# Demonstration of Simpson's Paradox in PISA 2015 Data: Confusing Differences between Boys and Girls

Gašper Cankar
National Examinations Centre, Slovenia

**Abstract:** This paper explores the occurrence of a Simpson's paradox in PISA 2015 science literacy data. Simpson's paradox, a case of contradicting interpretations when results are analysed by groups or aggregated as a whole, has both a practical and an academic significance. It is an interesting phenomenon that is far from theoretical and when it happens, it has profound effects on the interpretation and if left unidentified can cause confusion and misunderstanding. This paper demonstrates best ways to detect Simpson's paradox through appropriate tables and graphs. Actual occurrences of a Simpson's paradox and conditions leading to them are explored using PISA 2015 gender differences in science literacy data in five central European countries – Austria, Croatia, Czech Republic, Slovakia and Slovenia. In countries where the occurrence of a Simpson's paradox was detected, we provide correct interpretation of the results. Beside creating problems with interpretation an occurrence of a Simpson's paradox also provides new insight – it signifies that there is very different gender composition in different educational tracks which has important implications for the educational governance. We will discuss implications of these findings in context of Slovenian educational system.

**Keywords:** PISA; Simpson's paradox; gender differences; educational tracks; governance

Statistical paradoxes are usually not important when interpreting data from the international large scale assessments (ILSA). As Gardner (1982) points out, they are an interesting topic in itself, but they are more commonly viewed as a hobby of a retired statistician, a relaxing pursuit of students of statistics or as brainteasers intended to rouse curiosity and interest in the mathematics. Sometimes, however they also have profound implications on interpretation of a real data. In this paper we will focus on a Simpson's paradox, it's real life occurrences and implications for use and interpretation of data. As it turns out, the knowledge about a Simpson's paradox can be useful when interpreting results from the large scale assessments.

A Simpson's paradox is a situation where we get conflicting interpretations when same results are analysed at different levels of grouping. Or as Lesser (2001) puts it: "Simpson's paradox can be concisely defined as the reversal of a comparison when data are grouped." It was named a Simpson's paradox by Blythe (1972) after Edward Simpson, a British statistician who first wrote about it when he was still a post-graduate student (Simpson, 1951). Blythe neglected that another British statistician Udny Yule wrote about same paradox already in 1903 (Yule, 1903). To acknowledge this some authors nowadays also call it Yule-Simpson effect (Demers & Rossmo, 2015). We will use a shorter name throughout the paper.

https://doi.org/10.14712/23363177.2018.296
www.orbisscholae.cz

The paradox can be best explained through an example. Imagine two classes of students (Class A & Class B) learning same course on Mathematics and taking same test at the end. Both classes would consist of 30 students and Table 1 presents their average points achieved on test reported by gender.

**Table 1** Results on Mathematics achievement test for Class A and B.

| | Average (boys) | Average (girls) | Difference (girls–boys) |
|---|---|---|---|
| Class A | 23.6 | 20.3 | −3.3 |
| Class B | 13.7 | 10.4 | −3.3 |
| Total | 16.0 | 18.0 | +2.0 |

If we would compare boys and girls in Class A alone, we would conclude from difference that the boys on average perform better. Same conclusion would follow from the difference in Class B (3.3 points in favour of boys). But when we combine data from both classes, girls outperform boys for 2 points! This is called a Simpson's paradox and it is not an error in calculations. The reason for the observed phenomena is in the distribution of boys and girls in both classes as seen in Table 2.

**Table 2** Number of boys and girls in Classes A and B.

| | **Number (boys)** | **Number (girls)** |
|---|---|---|
| Class A | 7 | 23 |
| Class B | 23 | 7 |
| Total | 30 | 30 |

From Table 1 it was obvious that students in Class A on average performed much better then students in Class B. Therefore, the grouping of students into classes with regard to their Mathematics achievement was not random. The unequal proportions of boys and girls (7:23) combined with non-random grouping resulted in an observed paradox. In other words: in Class A the small number of high performing boys outperformed more numerous female peers. In Class B larger number of boys again outperformed smaller number of girls. Only when we join classes we discover the actual difference where girls on average performed better on the Mathematics test then boys. If we make conclusions only on averages from each class, we miss the real picture.

The example above is artificially constructed to explain the paradox. What about in real life? Is the paradox in practice really common or is it a rare finding that occurs only seldom? Judging from the amount of research literature the occurrence is certainly not uncommon. If we focus only on the recent research literature it can be found in different areas of science and life in general: medicine (Baker & Kramer, 2001; Rücker & Schumacher, 2008), administration (Demers & Rossmo, 2015) and even sports (Wright, 2012). In this paper we will explore its occurrence in large scale assessments in education.

## 1 State-of-the-art

Before we start with an analysis we will explore different ways to represent a Simpson’s paradox as such methods can help researchers to detect it and act accordingly.

To detect the Simpson’s paradox we can always calculate differences of averages in all subgroups and in a sample as a whole and see if it occurs as we did in Table 1 of our example. This however misses the point that there are many situations when we don’t get an actual Simpson’s paradox (reversal of difference between averages) but we get a substantial increase or decrease in the difference. Checking actual tables of averages may be a robust and concise way but it might be less visually appealing as a lot of tables makes results hard to read.

The best methods to spot a Simpson’s paradox in practice are graphical. This is due to the fact that a proper graphical representation accounts for different proportions of students in subgroups and difference in averages at the same time.

We will explore three ways to represent data: Bar-plot representation, Square representation, and Trapezoidal representation.

### 1.1 Bar-plot representation

This is a simple example trying to demonstrate on the same picture proportions of students and their average scores. Figure 1 shows a bar-plot of Class A and B students from our example.

Bar plot in Figure 1 fairly well shows differences in proportions but not differences in averages. It is simple to construct but it doesn’t warn us about a Simpson’s paradox on the first glance as there is no difference calculated. The reader must infer the inversion from comparison of averages as it is not readily visible.

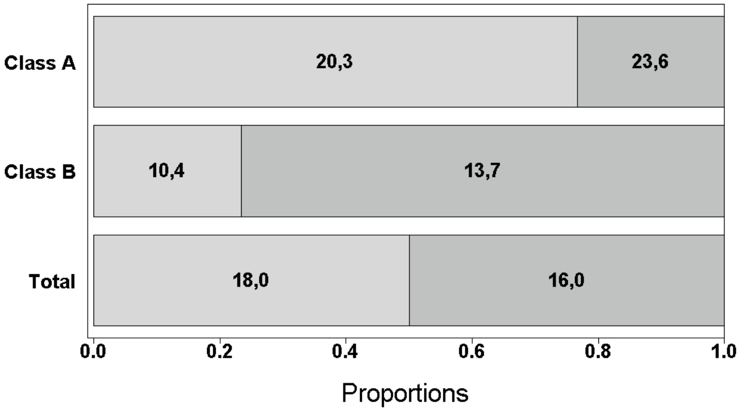
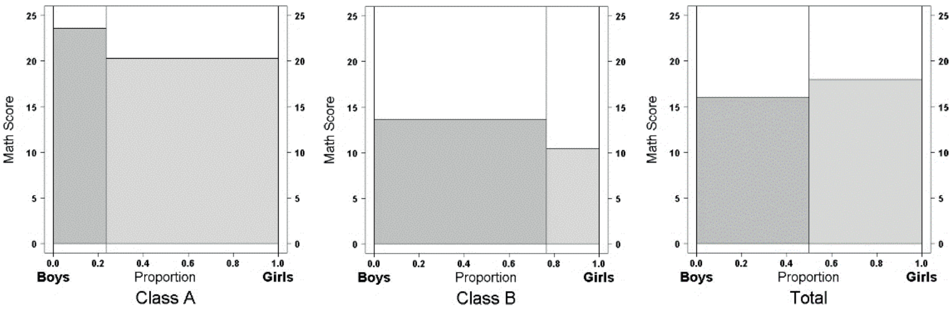


**Figure 1** Bar plot of proportions of girls (light) and boys (dark) in classes A and B with averages printed inside bars.

### 1.2 Square representation

This representation tries to capture differences in proportions and differences in averages in the same figure. It is adapted from unit square representation described by Lesser (2001). For each comparison (Class A, Class B and Total) we construct a square where one dimension represents proportions and the other dimension represents average scores. From series of three figures (for Class A, Class B and Total) we can observe what happened to average scores in subgroups and in total. When drawing the figure we first divide the square according to the proportions (in our example of boys and girls). Then we draw averages for each gender and shade each area respectively. Figures 2 to 4 show graphs for our example.

Square representation allows us to compare graphs for subgroups with the last graph that shows all subgroups together. The inversion of difference in the last graph (Figure 4) is now evident and it's easier to understand what happened. The downside is that you can't represent all information in just one graph but you have to compare several figures simultaneously.



**Figures 2–4** Square representations of proportions and average scores for Boys and Girls in classes A, B and in Total respectively.

### 1.3 Trapezoidal representation

Trapezoidal representation of a Simpson's paradox was first proposed by Tan (1986) who observed that "the length of any line segment which is parallel to the two bases and has its endpoints on the nonparallel sides of a trapezoid is the weighted mean of the lengths of the two bases". What this actually means is that we can plot all information on the same graph following this procedure:

- We start with square plot where x axis represents Proportions, left y axis represents Class A math score and right y axis represents Class B math score.
- On left y axis we mark Class A average score for boys. On right y axis we mark Class B average score for boys.
- We draw the line segment connecting both points (Class A and B boys' average score).
- On the x axis we mark the proportions of boys in Class A and Class B (from all the boys in Total).

– The vertical line delineating those two proportions actually intersects the line connecting both average scores right at the point of total average score for boys. Example is shown in Figure 5.

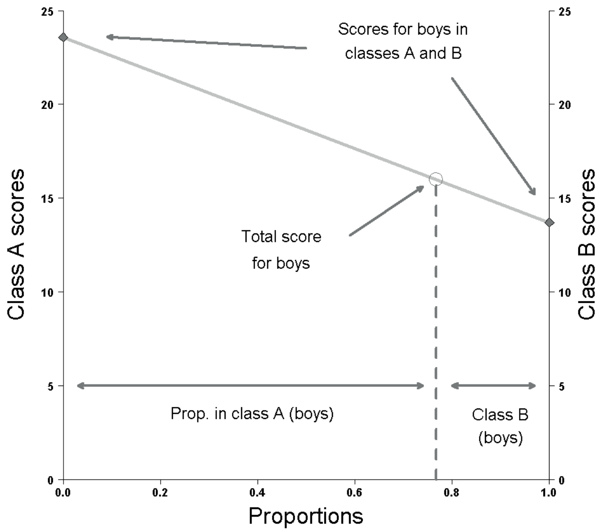


**Figure 5** Example for construction of trapezoidal representation for boys.

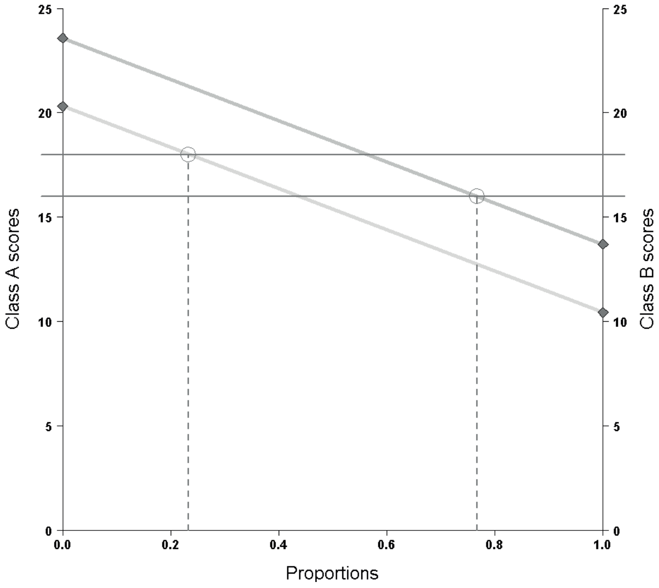


**Figure 6** Trapezoidal representation of our example of classes A and B. Left circle represents girls' average, right circle boys' average.

If we repeat same procedure for girls we can draw on the same graph another set of lines for girls. Then we can compare on the same graph differences in lines connecting averages and differences in heights at intersections (where the averages of all the boys and all the girls can be found). Our example of a Simpson's paradox can be seen in Figure 6.

Figure 6 shows very clearly that girls have lower average in both classes A and B. At the same time we also see that the average from both classes together is higher for girls than for boys. With trapezoidal representation we can show a Simpson's paradox in only one graph. There is a downside though − the method is suitable only when we have two subgroups like Class A and B in our example. If we would have three classes, the trapezoidal representation couldn't be applied.

We presented three graphical ways to explore the relationship between differences in subpopulations and in the total population and we mentioned their strengths and shortcomings. Trapezoidal representation seems most prudent as it clearly shows all information in just one graph, but it will be unusable for our purpose in this paper since we will be exploring occurrence of a Simpson's paradox between boys and girls in educational tracks. Most countries have their 15-year-old students in more than two educational tracks of formal education which suggests we should use graphical method that can accommodate more than two groups. One option would be to proceed with a Square representation but educational tracks present quite a challenge since they are a) numerous, which means a lot of graphs for each country; and b) not equal in size. Some educational tracks cover large portions of population of 15-year-olds other educational tracks include only small subgroups. Making them visually equal might again skew the interpretation.

To address this issue we will modify the Square representation by joining all educational tracks in the same graph and defining their widths according to the size of population in each track. Overall averages can be drawn as horizontal lines across whole graph. Examples are shown in the results section below.

### 1.4 Hypothesis

To focus our research, we state following two null hypotheses about differences between boys and girls in total and in subpopulations of each educational track (for each country):

*H01: Differences between boys and girls in PISA science results within educational tracks are equal to overall difference between boys and girls in each country.*

We also state stricter hypothesis that explicitly involves a Simpson's paradox (for each country):

*H02: Differences between boys and girls in PISA science results within educational tracks and in total don't show the pattern of Simpson's paradox (reversed difference) in each country.*

## 2 Method

This research draws data from the Programme for International Student Assessment (PISA) from 2015 cycle. Participants are students who were at the time of PISA main study 15 years old and still in formal education. To limit our exploration, we selected data from following countries: Austria, Croatia, Czech Republic, Slovakia and Slovenia. On this data we performed secondary data analysis to find out proportions of boys and girls in each educational track and their score on Science literacy.

As Smith (2008) points out secondary data analysis can be full of errors if it’s not done correctly. In case of ILSA we therefore consulted Technical report (OECD, 2017) where appropriate. All secondary data analyses were made using software IDB Analyzer 4.0.21 (IEA, 2018), using all 10 plausible values for Science literacy (PVSCIE) and the Final trimmed nonresponse adjusted student weight (W_FSTUWT). Plausible values are student’s results (in our case for Science literacy) prepared in such a way that researchers can calculate standard errors of any statistical parameter they estimate from them. This is very important since it helps us to interpret the data better and puts findings into a perspective. Student weights (W_FSTUWT) are ponders that reflect sampling procedure and enable us to calculate representative estimates for a whole population of 15-year-olds in a country even if only a sample participated in a study.

Proportions by an educational track and gender and PVSCIE averages as well as standard errors (for significance testing) were calculated using the module ’Percentages and means’. Missing values were excluded from analyses by default. Educational tracks were captured in a PISA variable PROGN and names of educational tracks for each country are taken from that variable. Graphical representations were made using a statistical environment R (R Core Team, 2017).

## 3 Results and interpretations

For each country’s results we will present PISA 2015 science results (PVSCIE) grouped by gender and educational tracks as noted in a variable PROGN. Students that participate in PISA can be in very different educational tracks; some are still in a comprehensive basic education, others already started in educational programmes leading to different secondary education outcomes. Educational tracks also differ widely in frequency – some are very popular and include large proportions of a whole population, others include only handful of students. Tables for each country are therefore not directly comparable. Educational tracks within the tables are ordered ascending according to average science score for each track.

To better understand proportions by gender and educational track each table also includes percentages of girls and boys and sums of student weights – they denote the size of a population captured in each statistic. Last column in each table presents a difference in science score between girls and boys in each educational track and in total (the last line). Positive difference means girls have higher average PISA 2015 science score than boys.

**Table 3** PISA 2015 science results by gender and educational track for Austria.

| | National Study Programme | $N_{GIRLS}$ (W_FSTUWT) | $N_{BOYS}$ (W_FSTUWT) | % (GIRLS) | % (BOYS) | PVSCIE (GIRLS) | PVSCIE (BOYS) | Difference (GIRLS-BOYS) |
|---|---|---|---|---|---|---|---|---|
| Pr.1 | Compulsory school | 1925 | 2553 | 42.99 | 57.01 | 366.49 | 395.01 | −28.52** |
| Pr.2 | Voc. sch. for apprentices | 4268 | 8782 | 32.71 | 67.29 | 417.38 | 442.13 | −24.75** |
| Pr.3 | Intermed. tech. and voc. schools | 6048 | 5224 | 53.66 | 46.34 | 428.29 | 451.63 | −23.34** |
| Pr.4 | Higher tech. and voc. college | 13011 | 11980 | 52.06 | 47.94 | 501.80 | 547.84 | −46.04** |
| Pr.5 | Academic secondary school | 11091 | 8497 | 56.62 | 43.38 | 544.53 | 572.68 | −28.15** |
| | Total | 36345 | 37034 | 49.53 | 50.47 | 485.53 | 504.37 | −18.84** |

** Differences are statistically significant at $p < 0.05$.

PISA 2015 science results for Austria in Table 3 on first glance present uniform picture – boys outperformed girls within every educational track and also on a country's level. We can note, however that overall difference is smaller than any difference within educational tracks. A Simpson's paradox didn't happen, but the data on a whole and grouped by educational tracks suggest slightly different conclusions. While differences within educational tracks suggest that boys outperform girls for more than 23 points and in case of most numerous educational programme for more than 46 points the total difference is actually only 18.84 points.

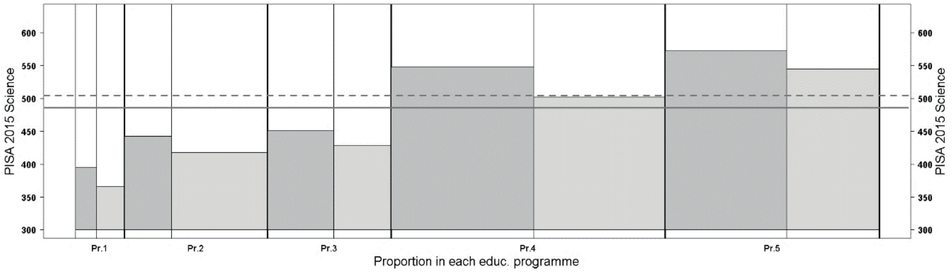


**Figure 7** PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Austria. Width of each programme corresponds to proportion of programme in a whole population. Lines show total average (dashed – boys, solid – girls).

Figure 7 shows the same trend of bigger differences in each educational track and smaller overall difference for Austrian data. A Simpson’s paradox didn’t occur but conclusions about the size of difference when examining data per country and within educational tracks are different.

**Table 4** PISA 2015 science results by gender and educational track for Croatia.

| | National Study Programme | $N_{GIRLS}$ (W_FSTUWT) | $N_{BOYS}$ (W_FSTUWT) | % (GIRLS) | % (BOYS) | PVSCIE (GIRLS) | PVSCIE (BOYS) | Difference (GIRLS-BOYS) |
|---|---|---|---|---|---|---|---|---|
| Pr.1 | Primary school – lower sed.+ | 54 | 34 | 61.89 | 38.11 | 339.88 | 402.30 | −62.42** |
| Pr.2 | Lower qualification voc. prog. | 40 | 37 | 51.58 | 48.42 | 340.00 | 344.23 | −4.23 |
| Pr.3 | Vocational prog. for crafts | 2492 | 4091 | 37.85 | 62.15 | 381.54 | 399.14 | −17.60** |
| Pr.4 | Vocational prog. for industry | 654 | 1638 | 28.52 | 71.48 | 382.80 | 403.85 | −21.05** |
| Pr.5 | Art programmes | 285 | 51 | 84.88 | 15.12 | 451.04 | 489.66 | −38.62 |
| Pr.6 | Four year vocational prog. | 9214 | 9039 | 50.48 | 49.52 | 454.76 | 483.49 | −28.73** |
| Pr.7 | Gymnasium | 8487 | 4783 | 63.96 | 36.04 | 527.78 | 563.63 | −35.85** |
| | Total | 21226 | 19673 | 51.90 | 48.10 | 472.59 | 478.42 | −5.83 |

sed+ – secondary education; ** differences are statistically significant at $p < 0.05$.

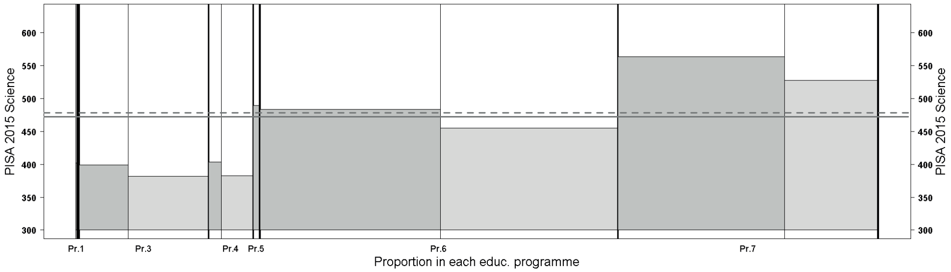


**Figure 8** PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Croatia. Width of each programme corresponds to proportion of whole population. Lines show total average (dashed – boys, solid – girls).

PISA 2015 science results for Croatia in Table 4 show similar trend than in Austria. Although boys outperform girls on whole and within every educational track we can still note that overall difference is rather low (5.83) compared to differences in most numerous educational tracks where boys outperform girls on average between 17 and 35 points! This is also evident in statistical significance results – overall difference is within the margins of ±1.96 standard errors while differences in most educational tracks are much bigger and statistically significant.

Figures of differences for educational tracks in Croatia give similar conclusion as Table 4 – reversal of differences didn't occur but it is much smaller on a whole compared to major educational tracks within the country.

**Table 5** PISA 2015 science results by gender and educational track for Czech Republic.

| | National Study Programme | $N_{GIRLS}$ (W_FSTUWT) | $N_{BOYS}$ (W_FSTUWT) | % (GIRLS) | % (BOYS) | PVSCIE (GIRLS) | PVSCIE (BOYS) | Difference (GIRLS-BOYS) |
|---|---|---|---|---|---|---|---|---|
| Pr.1 | Basic special schools | 680 | 813 | 45.55 | 54.45 | 361.18 | 348.96 | 12.22 |
| Pr.2 | Secondary special schools | 226 | 248 | 47.62 | 52.38 | 403.92 | 405.95 | −2.03 |
| Pr.3 | Voc\tech sed+ without maturate | 2850 | 4618 | 38.17 | 61.83 | 400.53 | 420.63 | −20.10** |
| Pr.4 | Basic school | 17140 | 21852 | 43.96 | 56.04 | 464.64 | 471.25 | −6.61 |
| Pr.5 | Voc\tech sed+ with maturate | 11532 | 8636 | 57.18 | 42.82 | 486.79 | 525.64 | −38.85** |
| Pr.6 | 4-year gymnasium | 4031 | 2157 | 65.15 | 34.85 | 567.80 | 595.70 | −27.90** |
| Pr.7 | 6, 8-year gymnasium and 8-year conservatory (lower secondary) | 2268 | 2717 | 45.49 | 54.51 | 581.31 | 605.98 | −24.67** |
| Pr.8 | 6, 8-year gymnasium (upper secondary) | 2400 | 2351 | 50.51 | 49.49 | 593.02 | 626.00 | −32.98** |
| | Total | 4112 | 43392 | 48.66 | 51.34 | 488.40 | 497.03 | −8.63** |

sed+ – secondary education; ** Differences are statistically significant at $p < 0.05$.

In Table 5 we present PISA 2015 science results by gender and educational track for the Czech Republic. Gender difference on country level (8.63) are similar to

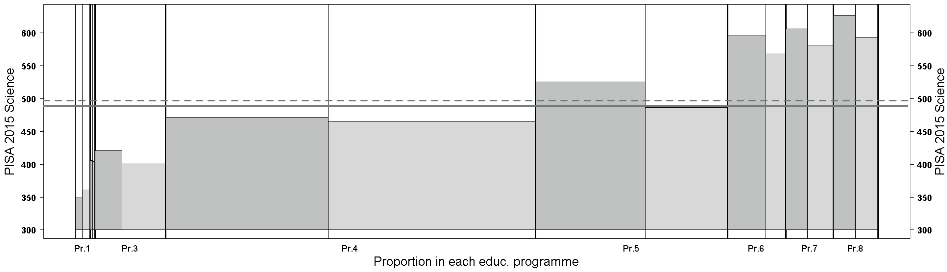


**Figure 9** PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Czech Republic. Width of each programme corresponds to proportion of whole population. Lines show total average (dashed – boys, solid – girls).

difference between students still in Basic schools. This makes sense since those students are still in comprehensive part of educational system. Differences increase drastically in secondary education where students choose educational track according to their abilities and preferences.

Figure 9 and Table 5 show that Simpson’s paradox didn’t occur in case of PISA 2015 data for Czech Republic but they also show that secondary education tracks show much larger differences than Basic schools and all tracks together.

**Table 6** PISA 2015 science results by gender and educational track for Slovakia.

| | National Study Programme | $N_{GIRLS}$ (W_FSTUWT) | $N_{BOYS}$ (W_FSTUWT) | % (GIRLS) | % (BOYS) | PVSCIE (GIRLS) | PVSCIE (BOYS) | Difference (GIRLS-BOYS) |
|---|---|---|---|---|---|---|---|---|
| Pr.1 | Vocational basic school | 580 | 683 | 45.94 | 54.06 | 306.16 | 306.92 | −0.76 |
| Pr.2 | Secondary vollege – without SLE | 1014 | 1807 | 35.94 | 64.06 | 355.58 | 377.77 | −22.19** |
| Pr.3 | Basic school | 9655 | 11518 | 45.60 | 54.40 | 431.51 | 440.72 | −9.21** |
| Pr.4 | Secondary college – with SLE | 6122 | 7237 | 45.83 | 54.17 | 453.48 | 466.97 | −13.49** |
| Pr.5 | High school | 5415 | 3293 | 62.19 | 37.81 | 538.46 | 559.27 | −20.81** |
| Pr.6 | Secondary school (ISCED2) | 603 | 494 | 54.96 | 45.04 | 540.15 | 558.69 | −18.54 |
| Pr.7 | Secondary school (ISCED3) | 682 | 549 | 55.40 | 44.60 | 557.04 | 566.11 | −9.07 |
| | Total | 24072 | 25582 | 48.48 | 51.52 | 461.22 | 460.36 | 0.86 |

SLE – school leaving examination; ** differences are statistically significant at $p < 0.05$.

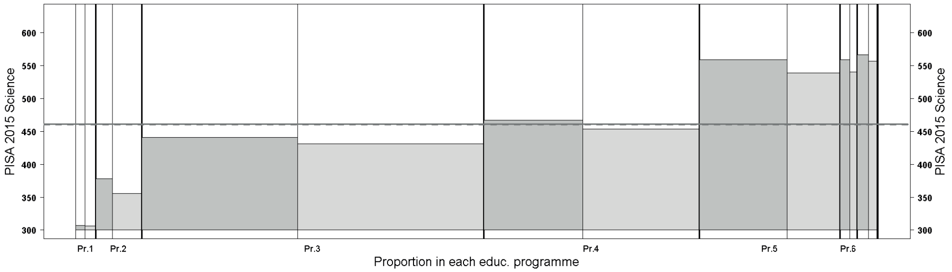


**Figure 10** PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Slovakia. Width of each programme corresponds to proportion of a whole population. Lines show total average (dashed – boys, solid – girls).

Pisa 2015 results for Slovakia in Table 6 are an example of a Simpson's paradox in real life data. While all educational tracks suggest that boys outperform girls, on whole results suggest otherwise.

Graphically Figure 10 clearly shows that great differences in each educational track (most of them are statistically significant at the *p*-value 0.05 and less) don't translate to overall difference. Here results between boys and girls are practically identical as they are well within margins of standard error ($SE_{GIRLS} = 3.31$; $SE_{BOYS} = 2.98$).

**Table 7** PISA 2015 science results by gender and educational track for Slovenia.

| | **National Study Programme** | **$N_{GIRLS}$ (W_FSTUWT)** | **$N_{BOYS}$ (W_FSTUWT)** | **% (GIRLS)** | **% (BOYS)** | **PVSCIE (GIRLS)** | **PVSCIE (BOYS)** | **Difference (GIRLS-BOYS)** |
|---|---|---|---|---|---|---|---|---|
| Pr.1 | Voc. ed. short duration | 42 | 121 | 25.58 | 74.42 | 356.10 | 380.83 | −24.73** |
| Pr.2 | Voc. ed. medium duration | 737 | 1786 | 29.20 | 70.80 | 403.94 | 423.81 | −19.87** |
| Pr.3 | Basic (elementary) education | 347 | 510 | 40.53 | 59.47 | 440.68 | 446.90 | −6.22 |
| Pr.4 | Technical ed. | 3207 | 3729 | 46.24 | 53.76 | 486.13 | 510.41 | −24.28** |
| Pr.5 | Sed+ – technical gymnasiums | 512 | 524 | 49.38 | 50.62 | 537.36 | 566.13 | −28.77** |
| Pr.6 | Sed+ – general gymnasiums | 3264 | 1993 | 62.09 | 37.91 | 576.78 | 596.31 | −19.53** |
| | Total | 8109 | 8664 | 48.34 | 51.66 | 515.77 | 510.14 | 5.63** |

sed+ – secondary education; ** differences are statistically significant at $p < 0.05$.

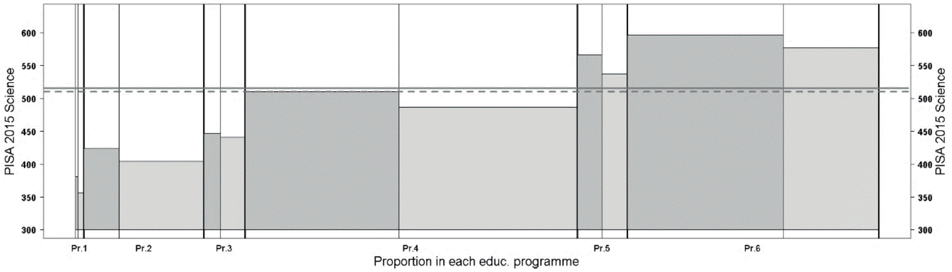


**Figure 11** PISA 2015 science scores for boys (dark) and girls (light) in different educational tracks in Slovenia. Width of each programme corresponds to proportion of a whole population. Lines show total average (dashed – boys, solid – girls).

PISA 2015 science data by gender and educational tracks for Slovenia demonstrates a Simpson's paradox. Since for Slovenia standard errors are quite small ($SE_{\text{GIRLS}}$ = 1.88; $SE_{\text{BOYS}}$ = 1.92) the difference of 5.63 points is statistically significant and shows that on average girls outperformed boys, while results in every educational track suggest otherwise.

Square representation graphs for Slovenia in Figure 11 show the remarkable shift of a Simpson's paradox. While generalizations from every educational track would implicate that boys outperform girls in PISA 2015 science literacy in fact the opposite is true!

We can summarize our findings with regards to our hypotheses as following:

*H01: Differences between boys and girls in PISA science results within educational tracks are equal to overall difference between boys and girls.*

Austria *CONFIRMED* – Overall difference and differences within educational tracks show same trend and are both statistically significant.

Croatia *NOT CONFIRMED* – Overall difference and differences within educational tracks show same trend but overall difference is not statistically significant.

Czechia *CONFIRMED* – Overall difference and differences within educational tracks show same trend and are both statistically significant.

Slovakia *NOT CONFIRMED* – Overall difference and differences within educational tracks don't show same trend and overall difference is not statistically significant.

Slovenia *NOT CONFIRMED* – Overall difference and differences within educational tracks don't show same trend and both are statistically significant in different directions!

*H02: Differences between boys and girls in PISA science results within educational tracks and in total don't show the pattern of Simpson's paradox (reversed difference).*

| Austria | Croatia | Czechia | Slovakia | Slovenia |
|---|---|---|---|---|
| *CONFIRMED* | *CONFIRMED* | *CONFIRMED* | *NOT CONFIRMED* | *NOT CONFIRMED* |

## 4 Discussion

A simple analyses of differences by gender or other characteristics are very common. Furthermore, due to the simplicity of calculating averages they are often not done and interpreted by statisticians alone but by people with wide variety of statistical knowledge. As Smith (2008) notes, secondary data analysis in general is often seen with scepticism because data, gathered for one reason is being used for another and this opens doors to errors. But even Smith (2008) recognizes great opportunities in using large scale data coming from well conducted research with good technical documentation. To avoid the pitfalls we must empower the researchers that use data. We demonstrated that researchers must be aware of possibilities for occurrence of Simpson's paradox and must pay attention against its effect on results and interpretations. This paper should empower researchers to keep guard and discover Simpson's paradox during analyses and thus provide correct interpretation of the findings.

Simpsons paradox can easily influence results of modern statistical analyses when we combine data sets from different sources and produce meta-analyses. Cohen and Moch (2017) warn researchers to be on guard and look for occurrences of Simpson's paradox when combining datasets. They provide examples from medicine, where samples are often small and the paradox occurs because different datasets are of different sizes. They cite cases where the results were different when datasets were analysed separately as when combined and conclude that only when researchers are prepared for the phenomenon of Simpson's paradox in advance can we avoid erroneous results and interpretations. Their results can be easily generalized outside medicine.

We should be aware that in case of Simpson's paradox it is not always straightforward which of the results is erroneous. In our PISA 2015 data the differences within educational tracks were misleading and difference in total dataset showed the real difference but it could easily be the case that total difference would be wrong and differences by subgroups would be correct. Baker and Kramer (2001) explored generalizations from studies of another set of medical interventions. They report on example where the treatment was better for males and females but when datasets were combined it appeared to be harmful to everyone!

The real examples from PISA 2015 data also provides several lessons. First lesson would be that it is important to follow proportions of boys and girls in different educational tracks. The proportions widely differ and the effects on educational systems in the long run can be profound.

Differences within educational tracks are interesting as they are heavily weighted by the proportions of boys and girls in each track and even more importantly by their preference for certain educational track. Boys and girls aren't allocated to educational tracks randomly but they rather select them according to their abilities and preferences. Some vocational and technical tracks can be more appealing to boys than girls and in other tracks situation might be reversed. From the point of educational governance, it is important to wonder if observed proportions are a reason to worry or not. Educational systems around the world are often aware of such differences and try to act upon them and govern their educational systems to address this mostly through the questions of equity. One good example are initiatives to attract more girls into STEM. Such initiatives can be found globally and are among others supported by UNESCO (2014) and EU (2016).

Another lesson from our analysis is also that we shouldn’t generalize findings from one educational track to others (or to educational system as a whole). It is often the case that countries have only data for one educational track (like specific leaving examinations that isn’t available in other educational tracks). Findings from secondary analyses of such data shouldn’t be generalized to the educational tracks where similar data doesn’t exist or to whole educational system. As shown on example of PISA 2015 data we should consider the analysis carefully to avoid misleading interpretations.

Situations where differences in proportions have substantial influence on results are important to note regardless of the fact if there was an actual case of Simpson’s paradox. In our PISA 2015 data Simpson’s paradox occurred only in Slovakia and Slovenia, but similar underlying tendencies of smaller overall difference were also detected in all other countries. This is important for interpretation as it reveals that boys and girls in same educational track are not directly comparable. In case of Slovenia data shows great differences in gender composition in different educational tracks and this finding should serve as basis for raising the awareness about the issue and future steps that would address it. Since effects of education are often very long term and profound such warning signs should not be neglected.

The topic of this paper focuses on two main parts revolving around Simpson’s paradox: theoretical and empirical one. Theoretical part warns the researchers to keep guard and spot Simpson’s paradox when it occurs so the interpretations of the data are valid. We have demonstrated that Simpson’s paradox isn’t a statistical amusement, it’s a real threat to validity of conclusions based on data and it’s a clear signal of neglected and overlooked factors influencing the data. This brings us to our empirical part where we use PISA 2015 Science data to demonstrate Simpson’s paradox but in the process we also uncover new insights. When gender of students is compared, many countries show differences in allocation of boys and girls to educational tracks, differences that raise questions of equity and fairness of each educational system, differences that can have long lasting effects in each country.

## References

Blyth, C. R. (1972). On Simpson’s paradox and the sure-thing principle. *Journal of the American Statistical Association*, *67*(338), 364–366.

Baker, S. G., & Kramer, B. S. (2001). Good for women, good for men, bad for people: Simpson’s paradox and the importance of sex-specific analysis in observational studies. *Journal of Women’s Health and Gender-Based Medicine*, *10*(9), 867–872.

Cohen, B. S., & Moch, P. L. (2017). Guarding against Simpson's paradox when combining data sets. *Curriculum and Teaching Dialogue*, *19*(1 & 2), 153.

Demers, S., & Rossmo, D. K. (2015). Simpson's paradox in Canadian police clearance rates. *Canadian Journal of Criminology and Criminal Justice*, *57*(3), 424–434.

European Commission. (2016). *She figures 2015*. Luxembourg: Publications Office of the European Union. Retrieved from http://ec.europa.eu/research/swafs/pdf/pub_gender_equality/she_figures_2015-final.pdf.

Gardner, M. (1982). *Aha! Gotcha: paradoxes to puzzle and delight*. San Francisco: Freeman.

Lesser, L. M. (2001), Representations of reversal: An exploration of Simpson's paradox. In A. A. Cuoco, & F. R. Curcio (Eds.), *The roles of representation in school mathematics* (pp. 129–145). Reston, Virginia: National Council of Teachers of Mathematics.

OECD. (2017). *PISA 2015 technical report*. Paris: OECD Publishing.

R Core Team (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. On-line: https://www.R-project.org.

Rücker, G., & Schumacher, M. (2008). Simpson's paradox visualized: The example of the Rosiglitazone meta-analysis. *BMC Medical Research Methodology*, *8*(1).

Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society*, *13*(2), 238–241.

Smith, E. (2008). *Using secondary data in educational and social research*. New York, NY: McGraw Hill/Open University Press.

Tan, A. (1986). A geometric interpretation of Simpson's paradox. *College Mathematics Journal, 17*, 340-341.

UNESCO (2014). *UNESCO's promise: Gender equality, a global priority.* Paris: UNESCO. Retrieved from http://unesdoc.unesco.org/images/0022/002269/226923m.pdf.

Wright, B. (2012). *Best of N contests: Implications of Simpson's paradox in tennis [Masters Thesis]*. The Florida State University.

Yule, G. U. (1903). Notes on the theory of association of attributes in statistics. *Biometrika*, *2*(2), 121–134.

Dr. Gasper Cankar
Kajuhova 32 U
SI-1000 Ljubljana
Slovenia
gasper.cankar@ric.si