How analysis strategy affects analysis results

Assessing results space and structure of Silberzahn et al. (2018) through model specification

MSc Dissertation

Sebastian Ploner, sploner1@sheffield.ac.uk, University of Sheffield

Contents

ABSTRACT Numerous studies have show that analytical choices researchers make throughout the process of conducting a study affect the outcome. Recent crowdsourcing approaches to data analysis show that even well-intentioned analytical choices can make the difference between find a significant and strong or a non-significant effect. IN PROGRESS

Introduction

Theoretical framework

The Covid-19 pandemic has reaffirmed the need for sound scientific research to inform decision-making on a scientific and societal level (Collins, 2021). Rigorous research builds upon a systematic and well-reasoned approach to solving a research problem. Based on available literature, a falsifiable research question is defined and a corresponding hypothesis is developed. An appropriate study is then designed and conducted. To draw inferences from the gathered data, statistical models are developed and the effect of each variable is assessed. Every one of these steps is influenced by the decisions researchers make, which are known as *researcher degrees of freedom* (RDF, Simmons et al., 2011; Wicherts et al., 2016). Usually there are many feasible analytical strategies to answering a research question and none of them are inherently right or wrong (Carp, 2012). This often creates uncertainty, for example, about what covariates to include and how to model them, which in turn leads to inconsistent findings (Ioannidis, 2008; Patel et al., 2015). Recently, efforts have been made to better understand the scope of variation induced by different analytical strategies.

These efforts include crowdsourcing approaches to data analysis (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). Its premise is to have a large number of researchers team up into to smaller, independent groups to investigate the same research question based on the same dataset. For instance, Silberzahn et al. (2018) had 29 teams investigate the effects of a football player's skin colour on the odds of being sent off the field. The variation between the analytical strategies was substantial. There were 29 different analyses with 21 different combinations of covariates. Twenty teams found a statistically significant effect. The authors also controlled for researchers' prior believes and experience as well as peer-rated analysis quality, none of which accounted for the variation of results. Botvinik-Nezer et al. (2020) made similar observations. Crowdsourcing projects excels at emphasising the substantial impact of different analytical strategies, but they are also extremely time and personnel intensive. The two mentioned studies lasted between 2 to 3+ years, and included 61 and 180 analysts, respectively. Additionally, despite covering a more diverse set of analytical strategies than a single conventional analysis, crowdsourcing is limited by the number of participating teams. It, therefore, still only covers a few selected strategies, not all possible ones.

Silberzahn et al. (2018) and Botvinik-Nezer et al. (2020), therefore, suggest using multiverse analysis (also known as specification-curve analysis) as an alternative to crowdsourcing. This approach requires identifying and running all plausible analytical strategies. Plausible strategies are defined as statistically valid, non-redundant, tests appropriate to the research question. Their aggregated results are then used to make inferences about the research question. Despite being statistically more complex and computationally more intense, this approach has the advantages of only needing one, or ideally a couple, researchers. Moreover, a researcher's inherent strategy bias is neutralised and noise is made transparent (Simonsohn et al., 2020). Multiverse approaches are therefore well suited for assessing the scope of variation induced by different

analytical strategies. However, to my knowledge, there is one aspect that has been neglected thus far. That is to what extend does the identified scope of variation cover the range of possible outcome. If the results space and its structure is unknown, it is impossible to tell the observed results were coincidental or in fact point to a true latent construct. The concept of statistical association provides a suitable analogy. Covariation and correlation both indicate the direction of relationships (positive, null, negative), however, only correlation allows for assessing the association's strength. Correlation is the standardised version of covariation, and only knowing the its limits gives the correlation coefficient meaning. The same principle applies the scope of variation and the range of all possible results.

Current project

Taken together, researchers make many analytical decisions throughout conducting a study introducing so-called *researcher degrees of freedom*. Crowdsourcing analysis presents a strategy to understand their induced variation, but is impractical due to its time and personnel requirements, and is still limited by the number of participating teams. Multiverse analysis present a promising, low-personnel, but computationally intense alternative to assessing the scope this variation. Despite observing substantial variation in crowdsourced data analysis projects, little is known about the extend to which these approaches cover the full range of possible results. Determining the full results space and its underlying structure is an important exercise to understand the impact of analytical strategies. To my knowledge no previous study has attempted to define and understand a study's full results space. In this project, running all possible analytical strategies, I, therefore, aim to define and understand the results space of Silberzahn et al. (2018). By doing so, I hope further the understanding of how an analytic strategy affects its result.

Analysis

Analysis plan

The project's objective is to define the results space and structure of Silberzahn et al. (2018) to draw conclusions about the mechanisms of multiverse analysis. The results space refers to the numerical interval between the lowest and highest possible outcome. It is created by running every possible analytical strategy. Hence, it is closely related to assessing an effect's robustness. Robustness refers to an effect's consistency under different model specifications. Patel et al. (2015) developed a standardised approach to assessing an effect's robustness. I will therefore use this standardised approach to define the results space. The approach essentially comprises two parameters: the statistical models and the covariates (or control variables). In Silberzahn et al. (2018) the analysts used numerous different statistical models like multiple linear regression, mixed-model logistic regression or Bayesian logistic regression. In total there were 29 different modelling approaches. Additionally, each team used a different set of covariates. Across all teams there 15 covariates

used. This gives 2^n i.e., $2^{15} = 32,768$ possible combinations of covariates. Adding all modelling possibilities to the equation gives a total of $29 * 2^{15} = 950,272$ combinations to run. This number exceeds the project's scope, hence, it needs to be reduce to a more manageable count. I am, therefore, focusing on one modelling approach. In particular, on the approach that produced the median outcome of all analyses. The median, being the middle number of a given set of values, is a reasonable starting point in order to estimate the results space. Nevertheless, even focusing on one analytical approach leaves $1 * 2^{15} = 32,768$ possibilities. Due to computational limitations I will run a sample of 1,000 combinations.

In Silberzahn et al. (2018) the median outcome was produced by Stafford et al. (2014) which will hereinafter be referred to as *team 23* due to it's designated team number in the original study. *Team 23* first transformed the data and then conducted a mixed-model logistic regression. The first step of this project will be replicating the transformation and the analysis. Replicating other researchers analyses has the benefit of checking their work and, if results are indeed replicated, increasing confidence in them. It also ensures that this project has the same starting point as *team 23* did. Given the team made all their scripts publicly available I expect this step to be straightforward. Next is drawing a random sample of covariates, without replacement. "Without replacement" ensures all covariate combinations are unique in the sample. The covariates will be appended to the base (or core) variables. Base variables are those variables that are primarily assessed to answer the research question. In this case whether a football player's skin colour affects the odds of being sent off the field. *Team 23* defined two interaction terms as the base: "skin tone X implicit bias" and "skin tone X explicit bias." (The variables are described in the "data" section.) Hence, all models will have the following structure:

$$RedCard = SkinTone * ImplicitBias + SkinTone * ExplicitBias + CovariateCombination_i$$

While running each model, the relevant statistics will be extracted. Those are the coefficient (or effect) of skin tone, its standard error, test-statistic and p-value. Just as *team 23* did I will then calculate the 95% confidence intervals (CI). The estimates and their CIs are then transformed to odds ratios (OR) through exponentiating them to the power of two. Odds ratios quantify the strength of association between two variables. If greater than one the dependent variable is more likely to occur given the independent variable, if lower than one it less likely.

These odds ratios will be visualised to assessed through a conventional specification curve plot, a rain-cloud plot and scatter plot. The following describes the three graphs in more detail:

(i) The specification curve plot was developed to be a descriptive plot i.e., raw outcome data being plotted without any aggregation done. Its objective is to describe the results space while allowing the reader to identify the model specifications for each outcome (Simonsohn et al., 2020). It therefore has two vertically stacked components. The top part shows a sorted scatter plot. Its horizontal axis lists the

specifications and its vertical axis displays the outcome measure. The scatter points are sorted from lowest to highest outcome measure. This way the lowest outcome measure is on the bottom left corner and the highest in the top right one. The bottom part of the plot represents a table. As in the top plot the horizontal axis lists the specifications, the vertical axis lists the covariates. This arrangement allows each cell to specify the presence or absence for each covariate in a given specified model. For the top and bottom plot to work together, it is therefore imperative that the horizontal axes are identically arranged. If so, for every point in the top plot (i.e., for every outcome) the specified covariates can be seen in the bottom plot. For an example check Figure 2. As the current project has 1,000+ specifications it gets very hard to identify specific covariates due to the space limitations. Hence, I to had come up with different more intuitive ways of visualisation.

(ii) The rain-cloud plot was developed by Allen et al. (2021). Its objective is to give an unbiased, transparent view on the raw data. It combines a density, box and scatter plot. If stacked vertically from top to bottom, respectively, they look like a cloud with rain drops, hence, its name. The advantage of this plot is to be able to assess the raw data including key statistics (median, 25th and 75th quartiles and CIs) as well as the probability density all in one succinct plot. The current project seeks to define and describe a results space, the rain-cloud plot is, therefore, the ideal choice of graphs. (For an example check Figure 3.) Unfortunately, the plot does not allow to identify the covariates giving rise to the outcomes or their specified effects. As the project also seeks to understand the structure of the results space it important to get insights about the specified covariates effects. Therefore a third plot is developed which is to be assessed in tandem with this plot.

(iii) The third plot is a sorted scatter plot. While there is nothing special about the plot itself its content is important. Given the large number of models, it does not make sense to assess each model and its effects in individually. Instead the effects for each covariate need to be aggregated. The goal of this plot is to visualise the specified effect for each covariate. Prior to doing so, the covariates first need to be aggregated and their specified effects need to be calculated. To this end, each covariate is recoded into a binary factor: included in the model yes/no. These newly defined factors are than used as an independent variables in an ANOVA. The previously calculated odds rations are used as dependent variables. The estimates of the fitted model are the specified effect for each covariates. These are visualised similarly to the top part of the specification curve plot. The covariates are on the horizontal axis, will the odds ratios are on the vertical axis. Here too, the lowest outcome is in the bottom left corner and the highest outcome in the top right one. Additionally the point are coloured indicating a statistically significant effect (or non-significant).

Taken together, the current project's objective is to define the results space and structure of Silberzahn et al. (2018) to gain a better understanding of how analysis strategy affects analysis results. The results space is created by running all possible model and covariate combinations. Applied to Silberzahn et al.

(2018) this means running roughly 1M models. Due to time and computational limitations, I am focusing on the model that produced the median outcome of all analytical strategies. Hence, I will run a mixed-model logistic regression with 1,000 randomly sampled covariate combinations. For each model relevant statistics will extracted and visualised. The conventional specification curve plot does not deal well with a large number of model specifications. Therefore, I will produce a rain-cloud and scatter plot, which visualise the results space (and its distribution) as well as the specified covariate effects, respectively. For the latter, I will first have to run an additional statistical model to calculate the specified effects. These graphs will give insight into the results space, its building blocks and ultimately into the mechanisms of multiverse analysis.

Code

This section provides a brief overview of the code and explains the reasoning behind it. As mentioned in the analysis section the first step was replicating the data transformation and analysis of *team 23* (Stafford et al., 2014). The team made their project folder publicly available. It contains three scripts relevant to this project: the data exploration, transformation and analysis. After duplicating their project folder on my local hard drive all scripts run without issues (only the working directories needed adjustment). The data exploration included the reasoning behind transforming the data and some cleaning. The data transformation restructured the data into a more intuitive format (more details on the topics of exploration and transformation are in the data section). The analysis script prepared the data by assigning variables classes (factors, numerical or boolean) and standardised a few variables i.e., they were centred around the mean. Finally, the models were specified. The team ran both frequentist and Bayesian models, however, this project covers the frequentist approach, only.

My code is based on Haessler et al. (2020) and Patel et al. (2015). The rough structure of the code is shown in the pseudocode table (see Algorithm 1). Here, I am briefly motivating two analytical choices I made. For a closer look please check the commented code itself:

(i) Choosing the function to run the statistical model. *Team 23* ran a mixed-model logistic regression. This model has two relevant components to this explanation: fixed effects and random effects. Fixed effects are those that are consistently observed in different situations because the construct is of direct interest to the research question. In this case, it is, for example, skin tone. Independent of the player, the game or the league skin tone is always observed. Their counter parts are random effects which change between situations. A player, for example, is just one "unit" a measurement, and is himself not directly relevant to answering the research question. The function used by *team 23* requires to specify a random effect. However, given the project seeks to sample from all possible covariate combinations, a random effect is not always included. Hence, the "non-random" counter part to this function needs to be utilised. For this project, it was important to check that both functions use the same estimation method (maximum likelihood or restricted maximum likelihood estimation) to calculate the statistics.

It is important to check estimation method to ensure the outcomes are comparable.

(ii) No multiple comparison. This is based on the fact that in statistics null hypotheses are one rejected if the probability of finding a false positive is below 5% (known as the significance level, α). The more statistical tests a run, the more likely it is that one of those tests is indeed a false positive. Hence, the are techniques that account of this issue. It is usually good practice to account for those "multiple comparisons." Here, I refrained from doing so though. This project seeks to simulate multiple researchers running different models. Therefore, for the the purpose of this project all models are run independently by individual researchers. Those would not have knowledge of the other analyses and, hence, would not account for them. To maintain the highest possible ecological validity, I therefore did not account of multiple comparisons.

Algorithm 1: Pseudocode overview–in progress

Data: Prepared data based on team 23 script
variables ← Define dependent, base variables and covariates
specifications ← Use *variables* to create matrix containing all possible covariate combinations
formula ← Paste *specifications* by row and append as column to *specifications*
formula.ranef; *formula.ef* ← Separate formulas based on including a random-structure
samlple.ranef; *sample.ef* ← Random, without replacement, sample from *formula.ranef*; *formula.ef*
wrapperfunction ← Extracting model statistics and stores them in a list
tapply

```
initialization
while not at end of this document do
    read current
    if understand then
        go to next section
        current section becomes this one
    else
        go back to the beginning of current section
    end
end
```

Result: List containing model statistics

The code is written in R (Version 1.4.1103) on macOS Big Sur (Version 11.4) and can be retrieved from my GitHub repository. The code from *team 23* (Stafford et al., 2014) can be retrieved from their OSF repository. Following R packages were used *here 1.0.1* (Müller, 2020), *data.table 1.14.0* (Dowle & Srinivasan, 2021), *tidyverse 1.3.1* (Wickham et al., 2019), *lme4 1.1-27.1* (Bates et al., 2015), *pbmcapply 1.5.0* (Kuang et al., 2019), *PupillometryR 0.0.3* (Forbes, 2020) and *cowplot 1.1.1* (Wilke, 2020).

Data description and preparation

The data was retrieved from Silberzahn et al. (2018). It contains information about football players, their encounters with referees and the received cards (yellow, yellow/red and red). Moreover, a player's position, age, club, league country, victories, ties, defeats and goals. In addition, a skin tone rating based

on two independent raters was included. The referees are numerically coded to protect their identify. The referees origin countries are also included as well as an implicit and explicit racism bias scores for their respective countries. The exploratory data analysis (EDA) of *team 23* showed that each row of the dataset represents a unique player-referee combination listing all their encounters as well as the couple's total number of received/assigned cards. *team 23* states that it preferred a different data format. More specifically, a format where each player-referee encounter is reflected by one row. This way each encounter had a maximum of one red card. To achieve this format the data had to be transformed more specifically, disaggregated. Further EDA showed that receiving a red card is highly unlikely, 0.008%, i.e., data is highly skewed. This property informed the teams decision to use a type of logistic regression (a statistical modelling technique equipped to deal with skewed/binary distributions). Finally, the team excluded all referees that did have at least 22 encounters. Their reasoning was a game includes (at least) 22 players, 11 players per team, if this were not the case referee most likely did not officiate in one for the four leagues of interest (England, Germany, Italy and Spain). This step excluded roughly 66% of the referees but retained 97.4% of the cases. The final dataset used for the current project contained 335,537 observations and 19 variables. Figure 1 shows the properties of the variables relevant to current project's analysis

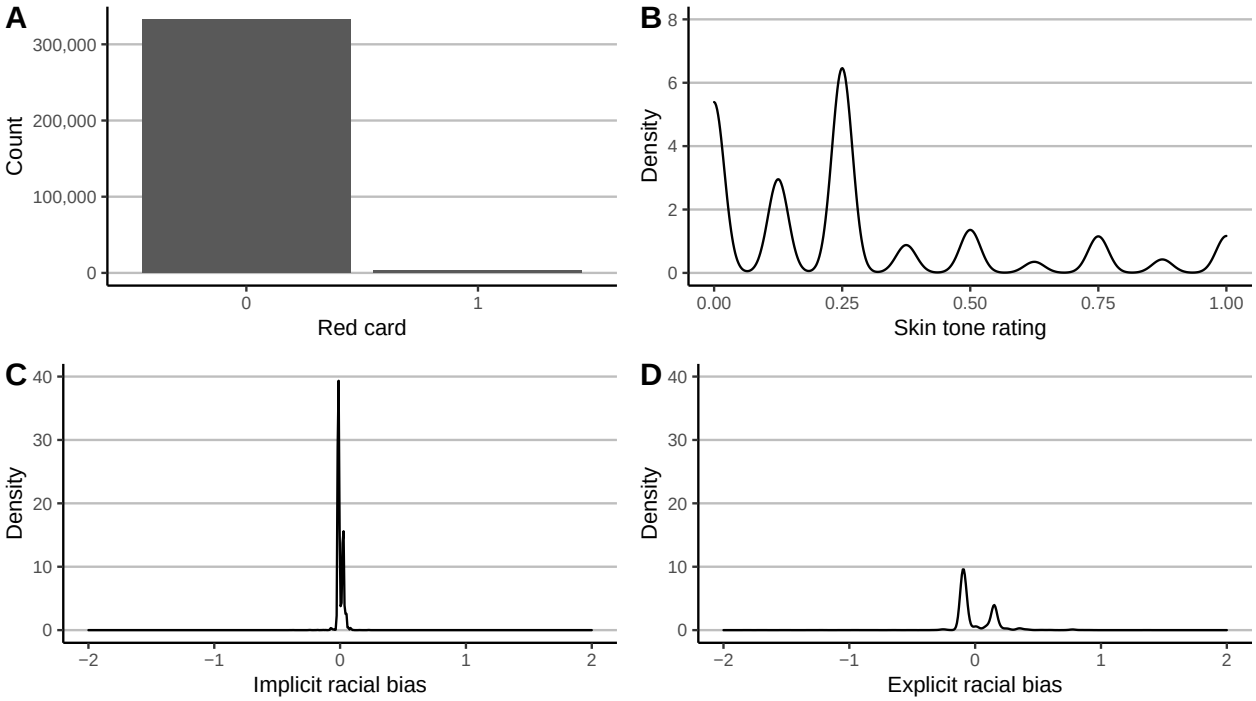

Figure 1: A) Showing the strong skew of receiving no red to receiving a red card. B) Showing most players had a lower skin tone rating i.e., most players were on the whiter side. C) Showing there is not much variation in implict racial scores. D) Showing more variation in explitic reacial bias scores. The latter two were centred around the mean, hence, zero does not mean there is no bias.

Results

Replicating *Team 23* analysis

As mentioned in the above section the first steps were replicating the *team 23* analysis using their provided scripts. The results are reproducible without any adjustment. Skin tone has a significant effect on the odds of being send off the field ($OR = 1.311$, 95%CI [1.099, 1.563], $p = 0.003$). This means while keeping all other variables constant for every unit increase in skin tone rating (darker skin tone), the odds of being send off the field increase by roughly 131%. The interaction terms of skin tone and implicit racial bias ($OR = 0.004$, 95%CI [0.000, 23.259], $p = 0.211$) as well as skin tone and explicit bias ($OR = 1.837$, 95%CI [0.493, 6.848], $p = 0.365$) are both non-significant. In accordance with the original analysis, there are also significant differences between leagues and positions as well as implicit racial bias scores. Explicit bias scores are on verge of being non-significant. This analysis only aimed at testing reproducibility, which proved to be true, further results details are forfeited.

Specification curve

Shown in Figure 2. The horizontal axes of the top and bottom chart show the specifications sorted based on their odds rations from lowest to highest. The odds ratios themselves are shown on the vertical axis of the top plot. The points in the top plot, which look like a line, each represent the outcome of one statistical model. Black points refers to statistical significant, red to non-significant outcomes. Overall are 89.7% of the outcomes significant. Except for the lowest outcome ($OR_{lowest} = 1.080$) are all points close to the their respective neighbours. This indicates that none of the specified models are responsible for a sudden increase in the outcome measure. The relatively smooth line suggest that sampling 1,000 specifications is, in fact, representative for all 33k models. The points can be divided into roughly three sections all non-significant ($\leq 57^{th}$), mixture of (non-)significance ($58^{th} - 247^{th}$) and all significance ($\geq 248^{th}$). The bottom table lists all covariates on its vertical axis, sorted from most positive to most negative impact. For example, player has the strongest positive impact, while club has the strongest negative impact. Each column of the table represents one mode. A coloured cell indicates the presence of the variable, an uncoloured cell its absence. Just like a above, black indicates the outcome measure was significant, red non-significance.

The top and bottom plot work in tandem. Each outcome (i.e., point) in the top corresponds to the indicated covariates in the bottom table. Regular specification curve analyses focus on all *reasonable* specifications, hence, their number of specifications is much lower. Keeping the intended purpose in mind one would, in fact, be able to tell the different model specifications apart. However, the current project's goal was to define the results space, which requires *all possible* specifications. The specifications are therefore huddled too closely which does not allow the plot to functions in its supposed way. What is interesting though are the top and bottom row of the table. The top row shows that player is particularly often

included in models with a higher outcome. Conversely, club is particularly often included in models with a lower outcome. Full disclosure, I only noticed this pattern after having calculating the specified covariate effects.

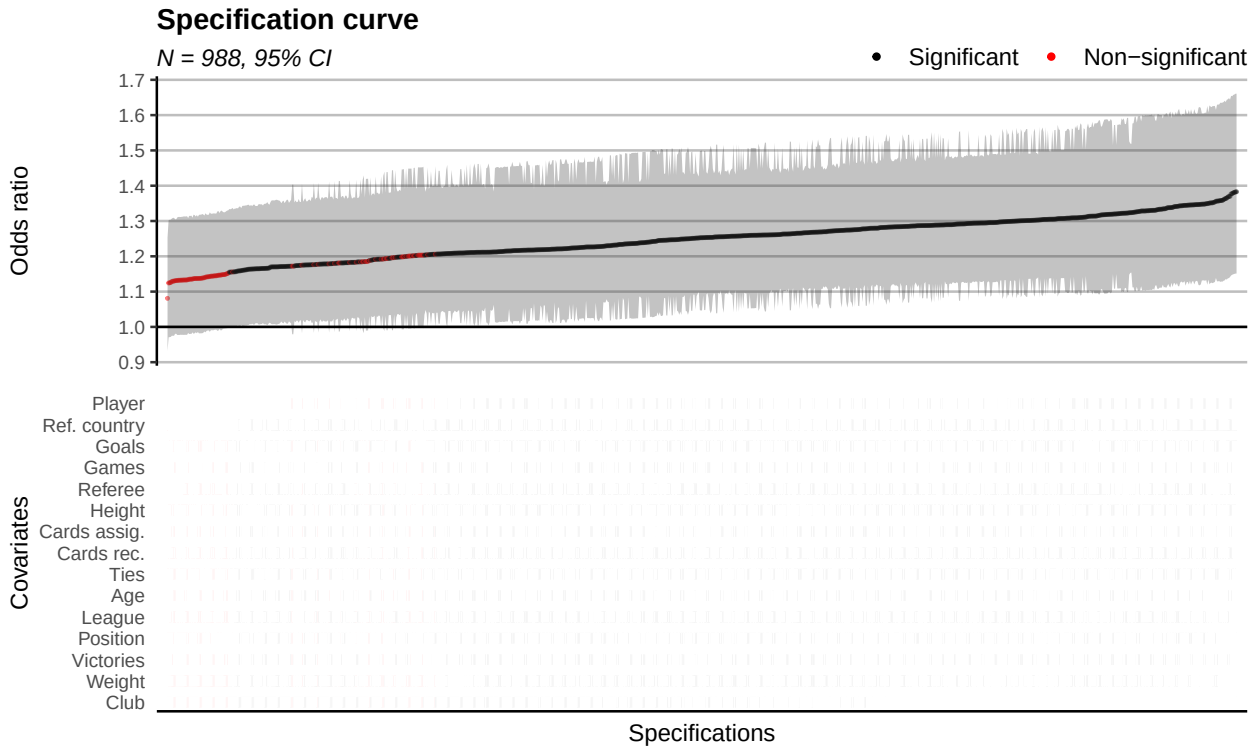

Figure 2: Convential specification curve. Each outcome (i.e., point) in the top corresponds to the indicated covariates in the bottom table. Due to the large number of specifications, this plot is not as informative as it is supposed to be. It is, therefore, merely making an argument for needing and alternative.

Results space

Shown in Figure 3. The horizontal axis shows the odds rations. From top to bottom the three component show the probability density distribution, the box plot and the raw data, respectively. The dashed line indicates the original results of *team 23*, which has also been the median outcome of all analytical strategies in Silberzahn et al. (2018). Based on the raw data points, the results space can be defined as the interval between 1.081 and 1.383. The 1st quatile equals to 1.206, the median equals to 1.248 and the 3rd quartile equals to 1.293. Therefore, as also evidenced by the probability density distribution, 50% of the lay withing a very narrow interval (between 1st and 3rd quartile). The outcome of *team 23* (1.31) is still included in the interval, but it is outside the middle 50% of the data.

The probability density distribution also has an interesting shape, having two peaks. This suggests a particular covariate combination structure causing two focal points. Based on this graph alone though it is not possible to decipher the underlying structure though. What also seems to be more apparent in this graph than in the specification curve is that the non-significant values also seem to revolve round to focal points.

One being around 1.13 and the other around 1.19. This also suggest some influential covariate or covariate structure that is cause non-significance.

Concering the results space, for transparency purposes, it also needs to be mentioned that eight points were excluded from all visualisation. Four of them were "infinite," one effectively approaching infinity, one being in the lower million range, one being 150, and a final being 4.6. I specifically chose not to include the latter one in the graph as it would despite is relatively close numerical proximity skewed the distribution, at least visually.

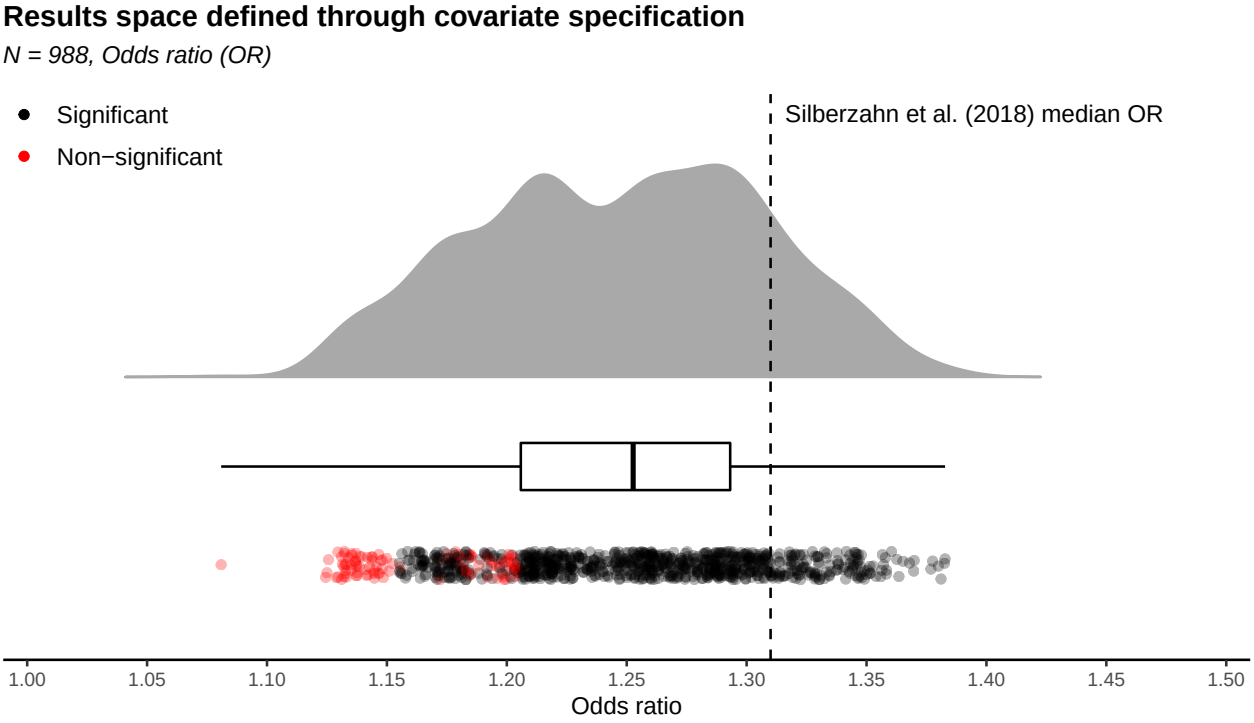

Figure 3: The rain cloud plot is comprised of three parts. The probability density distribution, the box plot and the raw data points. This allows for an transparent and bias assessment of the results space and its distribution.

Specified covariate effects

This graph shows the specified covariate effects. On the horizontal axis are the covariates, on the vertical axis the estimates. The covariates are sorted from negative to positive impact. Black indicates a significant effect, red non-significant. The error bars represent the 95% CI. Four out of 15 covariates are significant, roughly 27%. The significant covariates' effects seem relatively weak, ranging from -0.1 to 0.05, only. Interesting is which covariates are significant. The *team 23* analysis included league, position, referee and player as covariates. Of those only player is a significant covariate. Using crude mathematics, one might make the case that the added effect of players is the reason for moving the overall effect from the median, roughly 1.25 to 1.30. Though that's more speculation than causal reasoning. If

anything, this graph shows that the individual covariate effect is likely not as influential as interactions of covariates. As mentioned above, the roughly 90% of all outcome were statistically significant, given that only 27% of the covariates have an overall significant effects suggest that the effect is of skin tone itself is fairly robust.

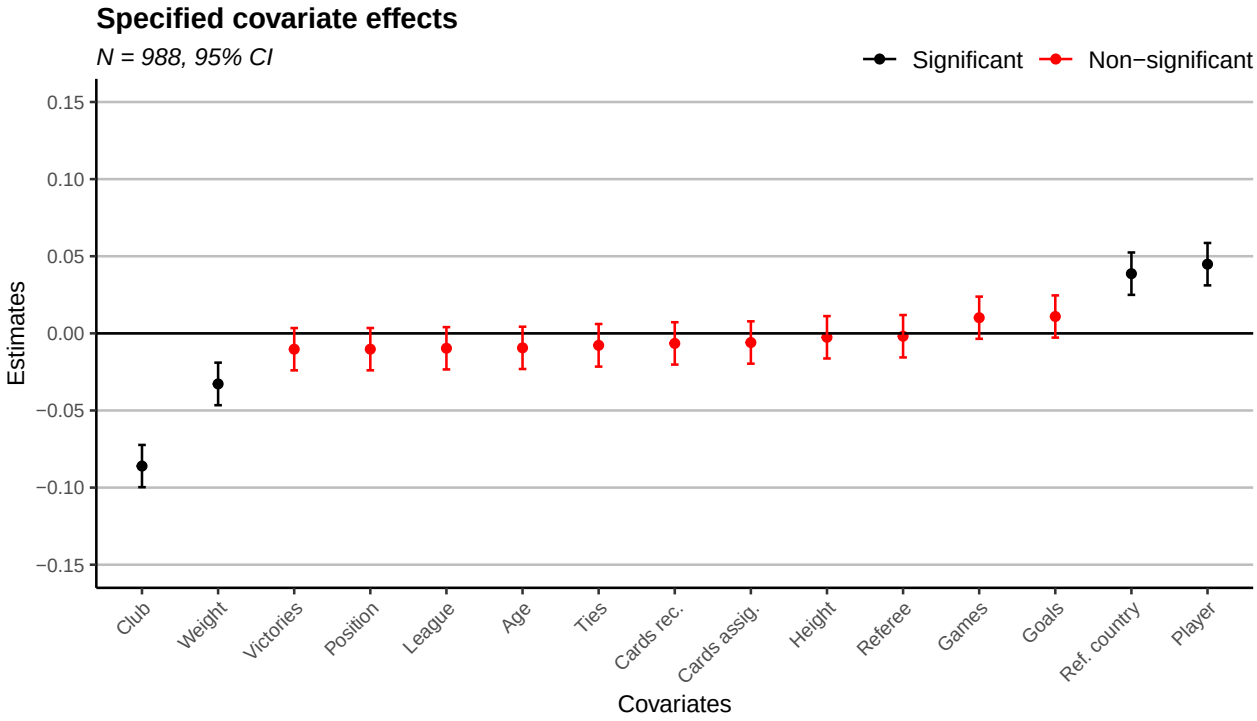

Figure 4: The specified covariate effects indicate the extend to which a covariate influences the outcome measures.

Systematic latent structure

This graph was not planned and yet it is too interesting not to include it in this report (though it will only serve to emphasise previous points and probe future research). I was very keen to investigating the underlying covariate structure assumed based on 3. I, therefore, went back to the previous mentioned Patel et al. (2015) study. Despite having a different objective, their methodology is similar to the current project's (rather the other way around). They visualised the results as so-called volcano plot (see Figure 5) which show the relationship between the p-value and effects size. So, I was interested how the graph would look like in the current project's case. It turns the graph evidences two distinct underlying constructs–perhaps even three. Particular salient is how distinct these "stream" are. Given that the specified covariate effects are relative weak, it is highly unlikely that only one is responsible for cause such structure. In fact, it is highly likely that a combination of covariates in play here. Considering the apparent influence of this deterministic structure is for future research to decipher the underlying covariate combination.

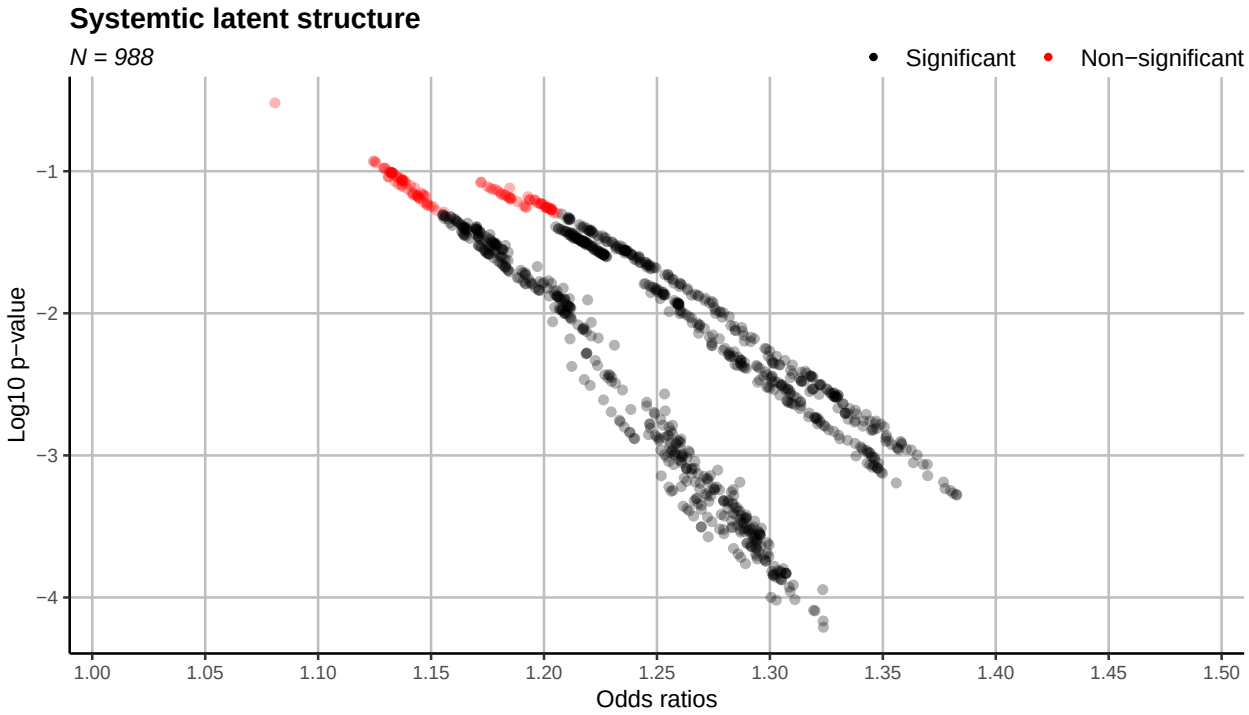

Figure 5: Volcano plot showing the the relationship between p-values and effect sizes. The degree of strictness in which the group are seperated suggest a strong deterministic latent covariate structure.

Discussion

The current project aimed at defining and understanding the full results space of Silberzahn et al. (2018). The rational was if the full results space is known, it possible to better estimate the extend to which analytical strategies influence their results. To this end, out all analytic strategies, every possible combination was supposed to be run in to determine the full results space. Due to time and computational limitations, the current project focused on one analytical strategy and sampled 1,000 covariate combinations out of all possible one. The chosen analytical strategy (i.e., the statistical model) was the one producing the median outcome of the original study.

This very analysis was first attempted to be fully replicated, yielding the identical results as the original team, *team 23*, did. This ensured an identical starting from all further analyses. After drawing the 1,000 unique combinations of covariates and running the respective models, the results were visualised. Removing unrealistic values (anything above 100 OR), yielded a relatively narrow results space ranging from roughly 1.1 to 1.4. Though my analysis showed one additional value of roughly 4.5, its plausibility is it to be determined. The outcome of 1.31 found by *team 23* was not included in the interval between the 1st and 3rd quartile, though its was part of the 95% confidence interval. It is, therefore, concluded that its specified covariates have an above averages effect i.e., are more influential than a random combination of covariates would have had. The distribution of the results space showed two peaks. To understand the distribution's underlying structure

the specified covariate effects be analysed. The results showed that 4 out of 15 covariates are statistically significant, though their effects are small. Given the outcome measures showed values that are beyond the apparent influence of a single covariates it is, therefore, assumed that specific covariate interactions have significant effect on the outcome. The final volcano plot supports this hypothesis by clearly outlining two (perhaps a weak third) structure. Unfortunately, is identifying these underlying structures beyond the scope of this project.

Taken together, considering the roughly 90% of all outcomes are significant it is no surprise that the original findings by *team 23* significant. In fact, it would have been more surprising if they had not found a significant effect. That being said, the effects is still strong than a seemingly random combination of covariates. The pressing question, therefore, is how did the authors arrive at this specific effect. IN PROGRESS.

All materials used for this analysis are publicly available and all results should be reproducible and expendable. The biggest constraint of this project were time and computational resources. However, reasonable decisions were made to achieve the best possible compromised between rigour and feasibility. The constraints and the findings of the current project both yield great potential for futures research. The first step would running all possible covariates combinations, not just a sample. It would highly interesting if the observed patter is facilitated or vanishes. Moreover, it would be highly interesting to decipher the underlying covariate structure that is responsible the observed "streams." Based on the present findings I would assumes there is not a single covariate but rather a combination of them responsible. Finally, running not mixed-model logistic regressions but all other models used in the original study by Silberzahn et al. (2018).

References

Allen, M., Poggiali, D., Whitaker, K., Marshall, T., Langen, J. van, & Kievit, R. (2021). Raincloud plots: A multi-platform tool for robust data visualization [version 2; peer review: 2 approved]. *Wellcome Open Research*, *4*(63). https://doi.org/10.12688/wellcomeopenres.15191.2

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Botvinik-Nezer, R., Holzmeister, F., Camerer, C. F., Dreber, A., Huber, J., Johannesson, M., Kirchler, M., Iwanir, R., Mumford, J. A., Adcock, R. A., Avesani, P., Baczkowski, B. M., Bajracharya, A., Bakst, L., Ball, S., Barilari, M., Bault, N., Beaton, D., Beitner, J., … Schonberg, T. (2020). Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*, *582*(7810), 84–88. https://doi.org/10.1038/s41586-020-2314-9

Carp, J. (2012). On the plurality of (methodological) worlds: Estimating the analytic flexibility of fMRI experiments. *Frontiers in Neuroscience*, *6*, 149. https://doi.org/10.3389/fnins.2012.00149

Collins, F. S. (2021). COVID-19 lessons for research. *Science*, *371*(6534), 1081–1081. https://doi.org/10.1126/science.abh3996

Dowle, M., & Srinivasan, A. (2021). *Data.table: Extension of 'data.frame'*. https://CRAN.R-project.org/package=data.table

Forbes, S. (2020). *PupillometryR: A unified pipeline for pupillometry data*. https://CRAN.R-project.org/package=PupillometryR

Haessler, T., Ullrich, J., Bernardino, M., Shnabel, N., Van Laar, C., Valdenegro, D., Sebben, S., Tropp, L. R., Visintin, E. P., Gonzalez, R., Ditlmann, R. K., Abrams, D., Selvanathan, H. P., Brankovic, M., Wright, S., von Zimmermann, J., Pasek, M., Aydin, A. L., Zezelj, I., … Ugarte, L. M. (2020). A large-scale test of the link between intergroup contact and support for social change. *Nature Human Behaviour*, *4*(4), 380–386. https://doi.org/10.1038/s41562-019-0815-z

Ioannidis, J. P. A. (2008). Why most discovered true associations are inflated. *Epidemiology*, *19*(5), 640–648. https://doi.org/10.1097/EDE.0b013e31818131e7

Kuang, K., Kong, Q., & Napolitano, F. (2019). *Pbmcapply: Tracking the progress of mc*pply with progress bar*. https://CRAN.R-project.org/package=pbmcapply

Müller, K. (2020). *Here: A simpler way to find your files*. https://CRAN.R-project.org/package=here

Patel, C. J., Burford, B., & Ioannidis, J. P. A. (2015). Assessment of vibration of effects due to model specification can demonstrate the instability of observational associations. *Journal of Clinical Epidemiology*, *68*(9), 1046–1058. https://doi.org/https://doi.org/10.1016/j.jclinepi.2015.05.029

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., Bahnik, S., Bai, F., Bannard, C., Bonnier, E., Carlsson, R., Cheung, F., Christensen, G., Clay, R., Craig, M. A., Rosa, A. D., Dam, L., Evans, M. H., Cervantes, I. F., ... Nosek, B. A. (2018). Many analysts, one data set: Making transparent how variations in analytic choices affect results. *Advances in Methods and Practices in Psychological Science*, *1*(3), 337–356. https://doi.org/10.1177/2515245917747646

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simonsohn, U., Simmons, J. P., & Nelson, L. D. (2020). Specification curve analysis. *Nature Human Behaviour*, *4*(11), 1208–1214. https://doi.org/10.1038/s41562-020-0912-z

Stafford, T., H.Evans, M., J.Heaton, T., & Bannard, C. (2014). *Crowdstorming team 23: There is definite racial bias in which soccer players are sent off, but its locus is unclear"*. https://osf.io/akqt4/

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., Aert, R. C. M. van, & Assen, M. A. L. M. van. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, 1832. https://doi.org/10.3389/fpsyg.2016.01832

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T. L., Miller, E., Bache, S. M., Müller, K., Ooms, J., Robinson, D., Seidel, D. P., Spinu, V., ... Yutani, H. (2019). Welcome to the tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Wilke, C. O. (2020). *Cowplot: Streamlined plot theme and plot annotations for 'ggplot2'*. https://CRAN.R-project.org/package=cowplot