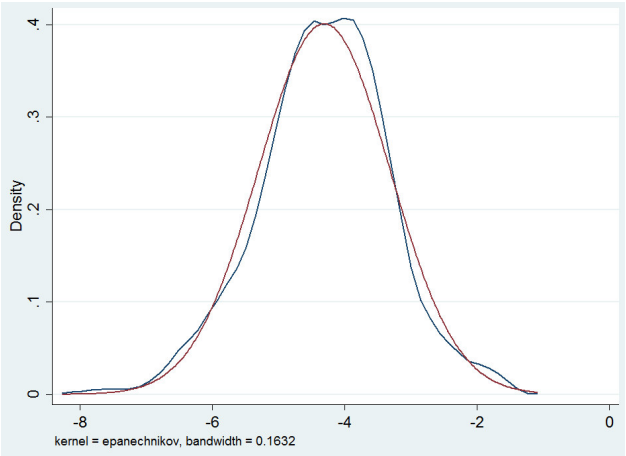
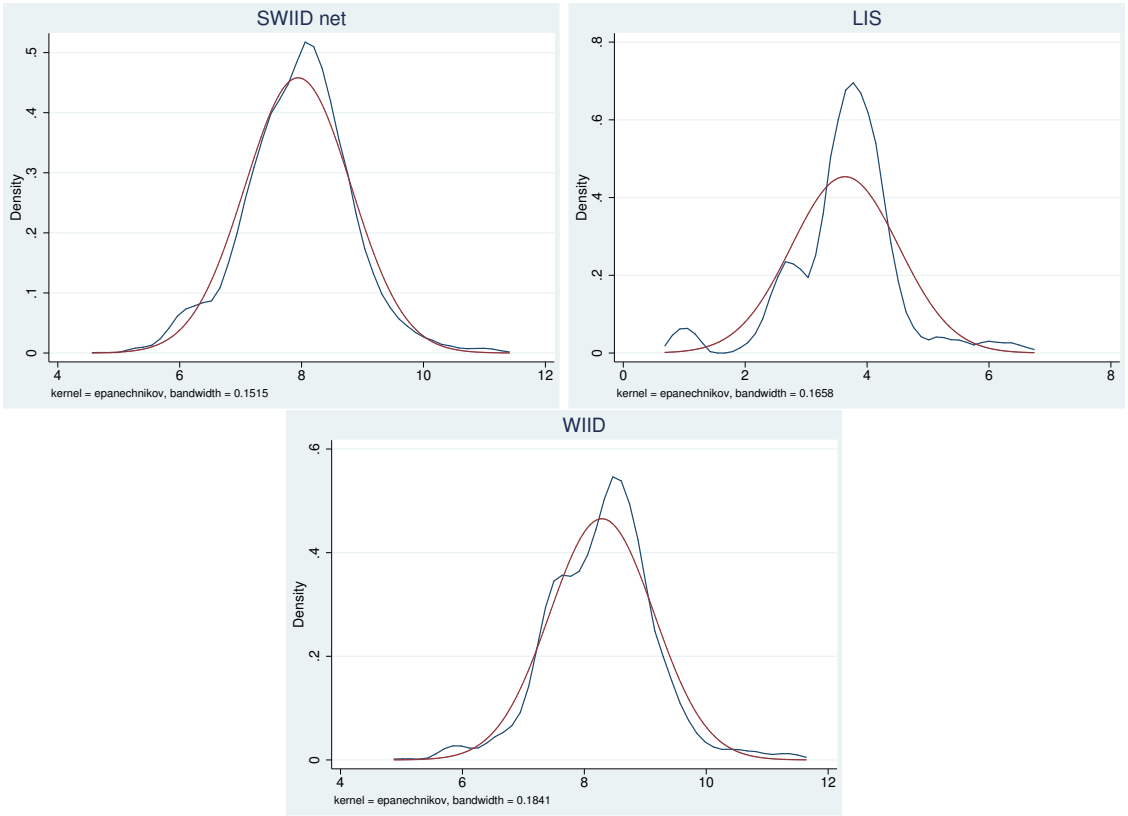


Figure A.4: Kernel densities of the log differences between the Theil index and Gini coefficients of inequality



3.B Appendix

Imputation using fitted values from linear regressions

Filling missing values with predictions obtained from a regression permits the exploitation of further available information provided in the UNIDO industrial statistics. For the index of inter-industry wage inequality, only the data on wages and employment are needed. However, the UNIDO dataset also contains information on other sector-level characteristics. For the prediction of missing values, I use data on the number of establishments and on output as additional explanatory variables. If only one of the two variables needed for the computation of the index is missing, the other one is used in the regression as well. A time trend is also included in the set of potential regressors.

Once the first set of fitted values has been obtained from all available regressors, the next step is to assess the plausibility of the obtained prediction of the missing. Checking for plausibility is crucial for two reasons: First, the Theil index uses logarithmic transformations, which does not allow the inclusion of negative values. Second, because the index is based on the *ratio* of shares, a too-large or too-small number in one variable has the potential to affect the sector's contribution rather substantially and lead to disruptions in the series which may be unwarranted. In other words, the aim of the imputation is to arrive at plausible values for the missings, which at the same time should keep the series of inequality statistics smooth.

A good fit of the surrounding data points provides some indication of the appropriateness of the underlying regression model for a particular missing value. While the R^2 seems like an obvious candidate to judge the general goodness of fit, a high R^2 can sometimes be misleading, especially when the time series is long. In several cases, the fit is very good for part of the data, but captures relatively little of the variation in other parts. Whether the fitted values are useful for imputation then depends on where the missings are located. Each and every fitted value is therefore checked individually, and the regression is adjusted if necessary.

If an imputed value is deemed implausible, there are 3 principal ways to adjust the regression: (1) changing the regressors, (2) changing the time period, or (3) changing the imputation method. Only the first two options are discussed in the following, whereas the

other imputation methods adopted are presented in the next subsections.

The first step is, of course, to identify "bad" fitted values, and predictions which are considered implausible for reasons other than a generally poor fit of the surrounding data points. Deviations of more than 30 percent of the fitted- from the actual values of the data points surrounding the missing are considered problematic and warrant changes in the regression model used. Then there are problems which arise occasionally despite a relatively good fit. The most obvious is a negative fitted value, which is not only problematic mathematically, but also conceptually impossible for wages or employee numbers. Along the same lines, even if the overall fit is good and the predicted value is positive, it can still be implausibly low or high. This basically happens when the values in the forcing variables suggest a value very different from the one obtained, and is mostly caused by large changes in one of the predictors to a level which does not occur elsewhere in the underlying data. Similarly, predicted values can be very different from their "surrounding" values and this is clearly not warranted by an extreme value in one of the forcing variables. These problems are of course related in many cases. In particular, negative values are just special case of an implausibly low fitted value. Similarly, their causes as well as the strategies for addressing them apply for several of the above cases.

Once a problematic fitted value has been identified, the next step is to check the coefficients of the individual variables to see whether a single regressor is driving the result.[58] Things that may indicate problems are negative coefficients (given that the initial reason for including the regressor was the assumption of a positive relationship) or a very large (or very small) size of an individual coefficient. Often, dropping the respective variable - which can also be the constant - solves the problem and yields a more realistic estimate of the missing value. However, it is not always possible to clearly identify an individual variable causing the problem. In many cases, all variables are useful in predicting a missing value, and it is not the set of variables but the time period which needs to be changed. This is especially true for long time series since the association between some of the variables is likely to not remain constant over a time span of 30 years

[58] While significance may seem like an obvious indicator of whether or not a regressor is useful, in many cases, the number of observations is too low to allow a judgement of which regressors to keep based on or the significance of the estimated coefficients - and whether or not they are robustly related to the regressand during the time period of interest.

or more, and sometimes changes visibly already in shorter time periods.

Some examples and illustrations with graphs and tables of the underlying data will help demonstrate the conjunctures encountered. Starting with the case of negative fitted values, suppressing the constant forces the regression line through the origin - an all but reasonable assumption, which helps to resolve the problem in many instances. An example is given below for missings in wages in Bolivia in sector 34 between 1971 and 1973, illustrated in table 3.B.1.[59]

Table 3.B.1: Example of Bolivia: suppressing the constant

					without constant		with constant	
Year	Empl.	Wages	Estbl.	Output	Fv	%dev	Fv	%dev
1970	59			166667	3588		**-112731**	
1971	65			166667	37132		**-102628**	
1972	25							
1973	33			50000	13956		-166461	
1974	166	100000		550000	112416	12	100131	0
1975	150	150000		750000	137362	8	90243	40
1976	318	750000		3900000	617896	18	641739	14
1977	352	650000		4300000	681558	5	733099	13
1978	313	550000		4800000	744133	3	710063	29
1979	478	1400000		8750000	1337179	4	1324724	5
1980	513	1200000		10280000	1560859	30	1514124	26
1981	577	1680000	22	15960000	2377494	4	2106231	25
1982	363	562500	14	4203125	670153	19	743361	32

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %.

As mentioned previously, implausible fitted values can be driven by outliers in one of the regressors. An example is the case of El Salvador shown in table 3.B.2 below, where the regression yields a very small number for the missing in wages in 1992 in sector 34. This is clearly due to the very low value of 22 of the explanatory variable "employees" in that year in comparison to the rest of the data for this sector, where employee numbers are always above 100. Obviously, the resulting wage number should also be substantially smaller than before, but is arguably not in the 3-digit range, as indicated by the value preceding the missing which is still around one third of the larger values in the later years. Here, suppressing the constant alone does not solve the problem. Only when the year 1998, containing substantially larger numbers for both wages and employees, is also

[59] An alternative would be to allow for a different functional form, e.g., by including a cubic term. However, due to the often few degrees of freedom, this is not always feasible. Given the theoretically valid assumption of a constant of zero, this approach is hence preferred.

omitted from the regression does it yield plausible numbers for the missing wages. In this case, plausibility is not only assessed through the deviations of the fitted values for the surrounding values, but also from observations with similar values for other regressors (in this case, establishments).

Table 3.B.2: Example of El Salvador (sector 34): suppressing the constant

					without constant				with constant	
					without 1998		with 1998			
Year	Empl.	Wages	Estbl.	Output	Fv	%dev	Fv	%dev	Fv	%dev
1991		124740	1	1995841						
1992	22		1		96187		773		234452	
1993	135	306463	5	3196204	242767	21	624579	104	413323	35
1994	112	520464	2	3403122	212973	59	539949	4	391284	25
1995	152	309781	16	259864	264881	14	783491	153	462313	49
1996	273	731696	10	4286122	421835	42	1448971	98	653002	11
1997	223	571249	11	3397116	357026	38	1223695	114	591077	3
1998	968	7154426	8	20503255	1323221	82	5139651	28	1703586	76

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %.

A generally bad fit is frequently caused by a particular data structure, wherein one can observe 2 different "regimes" in the development of wages and employees over time. Pooling these together in one regression yields a mediocre fit for both regimes. Excluding the years which display a different pattern from the one where the missings are located often solves the problem. Again using the example of El Salvador, in the first spell of data (pre-1985), wages are fairly stable, whereas in the second data spell (post-1990), they are much more dispersed and display high growth rates. Only post-1990 values are therefore used for the imputation of wages in 1992 in sectors 26, 27, 28, and 31. To give a better impression of this type of data structure, wages in these sectors are plotted in figure 3.B.1.

Figure 3.B.1: Log-normality of the (normalized) Theil index (wages in mn. USD)

That the association with a forcing variable is deficient - as indicated most clearly by a large negative coefficient - occurs repeatedly in the estimations. It is easy to detect, and the straightforward thing to do is drop the respective variable. This improves the result in most instances. An example is Mozambique, where the "establishments" variable has a negative and relatively large coefficient in sector 34 for explaining employee numbers, and produces a correspondingly poor result with partly negative fitted values. The exclusion of the variable leads to a substantial improvement of the fit and yields positive values. Table 3.B.3 contains the raw data, and table 3.B.4 displays the results with and without the exclusion of establishments.[60]

[60]Note that although the "output" variable also has a negative coefficient, its effect is much smaller and its exclusion does not lead to a better fit, nor does it solve the problem of negative fitted values. Also note that the variable turns positive once establishments have been excluded from the regression equation.

Table 3.B.3: Example of Mozambique (sector 34, 1997/98 employees): dropping a variable

					without estbl.		with estbl.		without output	
Year	Empl.	Wages	Estbl.	Output	Fv	%dev	Fv	%dev	Fv	%dev
1986	1240		5	7892849	1433199		4335966		1179593	
1987	2710			7160227	1317665					
1988	2920		10						980049	
1989	2760		9						916807	
1990	2487	1767323	9						843129	52
1991	2276	943206	9	6923114	909802	4	1220781	29	76946	18
1992	1102	590888	11	3145974	732922	24	600248	2	674896	14
1993	1138	640900	11	2064663	610363	5	613682	4	601218	6
1994	1049	875039	10	2618161	520744	40	888217	2	537976	39
1995	446	179071	11	2179552	411135	130	196929	10	45386	153
1996	911	327082	10	3541943	337816	3	313905	4	390618	19
1997			**24**	**4005683**	**246389**		**-8087087**		**170818**	
1998			**35**	**17402439**	**415559**		**-1.7E+07**		**-17670**	
1999			19						75647	
2000	475	68430	17						22842	67

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %.

Table 3.B.4: Example of Mozambique (sector 34, 1997/98 employees): regression output

Dep. var.: Wages	(1)	(2)	(3)
Output	0.0201	-0.195	
	(0.258)	(0.0366)	
Establishments		-579,534*	-10,437
		(50,212)	(72,769)
Year	-100,771	-197,187**	-73,679
	(102,847)	(15,080)	(68,392)
Constant	2.014e+08	4.004e+08**	1.476e+08
	(2.049e+08)	(3.038e+07)	(1.358e+08)
Observations	5	5	6
R^2	0.329	0.995	0.567

Notes. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The three columns contain the regression output for the three sets of fitted values/ %-deviations of table 3.B.3.

Individual variables can also cause a generally bad fit and dropping the variable often helps resolve the problem. An example is sector 36 in Bulgaria, where the inclusion of the variable "establishments" leads to large deviations and negative values in some years (see table 3.B.5). Here, the large coefficient on the variable (shown in table 3.B.6) is indicative of the problem. This example also demonstrates how a short time period restricts the options for achieving a better fit: excluding the later years containing substantially higher numbers for establishments would theoretically also be possible, but would leave even fewer years for estimation. Excluding the "establishments" variable is therefore preferable in this context.

Table 3.B.5: Example of Bulgaria (sector 37, 1996 wages): dropping a variable

					without estbl.		with estbl.	
Year	Empl.	Wages	Estbl.	Output	Fv	%dev	Fv	%dev
1996			26	736414	95838.74		771319.9	
1997	100	117882	26	1413299	103559.1	**12**	509868.9	**333**
1998	100	120108	25	1411643	102262.8	**15**	187921.1	**56**
1999	123	131675	18	2722748	118410.9	10	-433161.2	429
2000	303	370117	19					
2001	290	219709	24					
2002	426	385656	40					
2003	550	772205	46	47902118	713671.8	8	493296	36
2004	158	206335	57	5079010	143350.2	31	292953.3	42
2005	282	387516	64	7623259	175884.7	55	420798.6	9
2006	1211	2174736	62	2.53E+08	3431034	58	2952456	36
2007	1540	5682796	86	3.56E+08	4805079	15	5169121	9

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %.

Table 3.B.6: Example of Bulgaria (sector 37, 1996 wages): regression output

Dep. var.: Wages	(1)	(2)
Output	0.0133***	0.0119***
	(0.00260)	(0.00241)
Establishments		52450
		(31,883)
Year	-1274	-269478
	(94,069)	(182,145)
Constant	2.630e+06	5.373e+08
	(1.882e+08)	(3.634e+08)
Observations	8	8
R^2	0.909	0.946

Notes. Standard errors in parentheses, *** p<0.01, ** p<0.05, * p<0.1. The two columns contain the regression output for the two sets of fitted values %-deviations of table 3.B.5.

There are also reasons to ex ante exclude available regressors from the estimation. This often concerns the "establishments" variable, where the absolute numbers are sometimes very low (single-digits) and/or where there is little variation (e.g., in Lesotho, the number of establishments is constant and below 10 in several sectors for the 2001 to 2008 period and varies only by 1 in a few others). Another limitation to the inclusion of regressors is the number of observations. Due to the unbalancedness of the data, including an additional variable often reduces the years available for estimation. Again using the example of Lesotho, for predicting missing employee numbers in sectors 17 and 24 in 1980/81, the "output" variable is available only for years post-1981, but not for the earlier ones. As its inclusion would reduce the number of observations to a level where no degrees of freedom are left for estimation, it is dropped for the estimation. Omitting the constant also increases the degrees of freedom and performs better in other sectors where the output variable is more important in predicting the missing and is therefore retained.

Implausible values of any kind can of course also arise from a simple outlier in one of the forcing variables, in which case it suffices to exclude the respective year from the regression. An example from Puerto Rico is shown in table 3.B.7, wherein employee numbers drop to 650 in one year from a level of around 4000 in all other years, worsening the fit substantially.

Table 3.B.7: Example of Puerto Rico (sector 26, 1987/88 wages): outlier years

					with 1998		without 1998	
Year	Empl.	Wages	Estbl.	Output	Fv wages	%dev	Fv wages	%dev
1987	4520		162		2.33E+07		4.09E+07	
1988	4630		164		2.91E+07		4.87E+07	
1989	4950	34800000	167	5.00E+08	3.11E+07	11	5.75E+07	65
1990	3920	36900000	159	5.01E+08	5.81E+07	57	6.05E+07	64
1991	3620	92500000	156	5.02E+08	7.17E+07	23	6.66E+07	28
1992	3460	92100000	41	5.06E+08	9.42E+07	2	8.10E+07	12
1993	3370	93800000	151	5.25E+08	9.25E+07	1	8.06E+07	14
1994	3600	98300000	157	5.67E+08	9.57E+07	3	8.87E+07	10
1995	3620	1.03E+08	162	5.42E+08	1.03E+08	0	9.59E+07	7
1996	3610	1.06E+08	196	6.32E+08	1.07E+08	2	1.01E+08	4
1997	3760	1.09E+08	184	7.00E+08	1.14E+08	5	1.10E+08	1
1998	**650**	96400000	**37**	7.03E+08	1.92E+08	99	1.12E+08	16
1999	4010	1.19E+08	214	7.69E+08	1.22E+08	3	1.24E+08	4
2000	4340	1.31E+08	197	7.38E+08	1.26E+08	4	1.34E+08	2

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %.

If none of the described regression-based solutions yield any useful results, another type of imputation is applied. The idea is similar to a simple linear interpolation but still exploits some of the information contained in other variables. The method assumes a constant co-movement of the missing with another variable (that is, not using the "year" variable, which would be the case in a "normal" linear interpolation) and traces its development in the missing years. Missing employees in 1995 in sector 35 in Slovenia provide an example, shown in table 3.B.6. Here, the numbers rise so drastically for other variables in the years following the missing that no regression model can be found which results in acceptably low, but still positive, numbers. In this case, the problem stems from a lack of support for numbers of this magnitude in the data. Employee numbers have to be imputed differently and are assumed to move in accordance with wages[61] in the concerned years.[62]

Table 3.B.8: Example of Slovenia (sector 35, 1995 employees): linear interpolation

Year	Empl.	Wages	Estbl.	Output	Fv empl.	Fv wages	%dev wages 1997	%dev wages 1996
1995			33	84922771	**1373**	17911636		
1996	144		33	53110129	144	7478213		6
1997	143		43	39017597	143	2982585	4	25
1998	835	10097149	51	57697994	835	10097149	4	15
1999	2808	39551278	60	1.27E+08	2808	39551278	5	19
2000	2818	34440932	64	1.25E+08	2818	34440932	6	8
2001	1362	12833498	69	69103452	1362	12833498	5	3
2002	1373	17954450	87	1.02E+08	1373	17954450	3	32
2003	1317	20826801	94	1.16E+08	1317	20826801	11	36
2004	1332	9965221	119	67265244	1332	9965221	25	19
2005	2027	39793868	128	1.44E+08	2027	39793868	8	4
2006	2568	51433489	157	2.06E+08	2568	51433489	4	15
2007	2587	60224473	170	2.55E+08	2587	60224473	31	26
2008	1496	42479950	86	1.93E+08	1496	42479950	4	29
2009			84			17911636	4	
2010			81			7478213	5	

Notes. Fv is short for fitted values. %dev is the deviation of the fitted from the observed values in %. The estimation for wages in 1997 is based on a regression of wages on employees, output, establishments, and a time trend, whereas the 1996 value is based on output only.

Other imputation approaches

If no information is provided for other variables which would allow a regression-based

[61] The imputed values for wages are used, which have a fairly good fit in the early years (shown in the last two columns table 3.B.6).

[62] The relatively high correlation of 0.78 between the two variables supports this assumption.

imputation, a simple linear interpolation between the surrounding values is performed. This is equivalent to the above approach of imputation alongside another variable, but always using the "year" variable. For example, in the case of Bangladesh, there is no data in the years 1993 and 1994 but values for both wages and employees are available in 1992 and 1995. In sector 15, the number of employees is 107882 in 1992 and 126220 in 1995. The resulting values for 1993 and 1994 are calculated as (126220-107882)/3+107882 $\approx$ 113995, and (126220-107882)/3*2+107882 $\approx$ 120107. Imputation based on linear interpolation hence always implicitly assumes a linear development over time of the target variable for the missing years between the two surrounding data points.[63]

A disadvantage of the linear interpolation approach is that is requires both a start- and an end-observation for the missing time period. If a missing is located in the first or last year of the available data, the method cannot be applied. Instead, for missings located at the beginning or the end of a data spell, a time trend is used to extrapolate values when no other information is provided by the dataset - again exploiting the year variable.[64] The same procedure as with the regression-based approach is applied, including the option to drop the constant or change the time period when the fit is bad.

Again, there are several cases where it is not possible to find a good fit which would support extrapolation based on a time trend. In those cases, the first (or last) available value - which is in some cases an imputed one[65] - is then repeated in the missing years. This has, e.g., been done in Fiji in 1995, where the 1996 value is used to fill the missing in sector 20. Whether such a procedure is reasonable also depends on the development of

[63] In a few cases, means imputation is based on starting or ending values which are the result of a regression-based fitted value imputation (e.g., sector 27 in Fiji for the missings in wages in 1994-95, which use as the "starting" observation the fitted value of 1993). This is done because the alternative would be to start linear interpolation at the closest data points provided by the raw data, which means that the fitted value would be overruled by the linearly interpolated one. This runs contrary to the initial idea that regression-based fitted values are always favoured over linear interpolation, as they incorporate all of the information available from the data.

[64] One could also use a time trend to impute values missing "in the middle" as an alternative to the simple linear interpolation described above. Using a time trend is advantageous when there are outliers at the beginning or end of a linear imputation, as well as a discernible time trend in the data. Otherwise, it is less suppositional to assume that values do not range outside the value of the start and the end year of the gap. Given that the goal is to tamper with the data as little as possible, if no further information is provided in the data which would point towards the missing going into a particular direction, it is desirable to merely preserve the ratio of wage and employee numbers in order to maintain time coverage, but influence the inequality index as little as possible. Linear interpolation effectively means that the resulting contribution of the imputed missing will lie between that of the start and that of the end year, and is therefore the least intrusive option and preferred over the use of a time trend.

[65] An example for when an imputed value was used to fill missings in the first few missing years (1970-1973) is sector 27 in Indonesia.

the data in the preceding or following years: if they have been relatively stable, using the same values seems valid.[66]

Of course, the discrete procedure of imputing data on a case-by-case basis is inherently arbitrary. This applies not only to the imputation procedure, but also to the preceding decision of whether or not to impute in the first place. As a rule of thumb, no sector is used in the final index in which more than 50% of the data need to be imputed.

[66]There are only two cases with imputation approaches different from the ones described, but based on the same techniques. The first one is Bulgaria, which is the only case where a squared term has been employed to impute employee numbers post-2003 due to the clearly discernible inverse U-shaped development of employee numbers in sector 23. The second case is Tunisia, where the fitted values for sectors 22, 28, and 36 are located in a time period relatively far from another data spell containing support for both variables involved in the imputation. The fitted values are the result of an average of two imputation approaches with very different results, but an equally good fit in their respective (non-overlapping) parts of the data. The resulting fitted values line up nicely with the values of the time series following the missing.