How analysis strategy affects analysis results

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

MSc Psychological Research Methods with Data Science
Dissertation

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

Contents

Abstract

Crowdsourcing and multiverse analysis usually only assess reasonable analysis strategies. Little is known about how the results space of these *reasonable* analyses relates to the results space of all *possible* analyses. The current project, therefore, aimed at defining and understanding the results space and structure, respectively, of all possible analyses for Silberzahn et al. (2018). Due to time and resources limitations, the project focused on the analysis producing the median outcome and randomly sampled 1,000 covariate combinations out of all possible ones. The findings indicate that the full results space is narrower than the original confidence intervals suggest. If this were to be true for all analyses conducted in Silberzahn et al. (2018), the variability between the analysis strategies were likely smaller than previously though. Future studies have to confirm this hypothesis by first running the "median model" with all covariate combinations and for each statistical model.

Introduction

Theoretical framework

The Covid-19 pandemic has reaffirmed the crucial relevance of sound scientific research 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 that yields data. To draw conclusions from the data, researchers 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 ambiguity often creates uncertainty and inconsistency. Specifically, 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 odds 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 are well suited to 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 studies mentioned above. 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 specific 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, Simonsohn et al., 2020). This approach has been proposed by Botvinik-Nezer et al. (2020) and Silberzahn et al. (2018) as an alternative to crowdsourcing. In this approach, all reasonable analysis strategies are identified and performed. *Reasonable strategies* are defined as all statistically valid, non-redundant tests that are appropriate for the research question. The aggregated results of all reasonable 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.

Current project

Researchers make analytical decisions that introduce *researchers degrees of freedom*. Crowdsourcing analysis and multiverse analysis are promising tools to assess the variation of results induced by analysis strategies. However, these approaches do not assess the full range of results obtained by all possible analysis strategies but usually focus on a subset of reasonable analysis strategies. Therefore, it is unclear how the spaces of all reasonable and all possible analysis strategies relate to each other. In particular, it is unknown whether the space of all reasonable results covers the full range of possible results or whether it covers only a narrow sub-space of possible results. In this project, all reasonable and all possible analysis strategies were performed to define and understand the results space of Silberzahn et al. (2018). Furthermore, ways to visualise the space and its structure intuitively and succinctly were proposed. The project is, thus, intended to further the basic understanding of how analysis strategies influence research results and to provide tools for further studies on this topic.

Analysis

Analysis plan

The project's objective was to define the space and the structure of all possible results 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. This standardised approach to define the results space was used. The approach essentially comprises two parameters: the statistical models and the covariates (or control variables). In Silberzahn et al. (2018), the analysts used numerous 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, resulting in 2^n, i.e., $2^{15} = 32,768$ possible combinations of covariates. Adding all modelling possibilities to the equation yields a total of $29 * 2^{15} = 950,272$ possible analysis strategies. As this number of strategies exceeds the current project's resources, it was reduced to a more manageable number. Therefore, the project focused on the modelling approach producing 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 analysis approach leaves $1 * 2^{15} = 32,768$ possibilities. Due to computational limitations, a random sample of 1,000 analysis strategies were conducted.

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 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 current project was to replicate their transformation and analysis. Such replication can increase the confidence in the previous and the current approach. Moreover, it ensures that this project has the same starting point as *Team 23*. As team 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, a random sample of covariates without replacement was drawn. "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:

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

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* the 95% confidence intervals (CI) were calculated. 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.

To provide insights into the results space and its structure, the 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 plots in more detail:

(i) The specification curve is a descriptive plot that 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 data 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 of 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, the specification curve was complemented with other plots.

(ii) 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, summary 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 3). 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 essential to gain insights into how specific covariates influenced the outcomes. Therefore a third plot which complemented and extended the raincloud plot was developed.

(iii) 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, assessing all models and their effects individually did not make sense. Hence, the effects of each covariate were aggregated. To this end, each covariate was recoded into a binary factor: included in the model yes/no. These newly defined factors were then 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 was in the top right corner. Additionally, the colours of the points indicated the statistical significance of the effects.

In summary, the current 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 would result in about 1M models. Due to time and computational limitations, the project focused on the modelling approach that produced the median outcome of all analyses. Hence, a mixed-model logistic regression with 1,000 randomly sampled covariate combinations was run. For each model, relevant parameters were extracted and visualised. The specification curve is not ideal for large numbers of model specifications. It was, therefore, complemented and extended by raincloud and scatter plots, which visualise the results space and the specified covariate effects, respectively.

Code

This section provides a brief overview of the code and explains the reasoning behind it. First, the data transformation and analysis of *Team 23* (Stafford et al., 2014) were replicated. The team made their project folder publicly available. It contains three scripts relevant to the current project: data exploration, transformation, and analysis. After duplicating their project folder on the 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 current project focused on the frequentist approach.

The code was based on Haessler et al. (2020) and Patel et al. (2015). The basic structure of the code is shown in the pseudocode table (see Pseudocode 1). Here, two important analytical decisions are briefly motivated. For more details, please refer to the commented code:

(i) 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 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 measurement and is herself/himself not directly relevant to answering the research question. The function used by *Team 23* requires specifying a random effect. However, given that the current 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. Therefore, this project ensured that both functions used the same estimation method (maximum likelihood or

restricted maximum likelihood estimation) to ensure that the outcomes are comparable.

(ii) No correction for multiple comparisons. 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, α). The more statistical tests are run, the more likely it is that a result of at least one of the tests is a false positive. Hence, it is good practice to account for those "multiple comparisons." The current project, 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 no correction for multiple comparisons was made.

Pseudocode 1: Outlining high-level structure of the code. "Ran" in e.g. "FormulaRanEf" refers to "random-structure". "Ef" refers to "effect".

data ← prepared data based on team 23

$Variables$ ← Define dependent (dv), base variables ($basevar$) and covariates ($covar$)
$Specifications$ ← Use $Variables$ to create matrix containing all possible covariate combinations

$Formula$ ← Paste $Specifications$ by row and append as column to $Specifications$
$FormulaRanEf; FormulaEf$ ← Separate formulas based on including a random-structure

$SampleRanEf; SampleEf$ ← Unique, random sample from $FormulaRanEf; FormulaEf$

$ExtractorFunction$ ← Extracting model statistics and store them in object

$OutputsRanEf$ ← **lapply(data,** $ExtractorFunction$**(glmer(**dv**, x =** $basevar + SampleRanEf$**)))**
$OutputsEf$ ← **lapply(data,** $ExtractorFunction$**(glm(**dv**, x =** $basevar + SampleEf$**)))**

$ModelOutputs$ ← Merge $OutputsRanEf$ and $OutputsEf$
$ModelOutputs$ ← Compute odds ratios and confidence intervals

$JoinedData$ ← **LeftJoin(**$Specifications$**, **$ModelOutputs$**, by =** $Formula$**)**

Output: Rain cloud plot

$CovariateEffects$ ← **LinearModel(**$OddsRatios$**, x =** $Specifications$**, data =** $JoinedData$**)**

Output: Covariate effects plot
Output: SCA plot, sorted by $CovariateEffects$

Output: Volcano plot

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 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 identity. The referees' countries of origin 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 (0.8%), i.e. the data were highly skewed. The team, therefore, used a logistic regression which is a statistical modelling technique equipped to deal with skewed/binary distributions. Figure 1 shows an overview of the properties of the core variables used in their analysis.

Finally, the team excluded all referees who 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 in referees officiating games of minor leagues. As they focused on the major European leagues (England, Germany, France, and Spain), referees with less than 22 player-encounters were excluded. This step excluded about 66% of the referees but retained 97.4% of all player-referee combinations. The final dataset used for the current project contained 335,537 observations and 19 variables.

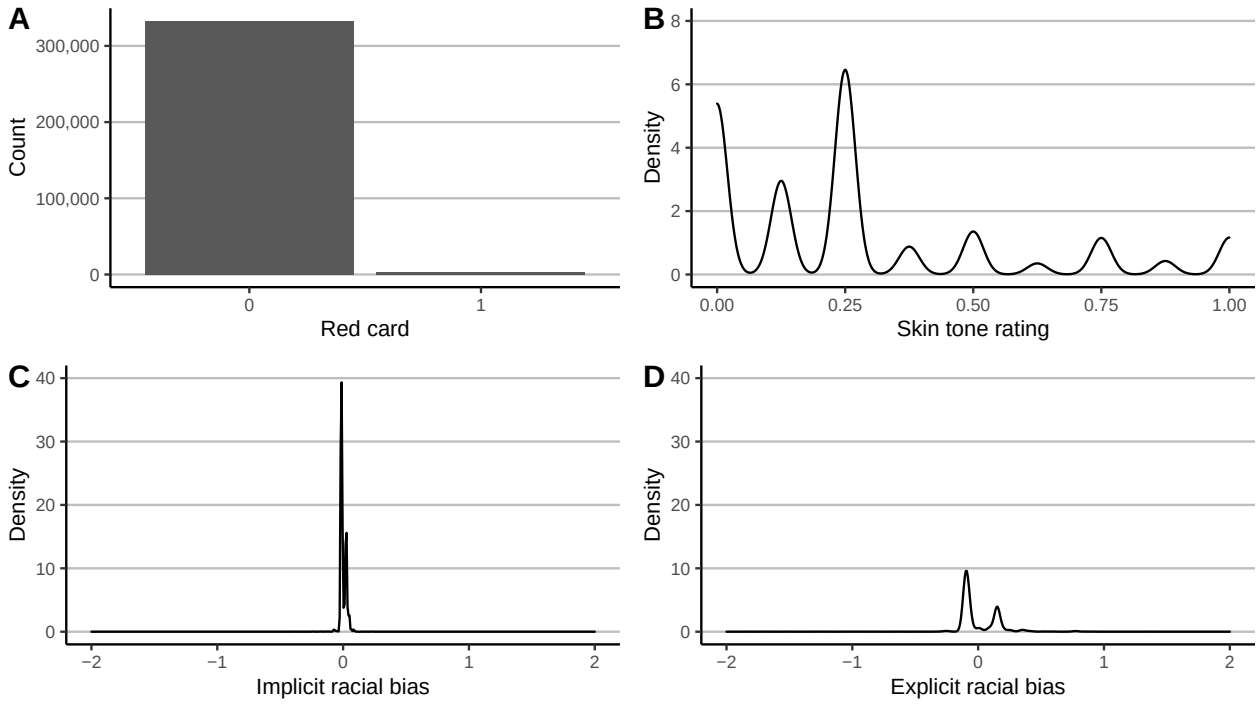

Figure 1: Team 23 base variables properties. A) There was a strong skew towards receiving no red card compared to receiving a red card. B) Most players had a lower skin tone rating, i.e., a brighter skin colour. C) The implicit racial bias scores hardly varied. D) The explicit racial bias scores varied slightly more than the implicit scores. (The latter two were centred around the mean; hence, zero does not mean there is no bias.)

Results

Replicating *Team 23* analysis

First, the analysis of *Team 23* was replicated using their scripts. Skin tone significantly affected the odds of being sent off the field ($OR = 1.311$, 95%CI [1.099, 1.563], $p = 0.003$). This means that when 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 ($OR = 0.004$, 95%CI [0.000, 23.259], $p = 0.211$) as well as skin tone and explicit racial bias ($OR = 1.837$, 95%CI [0.493, 6.848], $p = 0.365$) were both non-significant. In accordance with the original analysis concerning additional variables, there were significant differences of OR between leagues and positions as well as implicit racial bias scores. Explicit bias scores were on the verge of being non-significant. Thus, the results of *Team 23* were reproducible without any adjustment.

Exploring the results space - specification curve

Second, the specification curve was calculated to visualise the results space (Figure 2). 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 ($OR_{lowest} = 1.081$) all points were close to their neighbours. This indicates that none of the specified models was responsible for a sudden increase in the outcome measure. The bottom table lists all covariates on its vertical axis, sorted from most positive (top) to most negative impact (bottom). For instance, the covariate "player" had the most substantial positive impact, while the covariate "club" had the most substantial negative impact. Each column of the table represents one model. 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" was particularly often included in models with a higher outcome. Conversely, the covariate "club" was particularly often included in models with a lower outcome.

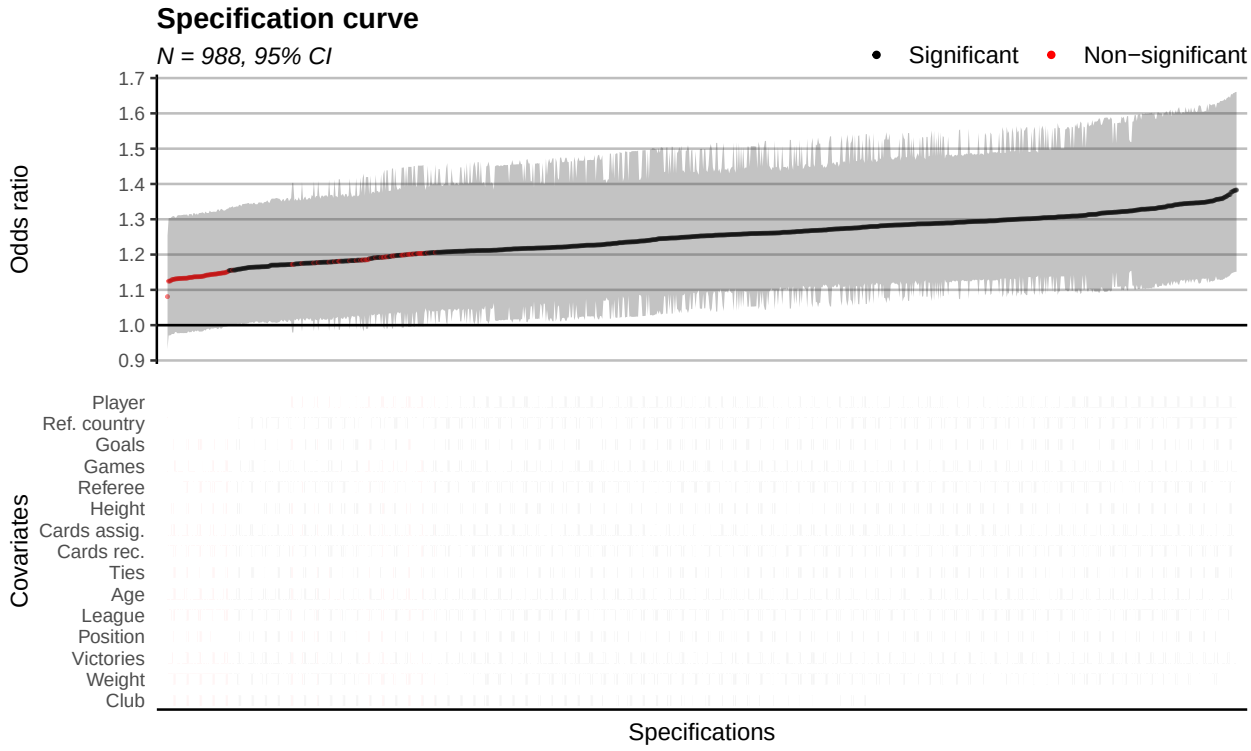

Figure 2: Specification curve. Results space ranges from about 1.1 to 1.4. However, due to the large number of specifications, the specification curve does not allow for identifying the relevance of the different covariates.

Exploring the results space – raincloud plot

Third, the results space was further explored by calculating raincloud plots (Figure 3). 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 were included was relatively narrow (1.206 to 1.293), indicating a low statistical dispersion. The dashed line indicates the original result of *Team 23*, which has been the median outcome of all analysis strategies in Silberzahn et al. (2018). The outcome of *Team 23* (1.31) was included in the results space but outside the middle 50% of the data. The CI calculated by team 23 ranged from 1.10 to 1.56. The lower bound of their CI was close to the lower margin of the results space of the present study. In contrast, the upper bound of their CI was notably higher than the present results space. This suggests a potential overestimating of the variability of the results in the original analysis.

The probability density distribution reveals another interesting feature of the data. The distribution has two peaks which suggest two latent covariate structures whose outcomes centre around two points. Those two focal points were located at about 1.13 and 1.19. Based on this graph alone, it is impossible 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 when

more specifications are run. In the presence of latent structures, the peaks should get more pronounced. In the absence of latent structures, the peaks should smoothen out, and the distribution should approach a normal distribution.

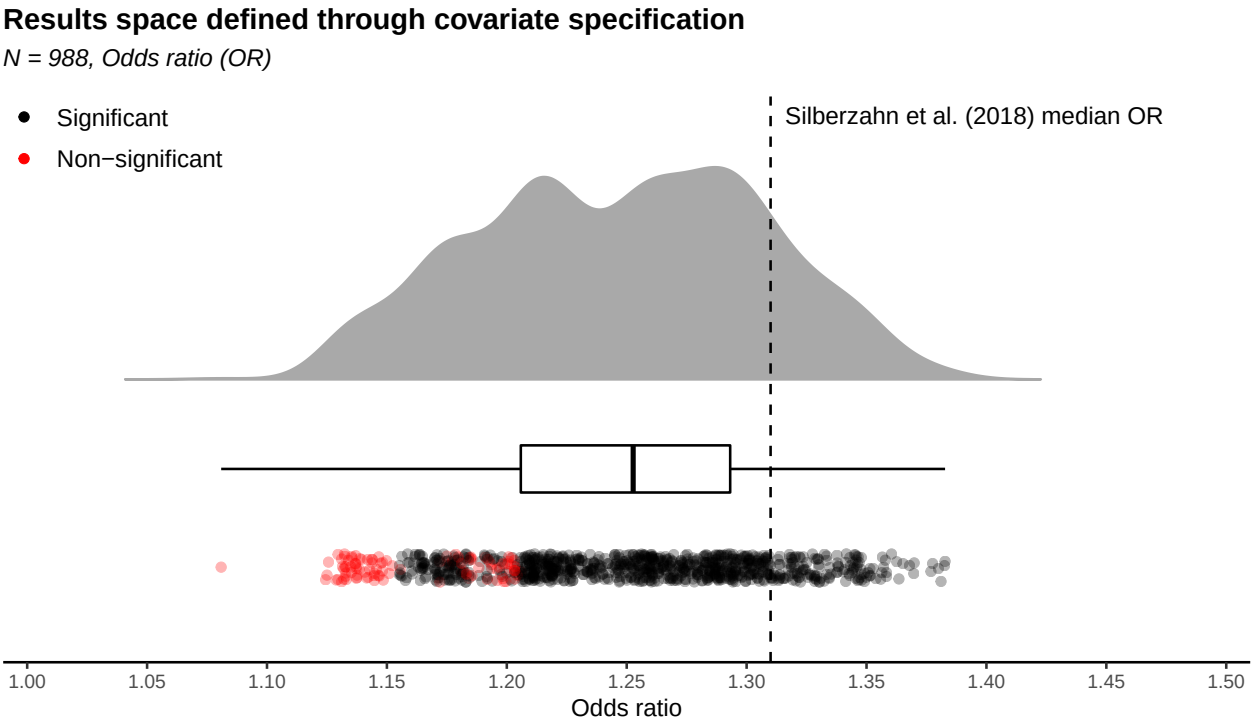

Figure 3: Raincloud plot. The results space ranges from about 1.08 to 1.38. The distribution showed two peaks which suggests an influential latent covariate structure. The original median effect was within the results space but outside the middle 50 percent of the data.

Specifying covariate effects – scatter plot

Fourth, to better understand the relevance of the covariates, another scatter plot was calculated (Figure 4). This graph shows the specified covariate effects. On the horizontal axis are the covariates; on the vertical axis are 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 effect sizes, including the significant covariates', were relatively weak. The CIs were relatively small, which was likely due to the 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 resulted in a maximum possible effect of 1.357. The maximum observed OR was 1.382, however. This difference suggests that there were more impactful covariate combinations or interactions that have not yet been identified. Vice versa, subtracting the sum of all negative covariate effects from the base model resulted in an OR of 1.070, whereas the minimum observed effect was an OR of 1.081. This suggests that there were also covariate combinations with a higher negative impact.

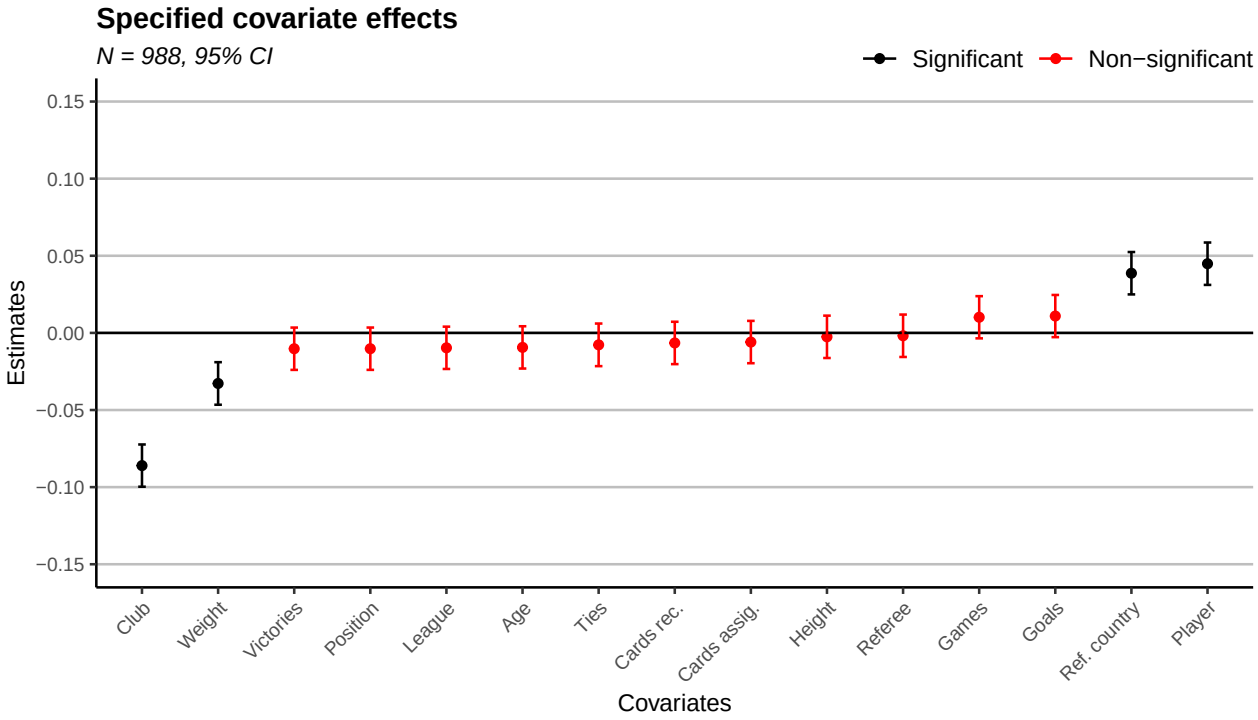

Figure 4: Scatter plot. Four out of 15 covariates were significant, although their effects were relatively weak. This suggests that covariates interactions were responsible for the higher outcome values.

Exploring the results structure – volcano plot

Lastly, another graph was produced despite not being planned. Therefore, it is only used to strengthen previous points and probe future research. The results above suggest one or more meaningful latent covariate combinations; these were further explored. Although Patel et al. (2015) had a different objective, their methodology was similar to the current project. They visualised the outcomes as a so-called volcano plot (see Figure 3), showing the relationship between the p-values and effect sizes. A similar graph was calculated for the current project. The two (potentially three) distinct "streams" provided further support for two (or three) distinct underlying constructs. Future research might specify which covariate combinations underly these structures.

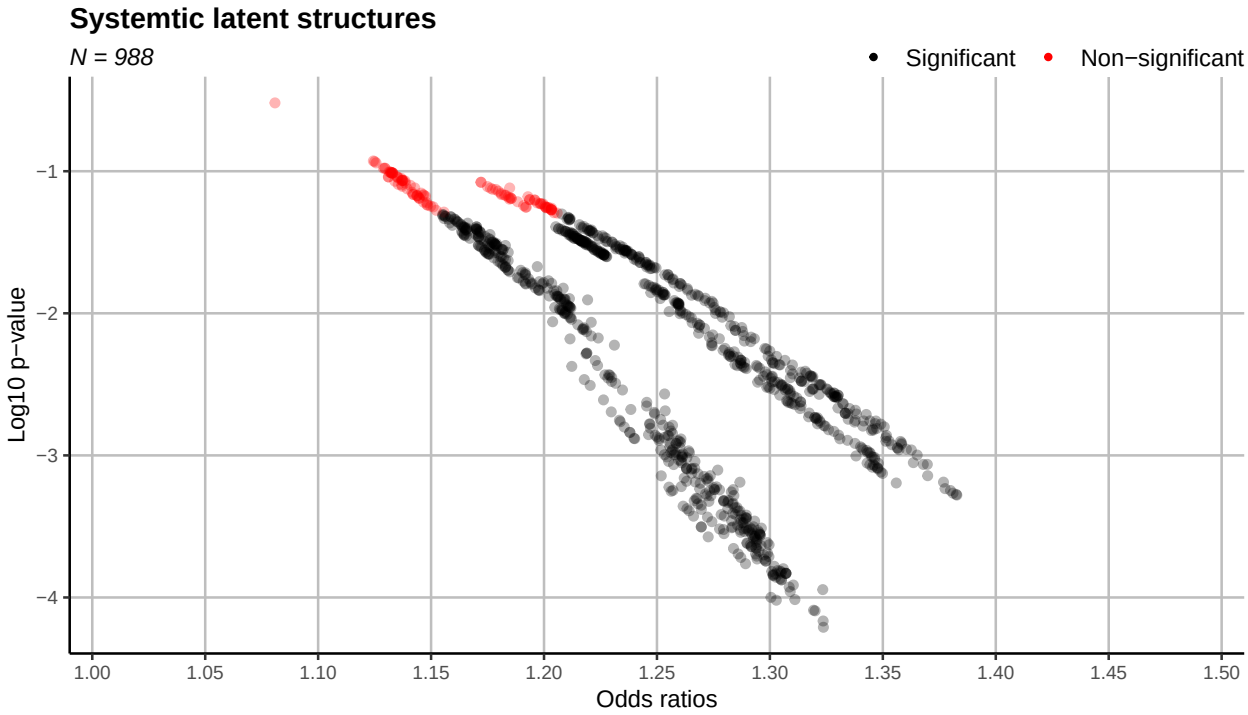

Figure 5: Volcano plot. The exploratory data visualisation supports the hypothesis of a systematic latent covariate structure with at least two underlying latent constructs.

Discussion

In the present multiverse project, not only all *reasonable* but for the first time all *possible* analysis strategies were performed to define and understand the results space of Silberzahn et al. (2018). The objective was to investigate whether the space of all *reasonable* results covers the full range of *possible* results or whether it covers only a narrow sub-space of *possible* results. Due to time and computational limitations, the current project focused on one analysis strategy and a random sample of covariate combinations. The chosen analysis strategy (i.e., the statistical model) was the one producing the median outcome in the original study.

First, the median analysis was successfully replicated, which ensured an identical starting point for all further analyses. Next, 1,000 unique covariate combinations were sampled, respective models were run, and results were visualised. After removing outliers, the data were distributed in a relatively narrow results space ranging from about 1.1 to 1.4. The median outcome of Silberzahn et al. (2018) was not included in the middle 50% of the data, though it was part of the 95% confidence interval. Therefore, it is concluded that the covariates of that model had an above-average effect, i.e. this covariate combination was more influential than a random combination of covariates.

Moreover, the distribution of the results space showed two peaks. To understand the distribution's underlying structure, the covariate effects were analysed. The results showed that 4 out of 15 covariates

were statistically significant, though their effects were small. Moreover, when adding the individual covariate effects to the outcome of the base model, i.e. the model without any covariates, the observed outcome values are always well within the results space. Thus, interactions of covariates likely contribute to the results space. The final volcano plot supports this hypothesis by showing at least two streams indicating at least two underlying constructs. Future research might investigate these latent structures. In short, the current project found that the full results space is narrower than the CIs of the original "median" analysis of Silberzahn et al. (2018) indicated. This suggests that the original study overestimated the variability. This was surprising as the full results space would have been expected to be larger than the estimated CIs. If the variability were overestimated for all analyses, this would suggest that there is more agreement between the analyses than previously assumed.

A common problem in research is p-hacking, i.e. tweaking the analysis to find significant effects, which are considered more "publishable." Recent preventive efforts include pre-registering the theoretical framework and analysis plan. The idea is to be transparent about the analysis strategy and holding oneself accountable to the pre-defined standards. Additionally, the data and analysis code are made publicly available. These effects also enabled the current project. Silberzahn et al. (2018) found that 2/3 of all outcomes were statistically significant. This poses the question of whether researchers fished for significant results or whether these outcomes reflect the true relationship. To scrutinise this question, Patel et al. (2015) suggested assessing the robustness of an effect, i.e. its consistency in different conditions. The current project built on their methodology and, therefore, also assessed the robustness of an effect. Considering that about 90% of all observed outcomes were significant, the effect of skin tone is considered to be robust. Hence, researchers have likely detected a true effect.

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 compromise between rigour and feasibility. The constraints and the findings of the current project yield great potential for future research. A first step could be to run not only a random sample but all possible covariate combinations. It would be interesting to know whether this procedure would confirm or correct the results space observed in the present study. Moreover, it would be highly interesting to decipher the underlying covariate structure responsible for 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 would be confirmed, further research might investigate how multiverse analysis can account for 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