**How analysis strategy affects analysis results**

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

Sebastian Ploner, [sploner1@sheffield.ac.uk](mailto:sploner1@sheffield.ac.uk), University of Sheffield

**ABSTRACT**

# Introduction

## Theoretical framework

     The Covid-19 pandemic has reaffirmed how important sound scientific research is for political and societal decision-making (Collins, 2021). Rigorous research builds upon a systematic and well-reasoned approach to solving a research problem. Based on current knowledge researchers define a research question and develop a hypothesis. To test the hypothesis, they design and conduct a research study which yields data. To draw conclusions from the data, they apply statistical models and assess how different variables have influenced the data. Each of these steps is influenced by the researchers’ decisions, which are known as *researcher degrees of freedom* (RDF, Simmons et al., 2011; Wicherts et al., 2016). In most cases, there is not only one but many feasible analysis strategies to answer a research question (Carp, 2012). This often creates uncertainty and inconsistency. For instance, researchers are often uncertain about which covariates to include and how to model them, which leads to inconsistent findings (Ioannidis, 2008; Patel et al., 2015). Recently, efforts have been made to better understand how different analysis strategies influence research results.

     Crowdsourcing is one approach to better understand the influence of analysis strategies on research findings (Botvinik-Nezer et al., 2020; Silberzahn et al., 2018). A large number of researchers team up into smaller, independent groups to investigate the same research question based on the same dataset. For instance, in Silberzahn et al. (2018) 29 teams investigated the effects of a football player’s skin colour on the odds of being sent off the field. The variation between the analysis strategies was substantial. There were 29 different analyses with 21 different combinations of covariates. Twenty teams found a statistically significant effect of skin colour on the risk of being sent off the field. The authors also controlled for researchers’ prior believes and experience as well as peer-rated analysis quality, but these factors did not account for the variation of results. Botvinik-Nezer et al. (2020) made similar observations. Such crowdsourcing approaches can excellently show how different analysis strategies influence research findings, but they are extremely time and resource consuming. For instance, it took two and more than three years and 61 and 180 analysts, respectively, to perform the two aforementioned studies. Furthermore, although crowdsourcing approaches can cover more analysis strategies than a single conventional analysis, they are limited by the number of teams. They therefore cover a certain selection but not all possible strategies.

     Another approach to assess the effects of analysis strategies on research findings is multiverse analysis (also known as specification-curve analysis). This approach has been proposed by Silberzahn et al. (2018) and Botvinik-Nezer et al. (2020) as an alternative to crowdsourcing. In this approach, all plausible analysis strategies are identified and performed. Plausible strategies are defined as all statistically valid, non-redundant tests that are appropriate for the research question. The aggregated results of all plausible analysis strategies are used to make inferences about the research question. Although the approach is statistically complex and computationally intense, it can be performed by one or a few 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 to better understand how different analysis strategies influence research results.

Multiverse analysis also presents an intriguing method to investigate questions posed by crowdsourcing approaches. Silberzahn et al. (2018) overserved many different analysis strategies and results. If any of these analyses were conducted as an induvial study, there would have been a 69% chance of observing a significant effect. What are the implications though? Are all analytical approaches unreliable? Or is 69% consensus sufficient to make a conclusive decision? To answer these questions, the variability between the analytical strategies, first, needs to be better understood. Crowdsourcing and multiverse analysis usually aim at scrutinising *reasonable*, never on all *possible* analytical approaches (which makes sense given the studies’ objectives). However, to better understand the variability between analytical strategies it would be interesting to know how the outcomes of all *reasonable* relate all *possible* analyses. This way an intuition would be established to what extend the observed results represent the full results space.

## Current project

     Taken together, researchers make many analytical decisions throughout a study which introduces *researcher degrees of freedom*. Crowdsourcing analysis represents one strategy to observe the variation induced by different analysis strategies. It is, however, extremely time- and resource-consuming and limited by the number of participating teams. Multiverse analysis represents a promising, low-personnel, but computationally intense alternative to assess the variation induced by analysis strategies. However, little is known about the extent to which these approaches cover the full range of *possible* results. Thus, determining the full results space and its underlying structure is important step to understand the impact of analysis strategies on research results. To my knowledge no previous study has attempted to define and understand the full results space of a study. In this project, I therefore apply all possible analysis strategies to define and understand the results space of Silberzahn et al. (2018). Furthermore, I propose ways to visualise the space and its structure in an intuitive and succinct way. This effort does not seek to be a conclusive project, rather it aims at laying a foundation for further studies into the variability of analytical strategies and results.

# Analysis

## Analysis plan

     The project’s objective was to define the results space and structure of Silberzahn et al. (2018). The results space refers to the numerical interval between the lowest and highest possible outcome. It is created by running every possible analysis strategy. Hence, it is closely related to assessing the robustness of an effect. Robustness refers to the consistency of an effect under different model specifications. Patel et al. (2015) developed a standardised approach to assess an effect’s robustness. I used 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, 15 covariates were used. These yield i.e., possible combinations of covariates. Adding all modelling possibilities to the equation yields a total of possible analysis strategies. As this number of strategies exceeded the resources of the present project, I reduced it to a more manageable number. I therefore focused on one modelling approach. I specifically focused on the modelling 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 to estimate the results space. However, even focusing on one analytical approach leaves possibilities. Due to computational limitations, I therefore performed a random sample of 1,000 analysis strategies.

     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 its 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 the present project was to replicate their transformation and analysis. Such a replication can increase the confidence in the previous and the present approach. Moreover, it ensures that this project has the same starting point as *team 23*. As *team 23* made all scripts publicly available this step was straightforward. The next step of the project was to define the results space and its structure. To this end, I drew a random sample of covariates without replacement. “Without replacement” ensured that all covariate combinations were unique in the sample. The covariates were appended to the base (or core) variables. Base variables were those variables that were primarily assessed to answer the research question. In this case the research question was 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 had the following structure:

The relevant outcome parameters of each model were extracted. These parameters were the coefficient (i.e., effect) of skin tone, its standard error, test-statistic and p-value. Similar to *team 23* I calculated the 95% confidence intervals (CI). The estimates and their CIs were transformed to odds ratios (OR) through exponentiating them to the power of two. OR quantify the strength of association between two variables. An OR greater than one indicates that the dependent variable is more likely to occur given the independent variable, if it is lower than one it is less likely to occur. Eight OR outliers were excluded from visualisation. Four approached infinity, one was in the lower million range, one was 150, and a final one was 4.6. These graphs will give insight into the results space and its structure.

     These OR were visualised and assessed by three complementary plots, (i) a specification curve plot, (ii) a raincloud plot, (iii) and a scatter plot. The following describes the three graphs in more detail:

1. The specification curve is a descriptive plot which shows raw outcome data without any aggregation. 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 has two vertically stacked components. The upper component is 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 lower component of the plot is a table. Similar to the upper component the horizontal axis lists the specifications, while the vertical axis lists the covariates. This arrangement allows each cell to specify the presence or absence for each covariate in a given model specification. For the integration of information from the upper and lower components, the horizontal axes are identically arranged. Thus, for each point in the upper component (i.e., for every outcome) the specified covariates can be seen in the lower component (see Figure 2 for an example). However, with the more than 1,000+ specifications of the current project it was hard to identify specific covariates. Hence, I complemented the specification curve with other plots.
2. The raincloud plot (Allen et al. (2021)) combines a probability density plot, a box plot and a scatter plot to give an unbiased, transparent view of the raw data. If the three plots are stacked vertically from top to bottom, they look like an eponymous cloud with rain drops. The strength of the raincloud plot is that it visualizes the raw data, key statistics (median, 25th and 75th quartiles and CIs) and the probability density in a single plot. As the current project seeks to define and describe a results space, the raincloud plot is an ideal tool (for an example check Figure ). However, the plot does not allow to identify which covariates cause which outcomes. As the project also sought to understand the structure of the results space it was important to gain insights into how specific covariates influenced the outcomes. I therefore developed a third plot which complements and extends the raincloud plot.
3. The third plot was a sorted scatter plot. The goal of this plot was to visualise the effects of each covariate. Given the large number of models, it did not make sense to assess all models and their effects individually. I therefore aggregated the effects of each covariate. To this end, each covariate was recoded into a binary factor: included in the model yes/no. These newly defined factors were than used as independent variables in an ANOVA. The previously calculated OR were used as dependent variables. The estimates of the fitted model were the specified effect for each covariate. These were visualised similar to the top part of the specification curve plot. The covariates were on the horizontal axis, while the OR were on the vertical axis. The lowest outcome was in the bottom left corner and the highest outcome in the top right corner. Additionally, the colours of the points indicated the statistical significance of the effects.

     In summary, the present project aimed to define the space and structure of the results of Silberzahn et al. (2018). The space of possible results is defined by running all models with all possible covariate combinations. Applied to Silberzahn et al. (2018) this means running roughly 1M models. Due to time and computational limitations, I focused on one model that produced the median outcome of all analysis strategies. Hence, I ran a mixed-model logistic regression with 1,000 randomly sampled covariate combinations. For each model, relevant parameters were extracted and visualised. As the specification curve is not ideal for large numbers of model specifications I complemented and extended it with raincloud and scatter plots, which visualise the results space and the specified covariate effects, respectively. For the latter, I first ran an additional statistical model to calculate the specified effects.

## Code

     This section provides a brief overview of the code and explains the reasoning behind it. I first replicated 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 the present project: the data exploration, transformation, and analysis. After duplicating their project folder on my local hard drive all scripts ran 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 variable 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, the present project focused on the frequentist approach.

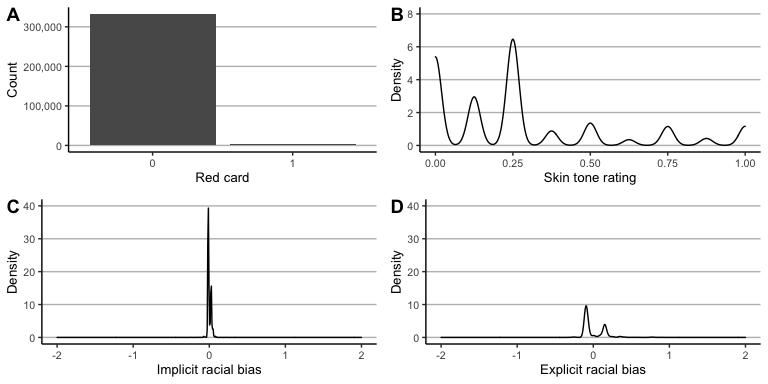
     My code was based on Haessler et al. (2020) and Patel et al. (2015). The rough structure of the code is shown in the pseudocode table (see Algorithm 1). Here, I briefly motivate two analytical choices I made. For a closer look please check the commented code itself:

1. Choosing the function to run the statistical model. *Team 23* ran a mixed-model logistic regression. This model has two relevant components: 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 counterparts are random effects which change between situations. A player, for example, is just one “unit” of a measurement, and is himself not directly relevant to answering the research question. The function used by *team 23* requires specifying a random effect. However, given the present project sought to sample from all possible covariate combinations, a random effect was not always included. Hence, the “non-random” counterpart to this function was used. For this project, it was important to check that both functions used the same estimation method (maximum likelihood or restricted maximum likelihood estimation) to ensure that the outcomes are comparable.
2. No multiple comparison. In statistics, it is common practice to reject a null hypothesis if the probability of finding a false positive is below 5% (known as the significance level, ). The more statistical tests are run, the more likely it is that one of the tests is a false positive. Hence, it is good practice to account for those “multiple comparisons.” Here, I nevertheless did not correct for multiple comparisons. This project sought to simulate multiple researchers running different models. Those researchers would not know the other analyses and, hence, would not account for them. To maintain the highest possible ecological validity, I therefore did not correct for multiple comparisons.

     The code was written in R (Version 1.4.1103) on macOS Big Sur (Version 11.4) and can be retrieved from my [GitHub repository](https://github.com/sebastianplnr/msc_dissertation_project). The code from *team 23* (Stafford et al., 2014) can be retrieved from their [OSF repository](https://osf.io/akqt4/). The 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 contained information about football players, their encounters with referees and the received cards (yellow, yellow/red and red). Moreover, it included a player’s position, age, club, league country, victories, ties, defeats, and goals. In addition, a skin tone rating based on two independent judges was included. The referees were numerically coded to protect their identify. The referees’ origin countries were also included as well as 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* stated that it preferred a different data 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, i.e., disaggregated. Further EDA showed that receiving a red card was highly unlikely (p=0.008%), i.e., the data was highly skewed. This property informed the team’s decision to use a logistic regression (a statistical modelling technique equipped to deal with skewed/binary distributions). A graphical overview of the properties of the core variables used their analysis can be seen in Figure 1. Finally, the team excluded all referees that did not have at least 22 encounters with players. Every football game includes (at least) 22 players, 11 players per team. If a referee has less than 22 player-encounters, there are missing cases. According to team 23 this was mostly the case of referees officiating games of minor leagues. They wanted to focus on the major European league though (England, Germany, Italy, and Spain). Therefore, referees with less than 22 player-encounters were excluded. 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 shows the properties of the variables relevant to the analysis of the present project.



Team 23 Base variables properties. A) Strong skew of receiving no red to receiving a red card. B) Most players had a lower skin tone rating i.e., most players were on the whiter side. C) Not much variation in implicit racial scores. D) More variation in explicit racial bias scores. (The latter two were centred around the mean, hence, zero does not mean there is no bias.)

# Results

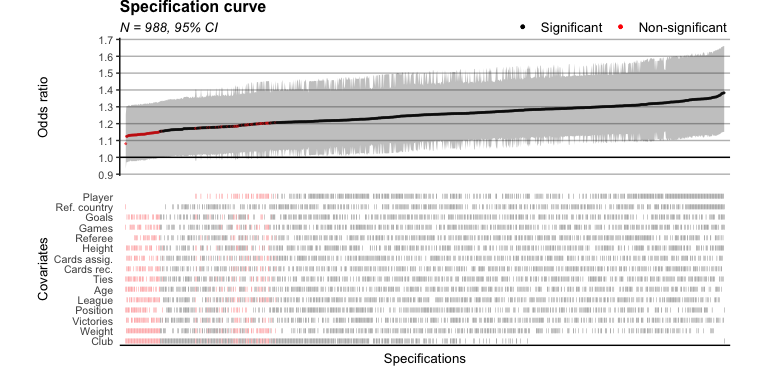
## Replicating *Team 23* analysis

    First, I replicated the *team 23* analysis using their scripts. The results were reproducible without any adjustment. Skin tone had a significant effect on the odds of being sent off the field (, 95%CI [1.099, 1.563], ). This means while keeping all other variables constant for every unit increase in skin tone rating (darker skin tone), the odds of being sent off the field increased by about 131%. The interaction terms of skin tone and implicit racial bias (, 95%CI [0.000, 23.259], ) as well as skin tone and explicit bias (, 95%CI [0.493, 6.848], ) were both non-significant. In accordance with the original analysis, there were also significant differences between leagues and positions as well as implicit racial bias scores. Explicit bias scores were on verge of being non-significant. This analysis only aimed at testing reproducibility, which proved to be true, further result details are, therefore, forfeited.

## Exploring the results space - Specification curve

     Second, I calculated specification curve to visualise the results space (Figure ) . The horizontal axes of the top and bottom chart show the specifications sorted by their OR from lowest to highest. The OR are shown on the vertical axis of the top plot. The points in the top plot (which together look like a line) each represent the outcome of one statistical model. Black and red points refer to statistically significant and non-significant outcomes, respectively. Overall, 89.7% of the outcomes were significant. Except for the lowest outcome () all points were close to their respective neighbours. This indicates that none of the specified models was responsible for a sudden increase in the outcome measure. The smooth line suggests that sampling 1,000 specifications was representative for the 33k possible models. The bottom table lists all covariates on its vertical axis, sorted from most positive to most negative impact. For instance, the covariate player had the strongest positive impact, while the covariate club had 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. Again, black and red colours indicate that the outcome measure was significant or non-significant, respectively.

     The top and bottom plots 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, the number of specifications is much lower than in the present study. With these lower numbers of specifications, it is possible to tease the different model specifications apart. However, the current project’s goal was to define the whole results space i.e., *all possible* specifications. With such a high number of specifications it is no longer possible to tease the different specifications apart. It is however interesting to look at the top and bottom rows of the table. The top row shows that the covariate player is particularly often included in models with a higher outcome. Conversely, the covariate club is particularly often included in models with a lower outcome.

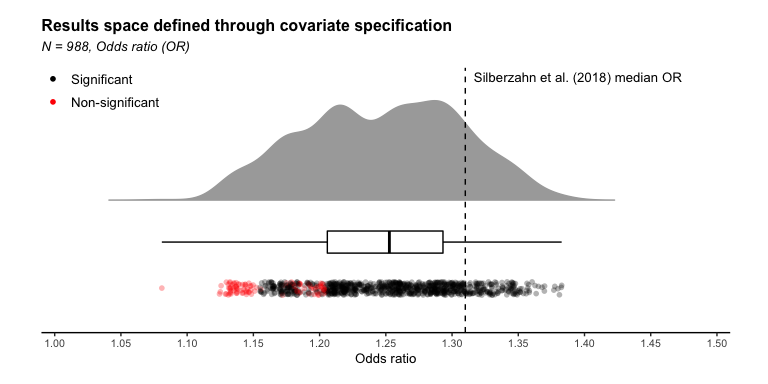


Due to the large number of specifications, the specification curve plot is not as informative as it is supposed to be. Results space ranges from roughly 1.05 to 1.4. Despite observing mostly significant results, covariates can also affect the results to be insignificant.

## Exploring the results space – Raincloud plot

     Third, I further explored the results space by calculating raincloud plots (Figure ) . The horizontal axis shows the OR. From top to bottom the three components show the probability density distribution, the box plot and the raw data. Based on the raw data points, the results space can be defined as the interval between 1.081 and 1.383. The 1st quartile was 1.206, the median 1.248 and the 3rd quartile 1.293. Thus, the interquartile range where 50% of the data are included was rather small (1.206-1.293) indicating a low statistical dispersion. The dashed line indicates the original results of *team 23*, which has been the median outcome of all analytical strategies in Silberzahn et al. (2018). The outcome of *team 23* (1.31) is included in the results space but is outside the middle 50% of the data.

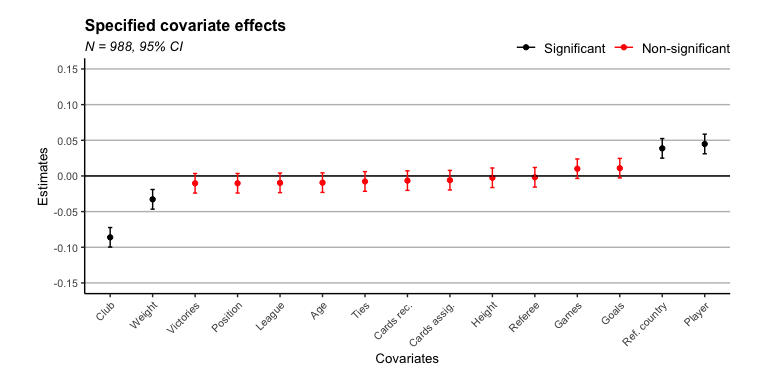
     The probability density distribution reveals another interesting feature of the data. The distribution has two peaks which suggests two latent covariate structures whose outcomes centre around two points. Those two focal points seem to be located around 1.13 and 1.19. Based on this graph alone it is not possible to decipher the latent structures i.e., determine the responsible covariates or combinations of covariates. To test the hypothesis of two latent structures it would be interesting to observe how the distribution evolves as more specifications are run. If latent structures were present, the peaks should get more pronounced. If they were none, the peaks would most likely smoothen out, approaching a normal distribution.



Results space ranging from roughly 1.05 to 1.4, possibly a higher lower bound due to a potential outlier. The distribution, showing two peaks, suggests an influential latent covariate structure. The original median effect is outside the middle 50% of the data.

## Specifying covariate effects – Scatter plot

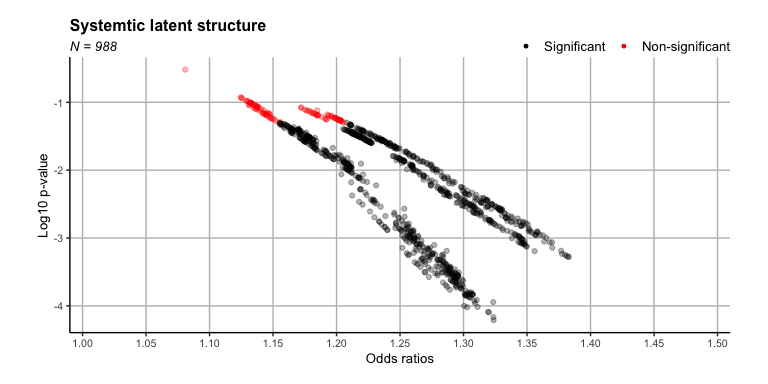
     Fourth; to better understand the relevance of the covariates I calculated another scatter plot (Figure ). This graph shows the specified covariate effects. On the horizontal axis are the covariates, on the vertical axis their estimated effects. The covariates are sorted from negative to positive impact. Black indicates a significant effect, red non-significant. The error bars represent the 95% CI. Four out of 15 covariates were significant. All effects sizes, including the significant covariates’, are relatively weak and have a narrow CI, which is likely due to large sample size. The OR of the model without any covariates (i.e., core variables only) was 1.262. Adding the sum of all positive covariate effects (games + goals + referee country + player, respectively) to the base model results in a maximum possible effect of 1.357. The maximum observed OR was 1.382 though. This difference suggests that there more impactful covariate combinations which have not yet been identified. Vice versa subtracting the sum of all negative covariate effects from the base model results in 1.070. Whereas the minimum observed effect is 1.081. This suggest that there are also some covariate combinations with a higher negative impact.



Four out of 15 covariates are significant, though they have a relatively weak effect. This suggests combinations of covariates are responsible for the higher outcome values.

## Exploring the results structure – Volcano plot

     Finally, I produced another graph which was not planned a priori, hence, it also is not included in the analysis plan. Nevertheless, it yielded interesting findings. As it was not planned, I will only use it to strengthen previous points and probe future research. The aforementioned results all suggest influential latent covariate combinations, I was very keen to explore these. I therefore went back to the Patel et al. (2015) study. Although that study had a different objective, their methodology was similar to the present project. They visualised the results as so-called volcano plot (see Figure ) which showed the relationship between the p-value and effect size. I calculated a similar graph for the present project. It indicated two distinct underlying constructs. Particularly salient is how distinct these “streams” are. Given that the specified covariate effects are relative weak, it is highly unlikely that only one covariate is responsible for this structure. Instead, it is much more likely that a combination of covariates causes this structure. Future research might specify which covariate combination underlies this structure.



Exploratory data visualisation supports the hypothesis of a systematic latent covariate structure. The strictness of the patters shown is notable.

# Discussion

     The current project defined and scrutinised the full results space of Silberzahn et al. (2018). The objective was to contribute to understanding the variability between different analytical approaches and their implications. To this end, out all analytic strategies, every possible combination was supposed to be run to define 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 replicated, yielding identical results as *team 23* did. This ensured an identical starting from all further analyses. After drawing 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 it 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 were statistically significant, though their effects were small. Given the lower and upper limits of the results space were beyond the cumulative influence of single covariates it is assumed that there are more influential latent covariate combinations. The final volcano plot supports this hypothesis by clearly outlining two (perhaps a weak third) structure. Unfortunately, identifying these underlying structures is beyond the scope of this project.

     In short, the present project found that the results space ranges from 1.08 to 1.38 with a median of 1.25. Team 23 determined the 95% CI to be ranging from 1.10 to 1.56. Despite their upper limit being higher, there is substantial overlap between the two intervals. This suggest that the original study overestimated the variability. This is highly interesting because I would have expected the full results space to be larger than the estimated CI. If the CIs were, in fact, overestimated for all analyses it would suggest that there is more agreement between the analyses than previously assumed.

     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 results space is further confirmed or corrected. Moreover, it would be highly interesting to decipher the underlying covariate structure that is responsible the observed “streams.” Moreover, it would be interesting to run all other statistical models including all covariate combinations. Based on the present findings, one might hypothesise that the full result spaces are narrower than the estimated 95% CIs. If this hypothesis were to be confirmed further research should be conduct into how multiverse analysis can account of such overestimation.

# 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>