UNIVERSITÀ DEGLI STUDI DI BARI ALDO MORO

Dipartimento di Economia e Finanza

Southern Europe Research in Economic Studies

Beyond the weights A multicriteria approach to evaluate Inequality in Education

Giuseppe Coco

Raffaele Lagravinese

Giuliano Resce

SERIES Working Papers n. 06/2020

Beyond the weights: A multicriteria approach to evaluate Inequality in Education*

Giuseppe Coco
Dipartimento di Scienze per l'Economia e l'Impresa, Università di Firenze

Raffaele Lagravinese†
Dipartimento di Economia e Finanza, Università di Bari

Giuliano Resce
Italian National Research Council (CNR)

Abstract

This paper proposes the use of a new technique, the Stochastic Multicriteria Acceptability Analysis (SMAA), to evaluate education quality at school level out of the PISA multidimensional database. SMAA produces rankings with Monte Carlo Generation of weights to estimate the probability that each school is in a certain position of the aggregate ranking, thus avoiding any arbitrary intervention of researchers. We use the rankings in 4 waves of PISA assessment to compare SMAA outcomes with Benefit of Doubt (BoD), showing that differentiation of weights matters. Considering the whole set of feasible weights by means of SMAA, we then estimate multidimensional inequality in education, and we disentangle inequality into a 'within' and a 'between' country component, in addition to a component due to overlapping, using the multidimensional ANOGI. We find that, over time, inequality within countries has increased substantially. Overlapping among countries, particularly in the upper part of the distribution has also increased quite substantially suggesting excellence is spreading among countries.

Keywords: Education inequality; PISA; SMAA; ANOGI; anywhere and somewhere;
JEL Classifications: : I14, C44.

*The authors wish to thank Paolo Liberati and the participants of the Workshop "Equity in Education held at " Faculty of Economics & Business, Katholieke Universiteit Leuven, Belgium. 30 November-1 December 2017 and the partecipants of the Italian Economic Society held in Palermo 24-26 October 2019.

†Corresponding Author: Department of Economics and Finance, University of Bari "Aldo Moro"Largo Abbazia Santa Scolastica, 70124 - Bari, Italy. email: raffaele.lagravinese@uniba.it

1 Introduction

In the last decade, there has been an increase of detailed international surveys on cognitive achievement tests. Among them, the Programme for International Student Assessment (PISA) is one of the most influential and used to measure student performances in different subjects (mathematics, science and reading). The performances in different dimensions are usually averaged in order to obtain a composite indicator to be used for ranking and comparisons among schools and states (e.g., Bloom et al., 2015). However, the average score can hide different attitudes and specialisations. Thus, a crucial issue is how to define a proper set of weights to aggregate different subjects. Decancq and Lugo (2013) distinguish three classes of approaches to weight dimensions into a composite index: data-driven, normative, and hybrid. The weights in data-driven approaches depend solely on the distribution of the elementary indices, normative approaches set the weights on the basis of value judgments, the hybrid approaches combine the information on the distribution of the elementary indices and the value judgments. In the absence of information about value judgments, as it is the case of PISA, the Benefit of Doubt - (BoD) methodology, has received considerable attention in the education sector (De Witte, López-Torres, 2017; Karagiannis and Paschalidou, 2017; De Witte, Schiltz, 2018). Some students may be more brilliant in mathematics, others in science or reading. Yet, some schools are more specialised in specific subjects[1]. The BoD model accounts for this by endogenously weighting the school outputs. In detail, BoD allows to aggregate linearly quantitative performance sub-indicators into a single composite one using the combination of weights that is the most convenient for the evaluated Decision Making Unit (DMU)(Cherchye et al, 2007). For each evaluated DMU, this is done by implicitly assigning less (more) weight to those subindicators or aspects of performance that the particular DMU is relatively weak (strong) in compared to all other DMUs in the sample (Karagiannis and Paschalidou, 2017). For this reason, decision makers should not complain about unfair weighting schemes, since each DMU is put in its most favourable condition, as any other weighting scheme would generate a lower composite index (Cherchye et al. 2008). Thanks to its low requests in terms of exogenous assumptions for setting weights, BoD evaluations represent the standard in recent literature on education systems evaluations among the non-parametric techniques (see De Witte, Lóopez-Torres, 2017 for a review). However, a relevant drawback of BoD, as well as of any proposed data driven approaches, is the uniqueness of weights vector to evaluate DMUs (Greco et al. 2018). This uniqueness requires the assumption of "representative agent", summing up in itself the preferences of all the individuals potentially interested in the composite indicator. Since in a group of people each one may assign a radically different importance to the considered dimensions, in order to ensure that the composite indicator is meaningful, the diversity of existing viewpoints should be considered (Decancq et al. 2013). Moreover while apparently being judgment free, the BoD approach implicitly favours more specialized environments, in particular where most students have the same specialization. Althought implicit this hides a value judgment where homogenous specialization at school level is desirable.

Compared to the BoD methodology, our proposal is to aggregate the schools' attainments

[1]For instance, in Italy the education offered by a liceo (*lyceum*) is mostly academic. Individual lyceums will cover the core subjects and specialise in specific fields of study; this may be the humanities, mathematics and science, or art.

on mathematics, reading, and sciences considering not only a single weight but the whole space of feasible weights vectors. From a methodological standpoint, we use the idea of Greco et al. (2018), where the Stochastic Multicriteria Acceptability Analysis approach (SMAA) is used to take into account a large sample of randomly extracted vectors of weights to rank alternatives. According to this methodology, each Decision Making Unit (school in our case) is assigned a probability of being in a given position in the rank in terms of the composite index. With this innovative approach, we propose to summarise the multidimensional education's outcome without any a priori judgement on specific vectors of weights. To the best of our knowledge, this is the first application of SMAA in education. In order to better understand the differences between these two methodologies, we apply both BoD and SMAA to four waves of PISA (2006, 2009, 2012, 2015) assessments to produce an overall (probability) ranking of school with the aim of evaluating the inequality within and across countries in each wave and then to identify trends in education inequality over this period. PISA scores are intrinsically unsuited to identify overall trends in inequality as the distribution of scores in each wave is normalised and the information is multidimensional. However, by using the ANOGI proposed by Yitzhaki (1994) and its multidimensional generalisation proposed by Lagravinese et al. (2019) on SMAA outcomes, we can evaluate changes in different components that explain the overall inequality.

Some relevant differences can be observed between the ranking obtained by BoD and that obtained by SMAA, suggesting that differentiation of weights matters even using international standardised surveys in the education domain. Exploring the whole set of feasible weights, within-countries inequality has substantially increased over the period 2006-15, while between-countries inequality decreased (bear in mind that only OECD countries are surveyed). This suggests a relative convergence of education systems, but also more inequality within national systems. In particular, we find that overlapping among countries in the distribution of excellent schools (top 20% performers in the world) has increased quite substantially. This suggests that in every country a certain share of the population is building up world-class human capital, potentially useful across borders.

Our findings are particularly meaningful for a political interpretation as they demonstrate for the first time that educational inequality within countries increased during and after the financial crisis, potentially fuelling the new electoral divide. In particular, the result of increasing overlapping of the excellent section of schools across countries lends some credit to the theory according to which in the last decades advanced societies have been experiencing a divide between 'Somewhere' and 'Anywhere' individuals, the latter being a sovranational class (Goodhart, 2017). Education inequality has been identified as one, if not the main, of the drivers of the recent populist backslash around the world both by scholars pointing to economic causes and by those pointing at a cultural divide (Picketty, 2018, Inglehart and Norris, 2016). Moreover, it has been found to be the best predictor of a populist electoral choice in many countries (again Picketty, 2018, Kriesi, 1999, Goodhart, 2017). The segregation of secondary school students into different schools may have important consequences for educational inequality, social cohesion and intergenerational mobility (Gutierrez et al. 2019). The rest of the paper is organised as follows: the next section presents the PISA database, section 3 deals with methodological topics, section 4 shows the results using both BoD and SMAA, section 5 shows multidimensional inequality in education, and section 6 concludes.

2 Data

In recent years, the PISA databases have been widely used in order to investigate the inequality in education in different countries. Measuring inequality in the educational sphere has been the aim of many recent contributions, often focusing either on opportunity for access to a given level of studies (e.g., Paes de Barros et al. 2009; Vega et al. 2010), or on opportunity in terms of educational achievement (e.g., Checchi and Peragine, 2005; Bailey and Borooah, 2010; Ferreira and Gignoux, 2014; Gamboa and Waltenberg, 2012; Agasisti et al. 2018).

Our analysis is conducted using data at school level collected by the PISA surveys in 2006, 2009, 2012 and 2015. The time span investigated allows to capture possible changes in the distribution of school achievement and performances across countries during a period characterised by a global economic recession. The database contains 9,955 schools in 2006, 10,867 schools in 2009, 11,605 schools in 2012, and 9,193 schools in 2015, and covers 34 OECD countries (see Table 1 for descriptive statistics). Overall, a substantial share of the cognitive items across reading, mathematics, and science domains requires manual coding by trained coders. It is crucial for comparability of results that students' responses are scored uniformly from coder to coder, and from country to country. Comprehensive criteria for coding, including many examples of acceptable and unacceptable responses, prepared by the OECD are provided to National Service Providers (NSP) in coding guides for each of the three domains: reading, mathematics, and science. Students' competencies are expressed in terms of "plausible values", which are obtained via a two-step procedure. The first step deals with the distribution of the students' latent abilities, which is obtained by adopting the item response theory (IRT) statistical technique. In the second step, a new distribution is derived by applying an affine transformation to the distribution that was generated in the first step. This process produces an arbitrary metric for test scores, which are then typically standardised to some arbitrary mean and standard deviation which are set (by OECD) to 500 and 100, respectively. In sum, the scaling methodology in PISA waves remained the same as for trend comparisons, making the analysis consistent between among cycles and comparable with different PISA waves. As is shown in table 1, there were consistent changes among countries in the various subjects over time. Among the OECD countries, in the four waves Japan and Korea were the best performing countries in math followed by Netherland and Poland. Finland, Estonia, Ireland and Japan, on the other hand, are the countries with the best performances in reading. States like Japan, Estonia, Finland and Canada, are also the four highest performing OECD countries in science. During the 4 waves, many countries recorded a reduction in some subjects and a performance, significant improvement occurred only in few countries: Chile, Israel, Norway, Portugal and Sweden. Looking at the performance of individual disciplines is very important, due to the effects that can be had on future growth and earnings. For instance, Murnane et al. (2000) suggest that a 1-standard-deviation increase in mathematics performance at the end of high school translates into 12 percent higher annual earnings. Also the evaluation of schools and the quality of performance in standardized tests are very useful to promote the growth of the economy (Hanushek and Raymond, 2006). For all these reasons, analyzing the performances of individual subjects and their evolution at school level is an aspect to be carefully evaluated.

Table 1: Descriptive Statistics

	2006			**2009**			**2012**			**2015**		
State	Mean	S. dev.	Freq.	Mean	S. dev.	Freq.	Mean	S. dev.	Freq.	Mean	S. dev.	Freq.
AUS	513.18	48.58	356	510.99	48.86	353	500.57	58.55	775	492.33	53.16	758
AUT	486.65	77.24	199	467.82	75.87	282	484.19	70.51	191	479.07	70.3	269
BEL	509.61	76.92	269	499.87	84.69	278	502.08	78.88	287	493.39	74.1	288
CAN	514.11	49.43	896	512.97	46.62	978	509.92	44.34	885	508.35	41.95	759
CHE	501.95	57.12	510	502.37	54.09	426	500.92	53.7	411	497.57	61.08	227
CHL	417.95	74.57	173	424.47	63.43	200	443.26	66.8	221	447.23	68.37	227
CZE	516.92	82.76	245	495.77	78.29	261	503.07	71.1	297	486.94	69.89	344
DEU	496.68	84.8	226	500.54	79.55	226	507.86	73.25	230	502.48	69.39	256
DNK	501.59	43.94	211	483.88	42.58	285	481.76	46.39	341	490.47	42.43	333
ESP	493.02	37.67	686	489.25	42.84	889	493.9	43.8	902	493.85	33.85	201
EST	514.9	41.77	169	512.4	41.11	175	523.51	39.8	206	520.89	41.22	206
FIN	552.7	25.49	155	540.01	35.81	203	516.16	46.53	311	520.85	41.07	168
FRA	489.42	74.29	182	492.07	77.64	168	493.01	79.45	226	487.23	75.68	252
GBR	499.45	52.49	502	495.9	48.38	482	497.52	50.49	507	495.41	46.15	550
GRC	448.31	69.78	190	467.04	59.32	184	451.74	64.74	188	451.02	64.93	211
HUN	469.98	79.98	189	475.4	77.74	187	471.73	75.41	204	458.64	75.12	245
IRL	507.87	40.25	165	495.05	46.46	144	512.33	42.13	183	505.8	35.79	167
ISL	500.65	52.32	139	500.74	48.89	131	480.99	45.26	134	479.17	36.22	124
ISR	442.54	69.33	149	456.29	71.96	176	472.72	73.02	172	467.83	72.03	173
ITA	467.03	75.1	799	482.21	68.55	1097	478.69	73.14	1194	483.02	62.3	474
JPN	515.23	69.77	185	527.98	71.29	186	538.02	68.22	191	527.61	61.84	198
KOR	539.63	56.61	154	538.56	49.71	157	540.29	54.83	156	515.9	51.55	168
LUX	487.8	53.21	31	482.69	60.64	39	488.85	57.12	42	486.26	58.7	44
LVA	485.96	42.34	176	483.69	41.6	184	489.34	45.51	211	480.63	40.17	250
MEX	419.05	52.15	1140	418.26	51.07	1535	413.84	48.66	1471	411.37	45.52	275
NLD	524.04	72.45	185	525.33	71.88	186	512.85	75.67	179	506.63	75.73	187
NOR	489.37	37.1	203	500.84	36.97	197	497.34	41.67	197	504.46	32.93	229
NZL	521.01	44.59	170	522.59	49.38	163	507.67	56.02	177	499.59	48.84	183
POL	525.33	60.56	221	506.99	40.96	185	529.76	56.69	184	509.46	39.18	169
PRT	470.3	54.19	173	483.44	49.36	214	479.9	55.02	195	478.97	54.89	246
SVK	477.76	66.96	189	478.95	62.36	189	461.15	73.68	231	451.59	66.99	290
SVN	462.56	76.27	361	455.26	73.5	341	457.95	75.53	338	470.66	71.31	333
SWE	506.98	42.43	197	502.01	49.09	189	485.79	51.38	209	499.18	47.48	202
TUR	431.75	64.4	160	443.92	69.44	170	450.05	69.49	170	409.26	58.63	187
Mean	485.47	68.87	9955	490.66	58.43	10876	483.2	67.25	11816	486.35	62.24	9193

Source: Authors' elaboration on data from OECD (2017a)

3 Methodology

3.1 The multidimensionality of education outcomes.

In order to compare the multidimensionality of education outcomes using different weights associated for each subject (mathematics, reading and science) we perform two no-parametric methodologies: BoD and SMAA.

The general framework in PISA sample is as follows. We have the set of schools A to be evaluated on the set of the average student's attainments on mathematics, reading, and sciences G (in line with previous evaluation on PISA, e.g. De Witte, Kortelainen, 2013, Lagravinese et al. 2020, we use the plausible values 1):

$$A = \{a_1 ....., a_m\} \tag{1}$$

$$G = \{g_1 ....., g_n\} \tag{2}$$

The school-level function that aggregates attainments in different subjects can be assumed as the weighted average of the three scores multiplied by the weights associated to each of the subjects. For each school $a_K \in A$ we can estimate the following individual CI of performance depending on a set of weights w:

$$CI\ (a_{k,}w) = \sum_{i=1}^{n} w_i g_i(a_k) \tag{3}$$

where w_i reflect the importance that that we give to the subject i, and $g_i(a_k)$ is the average score in the school a_k for the subject i. The main problem is that the order of importance given to different attainments is a subjective choice, which implies that one single objective vector of w does not exist. It poses the problem about the choice of weights in the absence of a priori information (Lagravinese et al. 2019). Two main solutions have been proposed to this issue: data-driven weights (such as Data Envelopment Analysis (DEA), Charnes et al. 1978); and a large set of random weights (such as Stochastic Multicriteria Acceptability Analysis, Lahdelma et al. 1998; Lahdelma and Salminen, 2001).

3.2 Considering data-driven weights-BoD

To avoid a set of weights reflecting a merit good approach, DEA without input (BoD), have been extensively employed as technique of aggregation in education (Decancq, Lugo, 2013; De Witte, Schiltz, 2018; Greco et al. 2018). Formally, BoD is a standard DEA where all inputs are assumed to be equal to 1 for all evaluated schools. This solution was originally proposed by Thompson et al. (1986), used by Melyn and Moesen (1991), and formalized under a DEA framework by Lovell and Pastor (1999). After Nardo et al. (2008) which defined BoD one of the methods to construct composite indicators, BoD has been used in many applications in different fields. In the education sector, De Witte, Schiltz (2017) used BoD for measuring and

explaining organisational effectiveness of school districts. The basic assumption of the BoD evaluations is that the status-quo is a choice of the local decision maker (Cherchye et al. 2007). On this assumption, the BoD estimates a composite index based on the combination of weights that is the more convenient for the evaluated school. Formally the model can be translated into the following linear program:

$$\begin{aligned} CI &= \max \sum_{i=1}^{n} w_i g_i(a_k) \\ \sum_{i=1}^{n} w_i g_i(a_j) &\leqslant 1, j = 1, ..., m \\ w_i &\geqslant 0, i = 1, ..., n \end{aligned} \tag{4}$$

A school is considered to be best performing if it obtains a score of one in the optimal solution of the linear program. A score less than one implies that the school is under-performing, the lower the index, the lower the effectiveness. The weights in the objective function are chosen automatically with the purpose of maximizing the score of the k-th school. The optimization ensures that each school is evaluated on the basis of its own best possible weights. In words, each school is put in its most favourable light, and any other weighting scheme would generate a lower score. Of course, the model does not resolve the fact that ranking of alternatives from A is heavily dependent on the considered weights $w_1 ... w_n$.

3.3 Considering large set of random weights-SMAA

In the methodological framework of composite indices, the question of uncertainness in weighting process was introduced by Lahdelma et al. (1998) and Lahdelma, Salminen (2001) with the SMAA. This methodology has been recently used in economics literature by Greco et al. (2018) and Lagravinese et al. (2019). Unknown preferences on the weights assigned to each dimension, are considered by the probability distributions $f_W(W)$ in the set of the feasible weights W defined as:

$$W = \left[(w_1 ... w_n) \ \in R_+^n, \qquad w_1 + ... + w_n \ = 1\right] \tag{5}$$

Lack of knowledge about weights is represented by a uniform weight distribution in the set of feasible weights W. To rank schools according to the composite index of educational attainments, the rank is defined as an integer from 1 to m (the number of schools). From the probability distributions $f\ \chi(\xi)$ on χ, where χ is the evaluation space (i.e. the values assumed by the the plausible values $g_i \in G$) Lahdelma and Salminen (2001) introduce a ranking function relative to the school a_k:

$$rank\ (k, \xi, w) = 1 + \sum_{h \neq k} \rho \left[CI(\xi_h, w) > CI(\xi_k, w)\right] \tag{6}$$

where ρ *(true)*=1, and ρ $(false)$=0. In words, the rank of school a_k, given a vector of weights w, is one plus how many times the weighted average of attainments of a_k is dominated by the weighted average of attainments of the other schools. Thus, the value assumed by the variable "rank" in equation (6) is one plus the number of schools that performs better than school a_k in terms of average attainments. It follows that the higher the value of $rank\ (k, \xi, w)$ the lower the performance of the school a_k

For each school a_k and for each value that can be taken by educational attainments $\xi \in \chi$, SMAA computes the set of weights for which school a_k assumes rank r:

$$W_k^r = (\xi) = [w \in W : rank\ (k, \xi, w) = r] \tag{7}$$

From equation (6), the rank acceptability index can be estimated as follows:

$$b_k^r = \int_{\xi \in \chi} f\chi(\xi) \int_{w \in W_k^r(\xi)} fw(w)\ dwd\xi \tag{8}$$

b_k^r gives the probability that the school a_k has the r-th position in the ranking. b_k^r is the ratio of the number of the vector of weights by which school a_k gets rank r to the total amount of feasible weights. Computationally, the multidimensional integrals are estimated by using Monte Carlo simulations. Our estimates are the result of 10,000 random extractions of vectors w from a uniform distribution in W. To this regard, Tervonen and Ladhelma (2007) have shown that 10,000 extractions allow to get an error limit of 0.01 with a confidence interval of 95%.

4 Results

In our analysis, we rank for each year the different schools in terms of the composite attainments. In what follows we first show country level average performances using using BoD and than we explore all the feasible vectors of weights using SMAA. In SMAA for each school an higher value of the rank implies a lower multidimensional education outcome. We present the aggregate results by means of cumulate rank acceptability indices. The focus here will be on two aspects: the country-level performance of schools using BoD (Section 4.1); the country-level distribution of rank acceptability indices using SMAA (Section 4.2).

4.1 The country-level performance of schools

As in the standard DEA model, using BoD presented in Section 3.1 a school can obtain an index between [0:1], the higher the index the higher the effectiveness. In columns 2, 3 and 4 of Table 2 we report the average of BoD estimated at school level for the 34 Countries in our sample in the interval 2006-2015. Focusing on 2006, higher average performances (more than 0.7) are found in Korea, Finland and Poland. On the contrary, Mexico, Chile, Israel, Turkey and Greece show lower average performances (less than 0.6). In 2015 Japan, Estonia, and Finland get the first three ranking and Mexico, Turkey, Chile, Greece, and Slovak Republic

are on the bottom of ranking. On the 2016-2015 interval, higher increases in the performances have involved Japan (0.128), Estonia (0.12), and Israel (0.119). Lower increases in the same interval can be found in the average performance of Turkey, Korea, and Czech Republic (0.032, 0.052, and 0.055 respectively). The distance between top and bottom performer countries is slightly increasing over time: 0.167 is the difference between Korea and Mexico in 2006, 0.168 is the difference between Korea and Mexico in 2009, 0.159 is the difference between Korea and Mexico in 2012, and 0.181 is the difference between Japan and Mexico in 2015.

These results are in line with previous studies on PISA dataset (Lagravinese et al. 2020), although in this study a different methodology (BoD vs. Conditional Slack Based Measure), a different time observation (2009-2015 vs. 2009-2012), and different unit of analysis (schools vs. students) are considered. The higher performances in schools located in Northern European systems have been explained by the relative greater homogeneity of social and cultural conditions (Esping-Andersen, Wagner, 2012). Good performances in less developed countries have been partially explained by percentages of "resilient" students and schools, i.e. DMUs from a disadvantaged socio-economic background who achieve relatively high levels of performance in terms of education (Agasisti and Longobardi, 2014; Lagravinese et al. 2020). The low performances in some South American countries have been found to be associated with institutional factors and inequality in different domains that are important in explaining the under-performances at school level (Chetty et al.2016; Raitano and Vona, 2016; Lagravinese et al. 2020).

Table 2: BoD average shools, average of indices obtained with weights from best school in Country, average of indices obtained with weights from best school overall, average of indices obtained with weights from worst school in Country, average of indices obtained with weights from worst school overall

	BoD Average shools by country				Weights Best School in Country				Weights Best School Overall				Weights Worst School in Country				Weights Worst School Overall			
Country	2006	2009	2012	2015	2006	2009	2012	2015	2006	2009	2012	2015	2006	2009	2012	2015	2006	2009	2012	2015
AUS	0.67	0.72	0.69	0.75	0.65	0.65	0.63	0.62	0.65	0.65	0.63	0.62	0.65	0.65	0.63	0.62	0.64	0.64	0.63	0.62
AUT	0.65	0.67	0.66	0.73	0.62	0.61	0.62	0.61	0.62	0.61	0.62	0.61	0.62	0.59	0.62	0.62	0.61	0.58	0.60	0.59
BEL	0.67	0.72	0.69	0.75	0.65	0.65	0.65	0.64	0.66	0.65	0.65	0.63	0.65	0.65	0.65	0.64	0.63	0.61	0.61	0.61
CAN	0.68	0.73	0.70	0.77	0.65	0.65	0.65	0.64	0.65	0.65	0.65	0.64	0.64	0.63	0.63	0.63	0.64	0.63	0.63	0.63
CHE	0.67	0.73	0.68	0.76	0.66	0.66	0.65	0.65	0.66	0.66	0.65	0.65	0.66	0.66	0.65	0.65	0.61	0.62	0.61	0.61
CHL	0.57	0.61	0.61	0.68	0.51	0.52	0.55	0.55	0.51	0.52	0.55	0.55	0.51	0.52	0.55	0.55	0.52	0.53	0.56	0.56
CZE	0.68	0.71	0.69	0.74	0.65	0.63	0.65	0.63	0.67	0.64	0.64	0.62	0.67	0.64	0.64	0.62	0.65	0.62	0.63	0.60
DEU	0.65	0.71	0.69	0.76	0.63	0.64	0.65	0.66	0.63	0.64	0.65	0.64	0.63	0.64	0.65	0.64	0.63	0.63	0.64	0.62
DNK	0.67	0.70	0.66	0.74	0.65	0.63	0.63	0.64	0.65	0.62	0.62	0.64	0.65	0.62	0.62	0.64	0.61	0.59	0.59	0.60
ESP	0.65	0.70	0.67	0.74	0.63	0.63	0.64	0.65	0.64	0.63	0.63	0.62	0.63	0.63	0.64	0.65	0.62	0.60	0.62	0.61
EST	0.67	0.72	0.71	0.79	0.65	0.65	0.67	0.67	0.65	0.65	0.66	0.66	0.65	0.65	0.67	0.67	0.65	0.65	0.66	0.65
FIN	0.72	0.77	0.70	0.79	0.70	0.69	0.65	0.65	0.70	0.69	0.65	0.65	0.71	0.70	0.67	0.69	0.69	0.68	0.65	0.65
FRA	0.65	0.71	0.68	0.74	0.63	0.63	0.62	0.62	0.63	0.63	0.62	0.62	0.63	0.63	0.62	0.62	0.60	0.61	0.61	0.60
GBR	0.65	0.70	0.68	0.75	0.63	0.62	0.62	0.62	0.63	0.62	0.62	0.62	0.64	0.64	0.65	0.65	0.63	0.62	0.62	0.62
GRC	0.60	0.68	0.63	0.68	0.56	0.59	0.56	0.57	0.56	0.59	0.56	0.57	0.56	0.59	0.56	0.57	0.56	0.57	0.56	0.55
HUN	0.62	0.68	0.65	0.70	0.60	0.62	0.62	0.59	0.60	0.60	0.59	0.59	0.60	0.60	0.59	0.59	0.59	0.59	0.59	0.57
IRL	0.68	0.71	0.71	0.76	0.68	0.65	0.68	0.68	0.64	0.62	0.64	0.64	0.62	0.62	0.64	0.61	0.62	0.62	0.64	0.61
ISL	0.66	0.72	0.66	0.72	0.65	0.64	0.62	0.62	0.65	0.64	0.62	0.62	0.65	0.64	0.62	0.62	0.62	0.61	0.59	0.58
ISR	0.59	0.67	0.66	0.71	0.56	0.57	0.59	0.59	0.56	0.57	0.59	0.59	0.56	0.57	0.59	0.59	0.56	0.56	0.58	0.57
ITA	0.62	0.69	0.66	0.73	0.61	0.63	0.62	0.63	0.59	0.61	0.61	0.62	0.61	0.63	0.62	0.63	0.58	0.60	0.59	0.59
JPN	0.67	0.75	0.73	0.80	0.66	0.67	0.68	0.68	0.66	0.67	0.68	0.68	0.65	0.66	0.67	0.66	0.65	0.66	0.67	0.66
KOR	0.73	0.77	0.73	0.78	0.72	0.70	0.70	0.67	0.69	0.69	0.70	0.66	0.72	0.70	0.70	0.67	0.64	0.66	0.66	0.63
LUX	0.64	0.69	0.66	0.73	0.63	0.62	0.64	0.63	0.63	0.62	0.62	0.62	0.60	0.60	0.60	0.60	0.60	0.60	0.60	0.60
LVA	0.64	0.69	0.67	0.72	0.63	0.63	0.63	0.63	0.62	0.61	0.62	0.61	0.63	0.63	0.63	0.63	0.60	0.60	0.61	0.59
MEX	0.56	0.60	0.57	0.62	0.55	0.55	0.55	0.55	0.53	0.53	0.52	0.51	0.55	0.55	0.55	0.55	0.51	0.51	0.51	0.51
NLD	0.69	0.75	0.69	0.77	0.67	0.67	0.66	0.66	0.68	0.68	0.66	0.65	0.67	0.67	0.66	0.66	0.65	0.65	0.63	0.62
NOR	0.65	0.72	0.69	0.76	0.63	0.64	0.63	0.64	0.63	0.64	0.63	0.64	0.64	0.66	0.66	0.67	0.60	0.61	0.61	0.61
NZL	0.69	0.75	0.70	0.76	0.68	0.68	0.67	0.66	0.66	0.66	0.63	0.62	0.66	0.66	0.63	0.62	0.65	0.65	0.63	0.62
POL	0.70	0.72	0.72	0.77	0.70	0.66	0.69	0.67	0.66	0.64	0.67	0.65	0.64	0.63	0.66	0.62	0.64	0.63	0.66	0.62
PRT	0.62	0.69	0.66	0.72	0.62	0.63	0.63	0.63	0.59	0.61	0.61	0.60	0.62	0.63	0.63	0.63	0.58	0.60	0.59	0.59
SVK	0.63	0.69	0.63	0.69	0.59	0.59	0.57	0.55	0.62	0.62	0.60	0.59	0.60	0.61	0.59	0.58	0.59	0.59	0.57	0.55
SVN	0.62	0.65	0.62	0.72	0.59	0.57	0.57	0.61	0.59	0.59	0.59	0.61	0.59	0.59	0.59	0.61	0.58	0.58	0.58	0.58
SWE	0.68	0.72	0.67	0.75	0.67	0.66	0.64	0.66	0.64	0.64	0.62	0.64	0.64	0.64	0.62	0.64	0.62	0.62	0.60	0.61
TUR	0.59	0.64	0.63	0.62	0.59	0.59	0.60	0.54	0.54	0.55	0.56	0.52	0.54	0.55	0.56	0.52	0.52	0.55	0.56	0.50
Mean	0.65	0.70	0.67	0.74	0.63	0.63	0.63	0.63	0.63	0.62	0.62	0.62	0.63	0.63	0.63	0.62	0.61	0.61	0.61	0.60

Source: Authors' elaboration on data from OECD (2017a)

What cannot be explored with BoD analysis as well as with any exercise using one vector of weights for DMU, is to what extend the results are due to the effect of weights. A relevant question in this context is if the ranking presented in columns 2, 3, 4, and 5 of Table 2 depends on the considered weights or it is robust changing assumption on weights.

It is worth noting that, since the numbers in columns 2, 3, 4, and 5 of Table 2 are the averages of school performances evaluated by applying to each school the best set of weights (with BoD), they represent the maximum index (in the interval [0:1]) the country can achieve given its schools' performances in mathematics, reading, and sciences. As a first representation of how assumptions on weights could influence the ranking, in the columns from 6 to 21 in Table 2 we report average indices at country level considering the following four cases: 1. We take the weights applied by BoD to the school with the highest index in each country, and apply the same set of weights to all other schools in the country (results in columns 6, 7, 8, and 9); 2. We take the weights applied by BoD to the school with the highest index overall, and apply the same set of weights to all other schools (results in columns 10, 11, 12, and 13); 3. We take the weights applied by BoD to the school with the lowest index in each country, and apply the same set of weights to all other schools in the country (results in columns 14, 15, 16, and 17); 4. We take the weights applied by BoD to the school with the lowest index overall, and apply them to all other schools (results in columns 18, 19, 20, and 21).

Considering the first case (evaluation with weights from the best school in the Country) the correlation with average indices obtained by BoD are between 0.92 in 2015 and 0.97 in 2006. The country-level differences with BoD can be interpreted as an approximation of how all schools diverge from the best performer, i.e. how different have to be the weights assigned to other schools in each country compared with the best performer in order to get the maximum index. Altought by construction all indices are lower than BoD, in some countries the differences between average index obtained with weights from the best school and average index obtained by BoD are more pronounced: Chile, Slovakia, Israel, and Greece loose more than 11 per cent (-14 per cent, -13 per cent, -12 per cent, and -12 per cent respectively), while Ireland is the only country loosing less than 6 per cent of the index obtained by BoD. Considering the evaluations obtained with weights from the best school overall (columns 10, 11, 12, and 13 in Table 2) the correlations with the average indices obtained by BoD range from 0.94 in 2012 to 0.98 in 2009. Chile, Turkey, Israel, and Greece have the highest difference between average index obtained by BoD and average index obtained by weights from the best school overall (-14 per cent, -13 per cent, -12 per cent, and -12 per cent respectively). On the contrary the lowest difference is in Switzerland (-7 per cent). Considering the indices obtained with weights from the worst school in the Country (columns 14, 15, 16, and 17 in Table 2), a lower correlation with BoD can be observed: the minimum is 0.90 in 2012 and the maximum is 0.95 in 2009. Countries loosing more than 11 per cent of index are seven: Chile (-14 per cent), Turkey and Ireland (-13 per cent), Poland, Israel, Canada and Luxembourg (-12 per cent). The lowest loss, -7 per cent, is observed in eight countries: Latvia, Korea, United Kingdom, Finland, Switzerland, Portugal, Mexico, and Norway. Finally, as for the average indices obtained by weights from the worst school overall, the correlations with BoD are between 0.94 in 2006 to 0.99 in 2015. The losses have a higher magnitude for Norway (-15 per cent compared with BoD), Estonia, Czech Republic, and Italy (-14 per cent). The minimum loss is the 10 per cent observed in Austria, Sweden, and Hungary.

Overall, the four different cases shown in table 2 highlight that there are systems which have education performances less sensitive to the weighting scheme, as it is the case of Ireland and Switzerland when the best schools are considered for selecting weights, and again Switzerland with some Scandinavian countries (mainly Finland and Sweden) when the weights from the worst school are considered for making evaluation. Such results evidence that these systems have homogeneous school's specialisation in the considered dimensions since changing the weights assigned to mathematics, reading, and science does not affect much the overall performance compared with BoD. In other words, in these countries the performance of schools are uniform in terms of mix (i.e. regardless the overall quality of school, the relative performance in the three subjects considered is uniform across all schools). On the contrary, the four different cases shown in table 2 also evidence that many systems have performances that widely depends on the heterogeneity of weights considered for making the evaluation. It is the case of Chile, Israel, and Greece when the weights from the best schools are considered, and again Chile, with Turkey, Estonia, Czech Republic, and Italy when the weights from the worst schools are used to estimate the performances of the remaining schools. The dependence of their performances from weights is an evidence of a heterogeneous performance within school. This means that the systems are heterogeneous at school level in terms of mix (i.e. in terms of relative performance in mathematics, reading, and science). Regardless the overall performance, the education system in these countries has school that specialise in differentiated subjects.

To what extent systems with schools specialised in similar subjects should be preferred to systems with differentiated schools specialisations is out of the scope of this study. Of course a uniform schools specialisation that strategically match the countrys specialisation will produce a more productive human capital. At the same time a differentiated school system would give more opportunity to students to specialise on what they want or are more talented at and would be more resilient to changes in the production systems. More broadly, table 2 clearly shows that small changes in weighting scheme result in wide differentiation in the country level mean indices, which in turn are often used for making performance evaluation in the education sector both in the scientific literature and for policy making worldwide (see De Witte, López-Torres, 2017 for a review). To over-come the shortcoming of referring to a unique set of weights into the evaluation, in the next Sections we present results of analysis using SMAA which explicitly considers differentiation in weights considering the whole space of positive feasible weights.

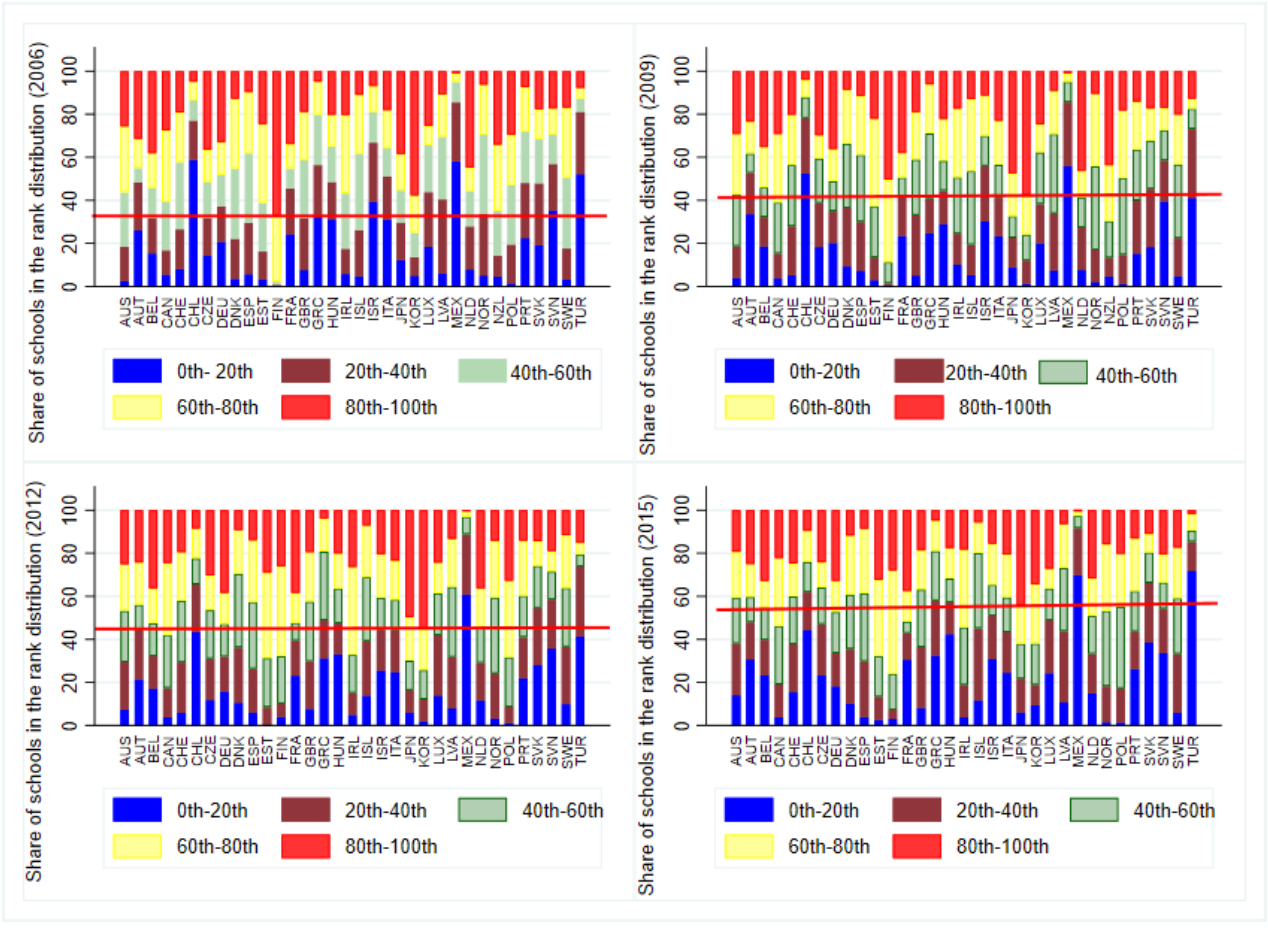
4.2 The country-level distribution of rank acceptability indices

The main outcome of SMAA, is a matrix with the school-level rank acceptability index for any rank and for each wave. The country-level rank acceptability index is given by the average of school-level rank acceptability indices. Taken a specific rank (we take the 20-th and the 80-th percentile of ranking), the country-level downward and the upward cumulative rank acceptability indices at country level can be interpreted as the share of good performer and bad performer schools respectively.

In order to analyse the distribution over time of the 'excellent' and 'shoddy' schools, we divide in 5 percentiles the rank distribution of the schools analysed in the PISA sample. In Figure 1 the red line represents for each country, the share of schools that falls in the 80th rank percentile. Interestingly, more than 60% of Finland schools was in this category in 2006, and

there has been a persistent reduction of this share over the years (see Chung 2015, for more detail on changes in the Finnish education system). A similar pattern can be seen in South Korea, which in the first two waves (2006 and 2009) recorded a high number of schools in the highest percentile. In these countries the share of excellent schools has drastically reduced by over 20 points, highlighting a significant reduction of the schools positioned in the top ranking position. This may signal the possibility that excellent schools are now more equally distributed among countries.

Figure 1: Share of Schools in the rank distribution (PISA 2006-2009-2012-2015)



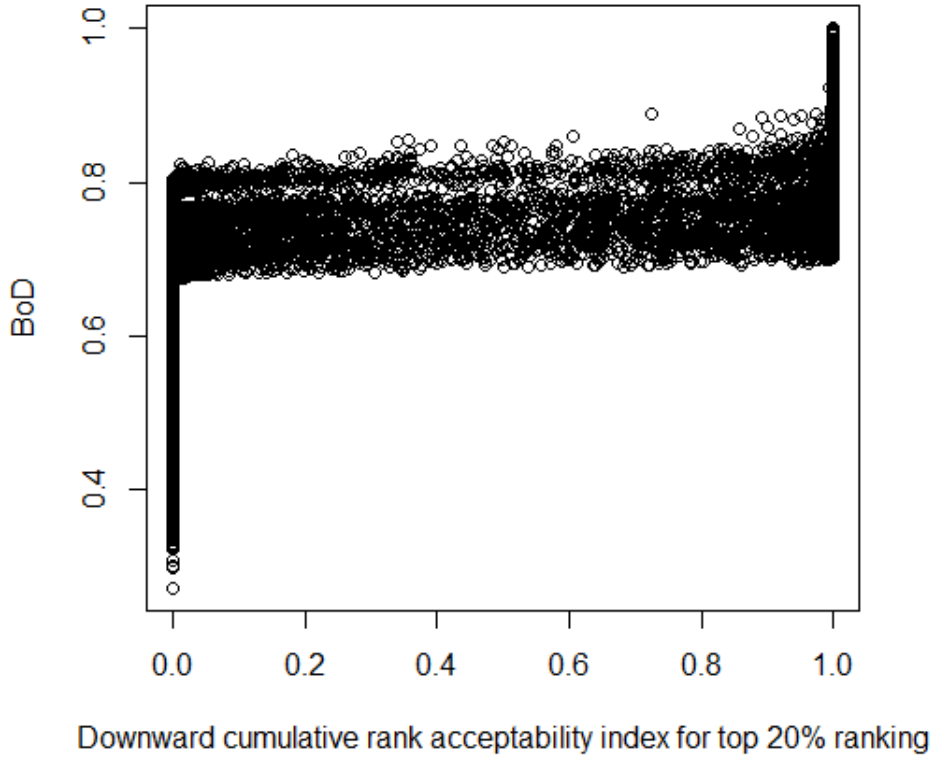
Authors elaboration on OECD (2017a)

In the 2006 – 2015 interval, our analysis shows that that Central-European countries (Belgium and Netherlands in particular) tend to be constantly good performers with high shares of excellent schools and low shares of shoddy schools. On average European countries (Finland in particular) lose a relevant share of 'excellent' schools in favour of Japan and Korea. Constant

bad performer countries are Mexico and Turkey, with really high share of 'shoddy' schools and high stratification. Indeed, the overlapping matrix shows that their school-level probabilities of being in the lowest ranks are not shared by many other countries. Constant high polarisation, with both high share of excellent and high share of shoddy schools is in some European Countries like Austria, France, Czech Republic, and Belgium.

Comparing these evidences with the results in Section 4.1, some relevant differences can be observed. The correlation between the school-level probability to fall in the 80-th rank percentile with the school-level BoD index is 0.69 in 2006, 0.691 in 2009, 0.682 in 2012, and 0.693 in 2015. This reveals that not all schools that perform well in BoD have the same results changing the weighting system. These evidences can be graphically explored in Figure 2 in which the downward cumulative downward cumulative rank acceptability index for top 20% ranking (probability to fall in 20-th percentile of ranking) is on the x axis and BoD index is on the y axis. It can be noted that some schools fall 100% of times in the 20-th percentile of ranking, and at the same time they have less than 0.7 index with BoD. Some other schools with the same or higher index on BoD have 0% probability to get the 20-th percentile of ranking. These school get higher (lower) index in BoD because of a higher (lower) performance in a specific dimension (math, language, and science), but when the whole set of feasible weight are explored by SMAA, their weakness in other dimensions show up.

Figure 2: BoD and downward cumulative rank acceptability index for top 20 per cent of ranking (PISA 2006-2009-2012-2015)



Authors elaboration on OECD (2017a)

5 The multidimensional inequality

A better way to look in the distribution of performances is by means of inequality measures. In the SMAA context, rank acceptability index b_k^r can be used to estimate multidimensional generalisation of the Gini index (Greco et al. 2018). These estimates can be obtained by first defining the upward cumulative rank acceptability index of rank l, i.e., the probability that the school a_k has a rank l or higher (Angilella et al. 2016). Formally:

$$b_k^{\geqslant l} = \sum_{s=l}^{m} b_k^s \tag{9}$$

Given $b_k^{\geqslant l}$, the Gini index of the upward cumulative rank acceptability index of rank l is (Greco et al., 2018):

$$G^{\geqslant l} = \frac{\sum_{h=1}^{m}\sum_{k=1}^{m}|b_h^{\geqslant l} - b_k^{\geqslant l}|}{2ml} \tag{10}$$

$G^{\geqslant l}$ measures how the probabilities of attaining rank l or higher are concentrated. For each l, the higher is $G^{\geqslant l}$ the more concentrated is the probability to be above this rank in terms of the composite index of educational attainments. In other words, $G^{\geqslant l}$ measures the dispersion of the probability that each school may have in occupying rank l or higher. An equal probability for all schools gives $G^{\geqslant l} = 0$, while a high level of $G^{\geqslant l}$ signals that this probability is heavily concentrated in few schools, and reveals great differences in the education outcome.

In the same way, the downward cumulative rank acceptability index of position l for school a_k is:

$$b_k^{\leqslant l} = \sum_{s=1}^{l} b_k^s \tag{11}$$

and the Gini index of the probability to attain rank l or lower is as follows:

$$G^{\leqslant l} = \frac{\sum_{h=1}^{m}\sum_{k=1}^{m}|b_h^{\leqslant l} - b_k^{\leqslant l}|}{2m(m-l+1)} \tag{12}$$

For each l the higher is $G^{\leqslant l}$ the more concentrated is the probability to be below this rank in terms of the composite index of attainments. As mentioned in Greco et al. (2018), $G^{\geqslant l}$ and $G^{\leqslant l}$ are generalization of the Gini because they allow to consider multidimensionality and all the possible vectors of weights, differently from previous proposals (Savaglio 2006; Weymark 2006).

The final aim of our analysis is to analyse education inequality not only among schools, but also among countries. To this aim, we use the ANOGI (Yitzhaki, 1994), as developed in Liberati (2015), and generalised in Lagravinese et al. (2019) to the decomposition of $G^{\geqslant l}$ and $G^{\leqslant l}$. The following decomposition will be used for the case of the Gini index of the upward cumulative rank acceptability index:

$$G^{\geqslant l} = \sum_i \underbrace{s_i G^{\geqslant l} p_i}_{\text{Standard WI}} + \underbrace{\sum_i s_i G^{\geqslant l} \sum_{j\neq i} p_i O_{ji}^{\geqslant l}}_{\text{Impact of overlapping on WI}} + \underbrace{G_{Bp}^{\geqslant l}}_{\text{Standard BI}} + \underbrace{(G_B^{\geqslant l} - G_{Bp}^{\geqslant l})}_{\text{Impact of overlapping on BI}} \tag{13}$$

The first term is the within-country inequality (WI) in the absence of overlapping, where s_i is the probability of schools within country i to be in rank l or higher and p_i is the share of population of country i. The second term is the impact of overlapping on within inequality, driven by the contribution of the overlapping index of each country with all other countries weighted by their population shares. The last two terms of equation (13) deal with the between-country inequality (BI). The term $G_{Bp}^{\geqslant l} = \frac{2cov(\bar{b}_i\ \bar{F}_i(b))}{\bar{b}}$ is based on the between inequality as

originally defined by Pyatt (1976), where the covariance is between the mean probability of each country $\bar{b}_i$ and its rank in the distribution of the mean probabilities of all countries $\bar{F}_i(b)$. This definition would imply that $G_{Bp}^{\geqslant l} = 0$ when all the mean probabilities are equal.

According to Yitzhaki and Lerman (1991), instead, one can alternatively define $G_B^{\geqslant l}$ as twice the covariance between the mean $\bar{b}_i$ of countries and the countries' mean ranks all schools, divided by overall expected rank acceptability index. The difference between the two definitions is in the rank that is used to represent the group (country): under Pyatt's approach it is the rank of the country-level mean $\bar{b}_i$ while under Yitzhaki-Lerman it is the mean rank of all schools belonging to the country. In this case, $G_{Bp}^{\geqslant l} = 0$ implies that the average rank of all countries in the OECD distribution would be equal.

These two approaches yield the same ranking if complete stratification occurs, $G_B^{\geqslant l} = G_{Bp}^{\geqslant l}$. This implies that in the absence of overlapping of probabilities, between-inequality would be uniquely defined by $G_{Bp}^{\geqslant l}$. With overlapping, instead, $G_B^{\geqslant l} - G_{Bp}^{\geqslant l} < 0$, which can be used as an indicator of the reduction in between inequality caused by the overlapping of probabilities.

With the same rationale, the downward cumulative Gini coefficient can be expressed as:

$$G^{\leqslant l} = \sum_i \underbrace{s_i G^{\leqslant l} p_i}_{\text{Standard WI}} + \underbrace{\sum_i s_i G^{\leqslant l} \sum_{j \neq i} p_i O_{ji}^{\leqslant l}}_{\text{Impact of overlapping on WI}} + \underbrace{G_{Bp}^{\leqslant l}}_{\text{Standard BI}} + \underbrace{(G_B^{\leqslant l} - G_{Bp}^{\leqslant l})}_{\text{Impact of overlapping on BI}} \tag{14}$$

with elements having the same meaning as in (13), but with respect to the probabilities of having rank l or below.

Among the advantages of Gini index compared with other decomposable measures of inequality such as the Theil (1967) index, one of the reasons for using Gini coefficients in SMAA context is computational. Differently from Theil index, in Gini does not matter that some values may be zero as it is the case of upward and downward cumulative rank acceptability indices defined in (9) and (11).

In Table 3 we show the ANOGI for the downward cumulative rank acceptability index for the top 20% of the ranking. Total inequality shown in the second column is quite constant over time (moves from 0.797 to 0.798 in the 2006-2015 interval). This means that, ignoring the countries of origin, overall school-level distribution tends to be constant if we consider the probability of being among the top 20%. Looking at the Gini components, we observe that the standard 'within' inequality without overlapping (third column in Table 3) moves from 0.029 to 0.03 in the 2016-2015 period. This component represents the 3.86 per cent of total inequality in average. The larger component of total inequality is the impact of overlapping on within inequality, representing 84.97 per cent of the total inequality observed among schools. Furthermore, this component suffered an increase of 8.27% in the 2006-2015 interval. This means that the school distributions of probability to be on the top 20% become more intertwined over time. In other words, some schools tend to converge to excellence beyond the national borders over time. An opposite trend can be observed in the fifth column where the between component of inequality is presented. This component decreases from 0.372 to 0.272 in 2006-2015. In line with the standard 'between' inequality, the impact of overlapping on between

inequality is also decreasing from -0.261 to -0.215. As robustness check, we find that these results are confirmed using weights from a uniform distribution around with mean 1/3 in W (see Table A1 in the Appendix).

Table 3: Multidimensional ANOGI of Downward cumulative rank acceptability index for the top 20 per cent of ranking

Year	Tot. Ineq.	Standard WI	Impact of overl. on WI	Standard BI	Impact of overl. on BI
2006	0.797	0.029	0.657	0.372	-0.261
2009	0.797	0.032	0.661	0.367	-0.263
2012	0.798	0.032	0.681	0.324	-0.24
2015	0.798	0.030	0.711	0.272	-0.215

Authors elaboration on OECD (2017a)

The ANOGI decomposition allows to explore the stratification of the country level performances by means of the overlapping matrix. In Table 4 we show the average Overlapping of downward cumulative rank acceptability index for the top 20% of the ranking by country. It is worth recalling that, if no school in country j lies in the range of the distribution of probabilities of schools in i, country i could be defined a perfect stratum and $O_{\bar{j}i}^{\leq 20\%} = 0$. It follows that the higher the values in Table 4, the lower is the stratification of the country. Regarding the downward cumulative rank acceptability index, in 2006 highly stratified countries are Finland, Korea, Spain, and Ireland. On the contrary, Netherlands, Germany, and Hungary are countries with lower levels of stratification. In 2015, highly stratified countries are Mexico, Slovak Republic, and Chile, while lower level of stratification can be found in Poland, Ireland, and Finland. So overtime, a massive decrease of stratification involves Finland, Korea, and Ireland, while Turkey, the Slovak Republic, and Mexico had an increase in their stratification. Finland becomes less stratified because of a significant decrease in the share of excellent schools.

In Table 5 we present the multidimensional ANOGI of upward cumulative rank acceptability index for the 80% of ranking. Overall, the inequality of school-level probabilities of being among the bottom 80% is quite constant around 0.798 in 2006-2015 (second column). Considering the Gini components, the standard 'within' inequality decreases from 0.035 to 0.024 (third column). Also in this case, the bulk of the total inequality is the impact of overlapping on within inequality, representing the 71.21 per cent of global inequality among schools in average and increasing from .551 to .617 in 2006-2015. Standard between inequality decreases from 0.499 to 0.440 while the impact of overlapping on 'between' inequality tends to be almost constant in the same interval. A robustness check using weights from a normal distribution around with mean 1/3 in W (see Table A2 in the Appendix) confirmed the main evidences provided here.

In Table 6 we show the average Overlapping of upward cumulative rank acceptability index for the bottom 80% the ranking by country. As before, the higher the values in Table 6, the lower is the stratification of the country. Regarding the upward cumulative rank acceptability index, there is a cell with missing values in Table 6. It represents the case in which all schools of the baseline country have the same probability of being in the bottom 80% of ranking. This happens in Finland in 2006, because all of their schools have zero probability of being in the

Table 4: Average Overlapping of downward cumulative rank acceptability index for top 20 per cent of ranking

	2006	2009	2012	2015
AUS	0.997	0.989	0.976	0.951
AUT	0.998	1.016	1.017	0.966
BEL	1.010	1.045	1.022	1.019
CAN	0.933	0.934	0.928	0.993
CHE	0.985	0.930	0.965	0.922
CHL	0.939	0.983	0.983	0.910
CZE	1.036	1.051	1.044	1.004
DEU	1.090	1.074	1.060	1.000
DNK	0.937	0.962	0.992	0.957
ESP	0.880	0.960	0.956	0.953
EST	0.940	0.911	0.871	0.960
FIN	0.762	0.820	0.873	1.077
FRA	1.028	1.018	0.984	1.014
GBR	1.017	1.011	1.008	0.941
GRC	0.962	0.897	0.918	1.003
HUN	1.076	1.061	1.036	1.017
IRL	0.894	0.875	0.860	1.055
ISL	0.971	1.035	0.932	0.965
ISR	0.980	0.949	0.949	1.009
ITA	0.965	0.960	0.983	0.971
JPN	0.987	0.994	1.027	0.983
KOR	0.823	0.961	0.952	0.998
LUX	1.069	0.983	1.049	1.009
LVA	0.970	1.024	0.976	0.988
MEX	0.984	1.017	0.997	0.831
NLD	1.077	1.062	1.101	1.011
NOR	0.980	0.949	0.917	0.930
NZL	0.911	0.911	0.950	1.000
POL	0.988	1.015	0.990	1.050
PRT	0.978	1.004	0.995	0.985
SVK	1.037	0.955	1.056	0.906
SVN	1.030	1.036	1.021	0.946
SWE	0.897	1.006	0.974	0.970
TUR	1.053	1.057	1.083	0.961

Authors elaboration on OECD (2017a)

Table 5: Multidimensional ANOGI of Downward cumulative rank acceptability index for the bottom 20 per cent of ranking

Year	Tot. Ineq.	Standard WI	Impact of overl. on WI	Standard BI	Impact of overl. on BI
2006	0.798	0.035	0.551	0.499	-0.286
2009	0.798	0.043	0.551	0.492	-0.288
2012	0.798	0.039	0.555	0.487	-0.282
2015	0.799	0.024	0.617	0.44	-0.282

Authors elaboration on OECD (2017a)

bottom 80% of ranking. In the other cases we observe really high stratification in Chile, Turkey, and Mexico over the whole period, which means that their school probabilities of being in the lowest rank are not shared by many other countries. In 2006 lower level of stratification are in Sweden, Iceland, and Poland. In 2015, Poland, Ireland, and Finland are the less stratified countries.

Table 6: Average Overlapping of upward cumulative rank acceptability index for bottom 20 per cent of ranking

	2006	2009	2012	2015
AUS	1.057	1.023	0.979	0.951
AUT	0.947	0.927	0.936	0.966
BEL	1.007	1.026	1.004	1.019
CAN	1.012	1.011	0.982	0.993
CHE	0.968	0.960	0.969	0.922
CHL	0.852	0.821	0.974	0.910
CZE	0.974	0.975	0.999	1.004
DEU	1.017	0.976	0.993	1.000
DNK	1.012	0.976	0.972	0.957
ESP	0.953	0.984	0.998	0.953
EST	0.954	0.920	0.870	0.960
FIN	n.a*	1.098	1.066	1.077
FRA	1.000	0.994	1.024	1.014
GBR	0.996	0.978	0.989	0.941
GRC	0.994	1.007	1.055	1.003
HUN	1.003	1.001	1.015	1.017
IRL	0.989	0.966	1.010	1.055
ISL	1.054	1.027	0.977	0.965
ISR	0.897	0.943	0.970	1.009
ITA	0.941	0.975	0.999	0.971
JPN	0.958	0.982	0.996	0.983
KOR	1.000	1.035	1.047	0.998
LUX	0.993	1.058	0.901	1.009
LVA	0.936	0.988	0.971	0.988
MEX	0.878	0.862	0.844	0.831
NLD	1.011	0.979	1.013	1.011
NOR	0.921	0.992	0.993	0.930
NZL	0.991	1.056	1.014	1.000
POL	0.981	0.981	1.079	1.050
PRT	1.002	0.977	1.000	0.985
SVK	0.950	0.956	0.961	0.906
SVN	0.948	0.952	0.935	0.946
SWE	1.039	1.045	1.001	0.970
TUR	0.829	0.894	0.836	0.961

Authors elaboration on OECD (2017a); * Represents the cases in which all schools of the baseline country have the same probability of being in the bottom 20 per cent of ranking

6 Concluding remarks

This paper investigated the evolution of education quality at school level in the OECD, during the time interval 2008-2015, out of the PISA multidimensional database. From a methodological standpoint we employ two non-parametric procedures to deal with weighting of the school's achievements in mathematics, reading, and science: the Benefit of Doubt (BoD) and the Stochastic Multicriteria Acceptability Analysis (SMAA). In line with previous studies on PISA, the results of the BoD evaluations show that schools located in Northern European systems have higher performances, which in the literature have been explained by the relative greater homogeneity of social and cultural conditions. Furthermore, results of BoD evidence low performances in some South American, which in the literature have been found to be associated with institutional factors and inequality in different domains.

As a first representation of how assumptions on weights could influence the ranking, we extend the BoD results creating four different scenarios in which weights from the best and the worst school within country are employed to evaluate all the remaining schools. The different cases highlight that there are systems which have education performances less sensitive to the heterogeneity of the weighting scheme across schools, as Ireland, Switzerland, Finland, and Sweden. Since changing the weights assigned to mathematics, reading, and science does not affect much the overall performance compared with BoD, in these countries the performance of schools are uniform in terms of mix: i.e. regardless of the overall quality of school, the relative performance in the three subjects considered is more uniform across all schools. On the contrary, many systems have performances that widely depend on the weights considered for making the evaluation. It is the case of Chile, Israel, Greece, Turkey, Estonia, Czech Republic, and Italy. The dependence of their performances from weights is an evidence of a heterogeneous performance in different subjects within school. Regardless the overall performance, the education system in these countries has schools that specialise in differentiated subjects. Overall, this is an evidence of the shortcomings of referring to a unique set of weights into the evaluation, and opens the way for the analysis with SMAA which explicitly takes into account differentiation in weights considering the whole space of positive feasible weights. Infact they prove that, although nominally data driven, the BoD assumption is infact based on a specific prior judgment of what is best for the education system, specialization at school level.

The SMAA evidence that in the time interval 2009-2015 that excellent schools become more equally distributed among countries. These evidences are confirmed by the multidimensional ANOGI which show that in the four different waves considered (2006, 2009, 2012, and 2015) there has been a convergence path between countries. However, as a result, inequality within countries (among schools) has increased substantially. This suggests that education inequality has followed a pattern similar to overall inequality, at least among relatively advanced countries. Our findings suggest that inequality at national level is a worrying phenomenon. It suggests increasing segregation at school level, leaving a large section of the population unable to face effectively the challenges of globalisation. It also suggests that policy efforts in advanced countries should be directed primarily at decreasing such inequality. Public policies are needed to foster virtuous paths to reduce disparities among students with different socioeconomic background. Our results are in line with a recent studies on school segregation on PISA databases and consistent with the evidence of most electoral analyses that identify the educational divide

as the primary explanation for the voting patterns in countries that have experienced a populist backlash in recent years.

Public authorities should develop supportive learning environments through concerted efforts of investing more in marginalized communities.

References

[1] Agasisti, T., Longobardi, S. (2014). Inequality in education: Can Italian disadvantaged students close the gap?. Journal of Behavioral and Experimental Economics, 52, 8-20.

[2] Agasisti, T., Longobardi, S., Prete, V., & Russo, F. (2018). Multidimensional poverty measures for analysing educational poverty in European countries (No. 73).

[3] Agasisti, T., Munda, G., Hippe, R. (2019). Measuring the efficiency of European education systems by combining Data Envelopment Analysis and Multiple-Criteria Evaluation. Journal of Productivity Analysis, 1-20.

[4] Angilella, S., Bottero, M., Corrente, S., Ferretti, V., Greco, S., & Lami, I. M. (2016). Non Additive Robust Ordinal Regression for urban and territorial planning: an application for siting an urban waste landfill. Annals of Operations Research, 245(1-2), 427-456.

[5] Bailey, M., & Borooah, V. K. (2010). What enhances mathematical ability? A cross-country analysis based on test scores of 15-year-olds. Applied Economics, 42(29), 3723-3733.

[6] Bidwell, C. E., & Kasarda, J. D. (1975). School district organization and student achievement. American Sociological Review, 55-70.

[7] Bloom, N., Lemos, R., Sadun, R., & Van Reenen, J. (2015). Does management matter in schools?. The Economic Journal, 125(584), 647-674.

[8] Charnes, A., Cooper, W. W., Rhodes, E. (1978). Measuring the efficiency of decision making units. European journal of operational research, 2(6), 429-444.

[9] Checchi, D. (1998). Povertà ed istruzione: alcune riflessioni e una proposta di indicatori, Politica Economica, 14(2), 245-282.

[10] Checchi, D., & Peragine, V. (2010). Inequality of opportunity in Italy. The Journal of Economic Inequality, 8(4), 429-450.

[11] Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T., Saisana, M., Saltelli, A., Liska, R., Tarantola, S. (2008). Creating composite indicators with DEA and robustness analysis: the case of the Technology Achievement Index. Journal of the Operational Research Society 59(2), 239-251.

[12] Cherchye, L., Moesen, W., Rogge, N., Van Puyenbroeck, T.(2007). An introduction to benefit of the doubt composite indicators. Social indicators research, 82(1), 111-145.

[13] Chetty, R., Hendren, N., Katz, L. F. (2016). The effects of exposure to better neighborhoods on children: New evidence from the Moving to Opportunity experiment. American Economic Review, 106(4), 855-902.

[14] Chung, J. (2015). International comparison and educational policy learning: Looking north to Finland. Compare: A Journal of Comparative and International Education, 45(3), 475-479.

[15] Costanza, R., Daly, L., Fioramonti, L., Giovannini, E., Kubiszewski, I., Mortensen, L. F., Wilkinson, R. (2016). Modelling and measuring sustainable wellbeing in connection with the UN Sustainable Development Goals. Ecological Economics, 130, 350-355.

[16] De Witte, K., & Kortelainen, M. (2013). What explains the performance of students in a heterogeneous environment? Conditional efficiency estimation with continuous and discrete environmental variables. Applied Economics, 45(17), 2401-2412.

[17] De Witte, K., & López-Torres, L. (2017). Efficiency in education: a review of literature and a way forward. Journal of the Operational Research Society, 68(4), 339-363.

[18] De Witte, K., & Schiltz, F. (2018). Measuring and explaining organizational effectiveness of school districts: Evidence from a robust and conditional Benefit-of-the-Doubt approach. European Journal of Operational Research, 267(3), 1172-1181.

[19] Decancq, K. and Lugo, M. A. (2013). Weights in multidimensional indices of wellbeing: An Overview. Econometric Reviews, 32(1):7–34.

[20] Decancq, K., Van Ootegem, L., Verhofstadt, E. (2013). What If We Voted on the Weights of a Multidimensional Well?Being Index? An Illustration with Flemish Data. Fiscal Studies, 34(3), 315-332.

[21] Esping-Andersen, G., Wagner, S. (2012). Asymmetries in the opportunity structure. Intergenerational mobility trends in Europe. Research in Social Stratification and Mobility, 30(4), 473-487.

[22] Ferreira, F.H.G., and Gignoux, J. (2014). The measurement of educational inequality: Achievement and opportunity, World Bank Economic Review, 28(2), 210-246

[23] Foster, J., McGillivray, M., Seth, S. (2009). Rank robustness of composite indices. OPHI Working paper 26.

[24] Foster, J.E., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures, Econometrica, 52(3), 761-766

[25] Foster, J. and Sen A. (1997) ?On Economic Inequality: After a Quarter Century? in A. Sen (ed.), On Economic Inequality, Extended Edition, Oxford University Press.

[26] Gamboa, L.F., and Waltenberg, F.B. (2012). Inequality of opportunity for educational achievement in Latin America: Evidence from PISA 2006-2009, Economics of Education Review, 31(5), 694-708.

[27] Greco, S., Ishizaka, A., Matarazzo, B., & Torrisi, G. (2018). Stochastic multi-attribute acceptability analysis (SMAA): an application to the ranking of Italian regions. Regional Studies, 52(4), 585-600.

[28] Greco, S., Ishizaka, A., Tasiou, M., & Torrisi, G. (2017). On the methodological framework of composite indices: A review of the issues of weighting, aggregation, and robustness. Social Indicators Research, 1-34.

[29] Gutirrez, G., Jerrim, J., Torres, R. (2019). School segregation across the world: has any progress been made in reducing the separation of the rich from the poor?. The Journal of Economic Inequality, 1-23.

[30] Hanushek, E. A., & Raymond, M. E. (2006). School accountability and student performance. Regional Economic Development, 2(1), 51-61.

[31] Herrero, C., Méndez, I., & Villar, A. (2014). Analysis of group performance with categorical data when agents are heterogeneous: The evaluation of scholastic performance in the OECD through PISA. Economics of Education Review, 40, 140-151.

[32] Hopfenbeck, T. N. (2016). The power of PISA – limitations and possibilities for educational research. Assessment in Education: Principles, Policy & Practice. 23(4). 423-426.

[33] Inglehart, R.F., Norris, P. (2016) Trump, Brexit, and the Rise of Populism: Economic Have-Nots and Cultural Backlash, Faculty Research Working Paper Series n. 16/026, Harvard Kennedy School.

[34] Karagiannis, G., Paschalidou, G. (2017). Assessing research effectiveness: a comparison of alternative nonparametric models. Journal of the Operational Research Society, 68(4), 456-468.

[35] Kriesi, H., (1999) Movements of the Left. Movements of the right. Putting the mobilization of the two types of movements into context. In Kitshelt, Lange, Marks, Stephens (eds.) Continuity and Change in Contemporary Capitalism, Cambridge University Press, New York, pp. 398-426.

[36] Lagravinese, R., Liberati, P., Resce, G. (2019). Exploring health outcomes by stochastic multicriteria acceptability analysis: An application to Italian regions. European Journal of Operational Research, 274(3), 1168-1179.

[37] Lagravinese, R., Liberati, P., Resce, G. (2020). The impact of economic, social and cultural conditions on educational attainments. Journal of Policy Modeling,42(1), 112-132

[38] Lahdelma R., Hokkanen J., Salminen P. (1998). SMAA-stochastic multiobjective acceptability analysis. European Journal of Operational Research, 106(1), 137-143.

[39] Lahdelma R., Salminen P. (2001). SMAA-2: Stochastic multicriteria acceptability analysis for group decision making. Operations Research, 49(3), 444-454.

[40] Lambert, P.J. (2001). The distribution and redistribution of income, Manchester and New York, Manchester University Press

[41] Liberati, P. (2015). The World Distribution of Income And Its Inequality, 1970–2009. Review of Income and Wealth, 61(2), 248-273.

[42] Lovell, C. K., & Pastor, J. T. (1999). Radial DEA models without inputs or without outputs. European Journal of operational research, 118(1), 46-51.

[43] Melyn, W., & Moesen, W. (1991). Towards a synthetic indicator of macroeconomic performance: unequal weighting when limited information is available. Public economics research papers, 1-24.

[44] Murnane, R. J., Willett, J. B., Duhaldeborde, Y., & Tyler, J. H. (2000). How important are the cognitive skills of teenagers in predicting subsequent earnings?. Journal of Policy Analysis and Management, 19(4), 547-568.

[45] Nardo, M., Saisana, M., Saltelli, A., Tarantola, S., Hoffman, A., Giovannini, E. (2008). Handbook on constructing composite indicators. Paris: OECD-JRC.

[46] OECD (2009). PISA Data Analysis Manual. OECD Publishing. Paris.

[47] OECD (2017a). Programme for International Student Assessment (PISA). Paris.

[48] OECD (2017b). Glossary of Statistical Terms. Paris.

[49] Paruolo, P., Saisana, M., and Saltelli, A. (2013). Ratings and rankings: voodoo or science? Journal of the Royal Statistical Society: Series A (Statistics in Society), 176(3):609–634.

[50] Permanyer I. (2011) Assessing the robustness of composite indices rankings. Review of Income and Wealth, 57(2), 306-326.

[51] Picketty, T. (2018) Brahmin Left vs Merchant Right: Rising Inequality & the Changing Structure of Political Conflict, WID.Working Paper Series, World Inequality Lab, N. 2018/7

[52] Porter M. E. (1990). The Competitive Advantage of Nations, Free Press, New York.

[53] Pyatt, G. (1976). On the interpretation and disaggregation of Gini coefficients. The Economic Journal, 86(342), 243-255.

[54] Raitano, M., Vona, F. (2016). Assessing students? equality of opportunity in OECD countries: the role of national-and school-level policies. Applied Economics, 48(33), 3148-3163.

[55] Ray, S. C., & Chen, L. (2015). Data envelopment analysis for performance evaluation: a child's guide. In Benchmarking for Performance Evaluation (pp. 75-116). Springer India.

[56] Roemer, J. E. (1998). Equality of Opportunity. Cambridge. Harvard University Press.

[57] Saisana, M., Tarantola, S., Saltelli, A. (2005). Uncertainty and sensitivity techniques as tools for the analysis and validation of composite indicators. Journal of the Royal Statistical Society A, 168(2),307-323.

[58] Sanchez, J.F., Sanchez, M.C., Badillo, R., del Carmen Marco, M., LLinares, J.V., and Alvarez, S. (2016). A new multidimensional measurement of educational poverty. An application to PISA 2012, mimeo.

[59] Savaglio, E. (2006). Three approaches to the analysis of multidimensional inequality. In: F. Farina, E. Savaglio (Eds.), Inequality and Economic Integration (pp. 264-277). London: Routledge.

[60] Sen, A.K. (1976). Poverty: An ordinal approach to measurement, Econometrica, 44(2), 219- 231.

[61] Sen, A. (1992). Inequality reexamined. Clarendon Press.

[62] Sharpe, A. (2004). Literature Review of Frameworks for Macro-indicators. Centre for the Study of Living Standards, Ottawa, CAN.

[63] Stiglitz, J. E., Sen, A., & Fitoussi, J. P. (2010). Report by the commission on the measurement of economic performance and social progress. Paris: Commission on the Measurement of Economic Performance and Social Progress.

[64] Tervonen, T., Lahdelma, R. (2007). Implementing stochastic multicriteria acceptability analysis. European Journal of Operational Research, 178, 500-513.

[65] Theil, H. (1967). Economics and information theory (No. 04; HB74. M3, T4.).

[66] Thompson, R. G., Singleton Jr, F. D., Thrall, R. M., & Smith, B. A. (1986). Comparative site evaluations for locating a high-energy physics lab in Texas. interfaces, 16(6), 35-49.

[67] Villar, A. (2016). Educational poverty as a welfare loss: Low performance in the OECD according to PISA 2012, Modern Economy, 7(4), 441-449.

[68] Weymark, J. A., (2006). The normative approach to the measurement of multidimensional inequality. In: F. Farina, E. Savaglio (Eds.), Inequality and Economic Integration (pp. 303-328). London: Routledge.

[69] World Bank. (1997). World Development Report 1997: The State in a Changing World, World Bank: Oxford University Press.

[70] Yitzhaki S., Lerman R. (1991), Income Stratification and Income Inequality, Review of Income and Wealth, 37, 313-329.

[71] Yitzhaki S., Schechtman E. (2013), The Gini Methodology – A Primer on a Statistical Methodology, Springer, New York.

[72] Yitzhaki, S. (1994). Economic distance and overlapping of distributions. Journal of Econometrics, 61(1), 147-159.

A Appendix

Table A1: Multidimensional ANOGI of Downward cumulative rank acceptability index for the top 20 per cent of ranking (Weights normal distributed around 1/3)

Year	Tot. Ineq.	Standard WI	Impact of overl. on WI	Standard BI	Impact of overl. on BI
2006	0.797	0.029	0.657	0.372	-0.261
2009	0.797	0.032	0.661	0.367	-0.262
2012	0.797	0.032	0.681	0.324	-0.240
2015	0.798	0.03	0.712	0.272	-0.215

Authors elaboration on OECD (2017a)

Table A2: Multidimensional ANOGI of Upward cumulative rank acceptability index for the bottom 20 per cent of ranking (Weghits normal distributed around 1/3)

Year	Tot. Ineq.	Standard WI	Impact of overl. on WI	Standard BI	Impact of overl. on BI
2006	0.798	0.035	0.551	0.499	-0.287
2009	0.798	0.043	0.551	0.492	-0.288
2012	0.798	0.039	0.555	0.487	-0.282
2015	0.799	0.024	0.617	0.440	-0.282

Authors elaboration on OECD (2017a)