


# Has the Credibility of the Social Sciences Been Credibly Destroyed? Reanalyzing the “Many Analysts, One Data Set” Project

Katrin Auspurg<sup>1</sup>  and Josef Brüderl<sup>1</sup>

Socius: Sociological Research for a Dynamic World  
 Volume 7: 1–14  
 © The Author(s) 2021  
 Article reuse guidelines:  
[sagepub.com/journals-permissions](https://sagepub.com/journals-permissions)  
 DOI: 10.1177/23780231211024421  
[srd.sagepub.com](https://srd.sagepub.com)  


## Abstract

In 2018, Silberzahn, Uhlmann, Nosek, and colleagues published an article in which 29 teams analyzed the same research question with the same data: Are soccer referees more likely to give red cards to players with dark skin tone than light skin tone? The results obtained by the teams differed extensively. Many concluded from this widely noted exercise that the social sciences are not rigorous enough to provide definitive answers. In this article, we investigate why results diverged so much. We argue that the main reason was an unclear research question: Teams differed in their interpretation of the research question and therefore used diverse research designs and model specifications. We show by reanalyzing the data that with a clear research question, a precise definition of the parameter of interest, and theory-guided causal reasoning, results vary only within a narrow range. The broad conclusion of our reanalysis is that social science research needs to be more precise in its “estimands” to become credible.

## Keywords

credibility crisis, crowdsourcing, reproduction, replication, causal reasoning, robustness analysis, multiverse analysis, sensitivity analysis, estimands

## Introduction

Are scientific findings credible? Basically, there are two approaches for answering this question. First, one might try to replicate findings. If findings can be successfully replicated, they are credible (valid). This approach has been implemented in various disciplines over the last decade with several large-scale replication audits. Results were quite disconcerting in that the audits reported an unexpectedly low replication rate (e.g., see Begley and Ellis 2012; Camerer et al. 2018; Christensen and Miguel 2018; Open Science Collaboration 2015). Across scientific disciplines, this has led to what is known as the “credibility crisis in science.”

A second approach, which has been used less often, is the crowdsourcing approach: Several researchers analyze the same research question with the same data (for a review, see Uhlmann et al. 2019). Science is credible if different researchers come up with a similar answer (this idea can already be found in Merton 1973; Popper 1959). A recent crowdsourcing study was the “many analysts, one data set”

project of Silberzahn and colleagues (2018a; hereafter, *crowdsourcing initiative* [CSI]). In this project, 29 teams analyzed the same research question on racial bias in soccer with the same data. The result was that the answers to the research question given by the 29 teams differed extensively. The authors concluded: “Any single team’s results are strongly influenced by subjective choices during the analysis phase. . . . Taking any single analysis seriously could be a mistake” (Silberzahn and Uhlmann 2015:191). They warned the public and politicians against trusting the results of single social science studies.

The CSI has had a huge impact. It seems to be a standard reference when discussing the (low) credibility of social

<sup>1</sup>Department of Sociology, LMU Munich, Munich, Germany

### Corresponding Author:

Katrin Auspurg, Department of Sociology, LMU Munich, Konradstr. 6, Munich, DE-80801, Germany.  
 Email: [Katrin.Auspurg@lmu.de](mailto:Katrin.Auspurg@lmu.de)



science research with observational data (see e.g., Damian, Meuleman, and van Oorschot 2019; Reed 2019; Young 2018). Many, especially in media articles and Internet discussions, even concluded that the social sciences may not be rigorous enough: Findings depend to a large extent on who did the research.

Sociology has been quite absent from discussions about a credibility crisis, which is probably mostly due to a lack of evidence on reproducibility (Freese and Peterson 2017; Auspurg and Brüderl Forthcoming). So far, large-scale replication audits have focused on experimental studies published in other disciplines, such as medicine, psychology, and economics. This literature says little about the situation in sociology, in which most studies are based on observational data instead of experimental data. And here comes in the CSI project: It is based on observational data and uses a typical sociological research question on racial bias. Thus, it was the first large-scale study that directly speaks toward the credibility of standard sociological research. And as already mentioned, its results cast serious doubts on the credibility of social research with observational data.

Therefore, it is important to explore the sources for the huge variation in results found in the CSI project. In this article, we start such an investigation. Our main argument will be that the CSI used an unclear research question for the crowd. The teams had to make up their minds on how to interpret the research question. We identify (at least) four different interpretations made by the teams. Consequently, the teams implemented (at least) four different research designs. This produced much variation in the results: It is no wonder that results differ if researchers investigate different research questions.

To demonstrate this, we reanalyze the CSI data with one clear research question. We apply theory-guided causal reasoning to define the parameter of interest precisely. Nevertheless, we all know that social research with observational data is like a “garden of forking paths” (Gelman and Loken 2014). The way from the data to the results is long and needs many decisions. Even with a clear research question and theory-guided causal reasoning, several decisions may not be obvious. Thus, there is model uncertainty. To simulate this, we allow for a huge and systematic amount of model uncertainty that is implemented by means of a multiverse analysis (Steen et al. 2016; also called *multimodel analysis*; Young and Holsteen 2017). We show that when estimating hundreds of different model specifications, there is variation in results, but within a much smaller range than found in the CSI.

Our main conclusion will be that the CSI did not “destroy” the credibility of sociological research in general: It only showed that *nonrigorous* social research may produce inconsistent results. We demonstrate that *rigorous* social research (based on a clear research question, a precise definition of the parameter of interest, and theory-guided causal reasoning) can be more consistent. Thus, to increase its credibility,

sociological research must become more precise in its “estimates” (Lundberg, Johnson, and Stewart 2021).

The article proceeds as follows: In the following sections, we summarize the CSI, discuss its design, present our reanalysis (first our analytic design, then data and results), continue our reanalysis by applying sensitivity analysis, give a summary of our results, and provide a discussion of what we can learn from the CSI and our reanalysis regarding the credibility of sociology. We also offer some suggestions for improving the practice of social science research in general.

## Many Hands Make Tight Work? Review of the CSI

In this section, we will shortly summarize the procedures and findings of the CSI (for a detailed description, see Silberzahn et al. 2018a). All researchers were asked to answer the same research question with the same data: Are soccer players with dark skin tone more likely to receive red cards from referees than players with light skin tone? The data were compiled by a sports-statistic company, with the sample being all soccer players in the 2012–2013 season in the first male leagues of four European countries ( $N = 2,034$  players). Information about the interaction of those players with all referees ( $N = 3,147$ ) whom they encountered across their professional career was obtained (until the time of data collection in 2014). The data were provided in an aggregated form with the single cases being dyads of the players and the referees ( $N = 146,028$ ). Variables were the number of red and yellow cards the players received in these dyads and some characteristics of the players at the time of the data collection (e.g., their body weight, age, the club they were playing for). Information on the player’s skin tone was provided in the form of two independent ratings based on visual inspections of photographs of the players.

The 61 researchers (in 29 teams) that participated until the end of the CSI were recruited by an open call. There were not any restrictions regarding their qualification. Status groups ranged from bachelor’s students to full professors. Substantive backgrounds included sociology, psychology, and economics. The teams worked independently from each other, but there were also some crowd discussions on statistical models, and all teams received feedback from other teams, which was thought to mirror the standard peer-review process for journal publications.

The treatment variable was a 5-point scale ranging from 0 (very light skin) to 1 (very dark skin). The outcome was the likelihood of receiving a red card. Results were reported in odds ratios (ORs). The median OR was 1.31, meaning that players with very dark skin color, compared to players with very light skin color, had 31 percent higher odds of receiving a red card. This is a moderate association of skin color with red cards. Most teams reported results close to the median. But there were also teams that found no association at all or teams that found much stronger effect sizes, reaching up to

nearly 200 percent higher odds reported by two “outlier” teams. The range of reported ORs was .9 to 2.9 (a kernel density plot of the estimates, excluding the outliers, is provided in Figure 2b). In addition, there was a large variance in the estimated standard errors, leading partly to different conclusions on statistical significance even when teams found similar effect sizes.

Thus, the main finding was a huge variation in the results produced by the different teams. Accordingly, the leading authors warned against relying on the analyses of one team only (as is standard in published research): The conclusions might depend to a large extent on who did the research.

### Why Was the Variation of Results Obtained in the CSI so Large?

Silberzahn and colleagues (2018a) tried to explain the large variation in results by analyzing the impact of different classes of statistical models (e.g., ordinary least squares [OLS]/logit/Poisson regressions; Bayesian methods yes/no) and numbers of covariates that were used by the different teams. However, the attempts to explain the variance by these modeling choices were not very successful (see Silberzahn et al. 2018a, in particular Table 4 and Figure 2).

So what was the main driver of the inconsistency in results? Our conjecture is that it was the unclear research question given to the teams. In rigorous research, one would first give a precise statement of the research question and define the “parameter of interest” (the “theoretical and empirical estimands” as it is called by Lundberg et al. 2021). After that, one would apply relevant theories and causal reasoning to arrive at an identifying research design. With observational data, this would result in the specification of a statistical model including relevant control variables (confounders and mediators; see Elwert and Winship 2014; Kohler, Sweet, and Class 2018). We show in the following that there was no precise research question defined in the CSI. Consequently, each team had to come up with its own interpretation of the research task.

Recall the research task that was set in the CSI: to find out whether players with dark skin tone are more likely to receive red cards than players with light skin tone. How would you interpret this research task? In our opinion, this verbal statement of the research question is quite diffuse. After analyzing the reports submitted by the different teams in the CSI (Silberzahn et al. 2018b; available at <https://osf.io/gvm2z>), we conclude that in fact, there were at least four different interpretations.

1. The literal interpretation of the verbal statement is that the parameter of interest is the mean difference in the risk of red cards between dark- and light-skinned players. This is the *bivariate association* of skin tone and red cards. A simple bivariate analysis answers this descriptive research question, and one should not include any controls.

2. One might interpret the verbal statement also as a question about discrimination or racial bias. Then the parameter of interest is the *direct causal effect* of skin tone that remains after netting out confounders and “productivity-relevant” mediators. Discrimination in our context means that players are treated differently *solely* because of their race that is signaled by their skin tone (for a general review, see Pager and Shepherd 2008). Dark-skin players are more often punished with a red card than light-skin players, *all else equal*. To correctly estimate this direct skin tone effect, one has to partial out all alternative mechanisms that would be in line with equal treatment, such as player’s position, minutes played, or different baseline rates at which individual players commit fouls (for similar arguments for basketball research, see Price and Wolfers 2010).<sup>1</sup>

Equations 1 and 2 formalize the research designs that correspond to the first two research questions (cf. Young and Holsteen 2017:19). The “treatment” variable (skin tone) is denoted with  $X$ , and  $Y$  is the dependent variable (red cards);  $i$  is an indicator for different games, and  $\varepsilon$  is an error term. The parameters of main interest are  $\beta$  and  $\beta^*$ .  $\beta$  measures the bivariate association;  $\beta^*$  measures the direct effect of  $X$  on  $Y$ .

$$Y_i = \alpha + \beta X_i + \varepsilon_i \quad (1)$$

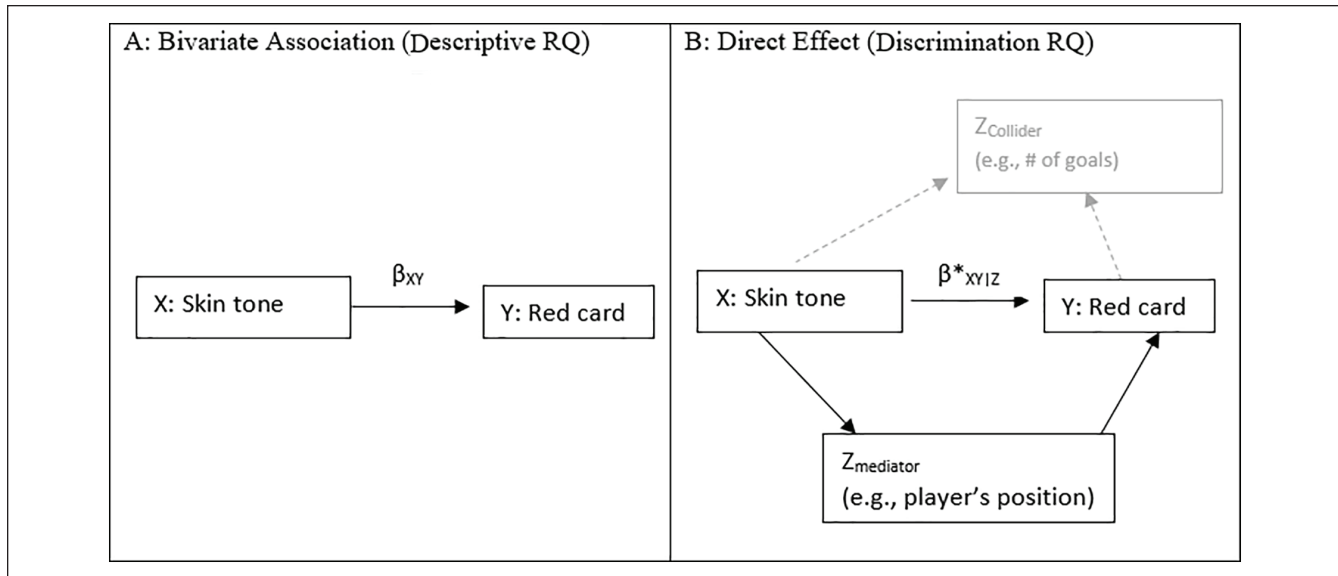
$$Y_i = \alpha^* + \beta^* X_i + \gamma W_i + \delta Z_i + \varepsilon_i^* \quad (2)$$

The difference lies in the inclusion of controls  $\mathbf{W}$  (confounders) and  $\mathbf{Z}$  (mediators) in the second equation (with  $\gamma$  and  $\delta$  denoting their regression coefficients). With their inclusion, the core parameter changes ( $\Delta = \beta - \beta^* \neq 0$ ). This is true as long as the controls are correlated with  $X$  and at the same time influence  $Y$ . Thus, it is no wonder that teams that tried to answer Research Question 1 arrived at different conclusions than those that answered Question 2.

The unbiased estimation of the “discrimination effect”  $\beta^*$  requires a careful selection of controls  $\mathbf{W}$  and  $\mathbf{Z}$ , based on the methodology of causal inference and substantive literature on discrimination and soccer research (as with any “ $X$ -centered” research strategy; Ganghof 2005). When estimating direct effects, the general rule is to control for both confounders and mediators but to not control for colliders (Elwert and Winship 2014; Young and Holsteen 2017:11).

Examples for mediators have been given before. In the case at hand, it is more difficult to imagine any confounders (i.e., variables that precede the skin tone and at the same time

<sup>1</sup>Mediators that represent discrimination mechanisms should not be controlled for (because this would induce an overcontrol bias). In the example at hand, it is, however, difficult to imagine any such mediator.



**Figure 1.** Causal diagrams for two different research questions (RQs).

Note: In Panel b, only the black variables should be included in the model specification. To obtain unbiased estimates of the direct effect, one has, in general, to also include confounders (in this case, composition effects).

impact the likelihood of red cards). But there might be composition effects (i.e., variables that correlate with skin tone and affect the likelihood of red cards). For instance, age is certainly not an effect of skin tone (i.e., it is not a mediator), but it is correlated with skin tone due to the fact that dark-skin-toned players entered European soccer leagues increasingly in recent years. Composition variables such as age should also be controlled for.

One must not, however, include controls that lead to a backward causal link with the outcome variable because these controls induce a collider or endogenous selection bias (Elwert and Winship 2014; Young and Holsteen 2017:11). In the example at hand, such an endogenous variable could be the number of goals scored by the player. This is because soccer players who receive a red card are sent off for the rest of the game and are also suspended for the next game(s), shortening the possible time in which players can score goals.

The causal reasoning behind the two research questions is best represented graphically (Figure 1). Whereas the bivariate association is estimated without any controls (Figure 1a), the model specification for estimating the direct effect should include mediators (in black, solid arrows) as controls but not any colliders (in gray, dashed arrows; Figure 1b). The use of such causal diagrams is very helpful to specify and communicate the causal assumptions underlying the identification of the causal effect of interest (for details, see Elwert 2013).

From the reports submitted by the different teams in the CSI (see the supplementary material in Silberzahn et al. 2018b), one can infer that some teams indeed tried to answer Question 1, whereas the majority of teams focused on Question 2. (We provide exemplary quotes on the teams' research questions in Table A1 in the Supplemental Material.)

In addition, there were two further interpretations of the research question:

3. Some teams aimed at “Y-centered” research (Ganghof 2005), meaning they did not try at all to partial out the effect of skin tone on red cards. Instead, they focused on maximizing explained variance in *Y* (likelihood of red cards). Their motivation was to see whether skin tone is among the explaining factors. For this kind of research, covariate selection is based only on statistics of model fit (such as  $R^2$  values in regressions). Model fit is often increased in particular by the inclusion of endogenous covariates (colliders) that are not allowed in the *X*-centered causal research to answer Question 2 (Elwert and Winship 2014; Young and Holsteen 2017).
4. Finally, still other researchers chose an “exploratory” research strategy (for exemplary quotas, see Table A1 in our Supplemental Material). By this strategy, they wanted to get at novel, extreme, or unexpected findings. In the CSI, the researchers focusing on this research strategy seemed to be mostly motivated by the fact of being part of a crowd: Given this setting, they wanted to enlarge the pool of findings that were gathered by the crowd. Thus, they often explicitly used very unorthodox statistical techniques. Their motto was “anything goes.”

Thus, we argue that the results obtained in the CSI showed such a large variation because the 29 teams pursued (at least) four different research questions and therefore used different research designs. Designs according to Questions 1 to 3 led



to models that greatly differed in the adjustment set, and accordingly, results differed. Teams that pursued Question 4 used “weird” models and thereby added some outlier results to the CSI.

## Analytic Strategy of Our Reanalysis of the CSI Data

Our basic hypothesis is that the CSI produced a large variation in results because the teams did not pursue one research question, but four. The most straightforward strategy to demonstrate this would be to group the 29 research teams according to their research question and to compute the variation within and between the groups. According to our hypothesis, the share of between-group variation should be relatively large. However, we will not follow this research strategy because in many cases, it is quite difficult to classify the teams in a definitive way from the team reports. Most teams did not specify any explicit research question, as would be common in standard research articles, but instead reported only the statistical approach (e.g., kind of regression model) and results. Therefore, the results obtained from this research strategy (grouping teams into four different research questions) would be very uncertain.

Instead, we pursue a different strategy. We reframe our basic hypothesis: The CSI would have produced much less variation in results if the teams would have investigated only one precisely defined research question. Therefore, we chose one of the four research questions and investigated the range of results one could possibly obtain. Instead of starting a new crowdsourcing exercise, we “simulate” crowdsourcing by allowing for (a reasonable amount of) model uncertainty. If the range of results obtained by this simulation is rather low, we would conclude that rigorous social science research is able to provide a consistent answer for the chosen research question.

For this strategy, we selected the arguably sociologically most interesting research question: Research Question 2 on racial bias. We suppose that Silberzahn and colleagues had this research question in mind. Moreover, this is an *X*-centered research question, the type most often examined in quantitative sociological studies (for numbers on the *European Sociological Review*, see Kohler et al. 2018). Finally, discrimination research is certainly central to sociology.

In an ideal world, given a precise definition of the parameter of interest, theory would suggest one optimal model specification, and all researchers would obtain the same result. However, sociological theories, including theories on discrimination, usually provide only a vague idea on the causal association between variables, and there may also be uncertainty about the correct operationalization of concepts or the functional form of regression models. Thus, in the “real” world of social science research, there will be model uncertainty, and researchers will arrive—depending on their specification decisions in the garden of forking paths—at different results. Identifying the effect of this uncertainty on results was exactly the goal of the CSI. Instead of relying on

the “manual” work done by crowdsourcing teams, we simulate the effect of model uncertainty drawing on computer algorithms that allow one to estimate a huge range of possible model specifications (multiverse analysis). This allows for an even larger range of specifications than manual crowdsourcing. Thus, this is an even more conservative test of our hypothesis than a crowdsourcing exercise would provide.

We do not allow for maximum model uncertainty; specifically, we do not allow for anything goes—Research Question 4—but only for reasonable model uncertainty. We argue that discrimination theories and causal methodology at least give some guidance on how to specify a model for investigating Research Question 2. This restricts the model space to a reasonable subspace, which, however, is larger than one because some uncertainty is left.

So, in the following, we basically (1) start from a precisely defined research question informed by discrimination theories and causal reasoning and delineate the space of reasonable model specifications. Then, we (2) apply multiverse analysis to the CSI soccer data to investigate how large the range of results is.

## Data and Results

The data recorded player-referee dyads, linking players with the different referees they encountered during their careers ( $N = 146,028$  dyads). The variables informed on the number ( $n_j$ ) of games in the dyad (with  $j$  indexing the different dyads), on the number ( $q_j$ ) of red cards the player received from the referee, and on some characteristics of the players and referees at the time of the data collection in 2014.<sup>2</sup>

One should condition analyses on the time spent on the playing field: Only active players are at risk of receiving a red card. The data at hand include only a proxy for time spent on the playing field: the number of games played. To adjust our analyses for the numbers of games, we expanded the player-referee data to player-referee-game data.<sup>3</sup> For 483 players, there was no information on the skin tone (or body weight or height). We excluded these players from the

<sup>2</sup>Variables that change across careers (e.g., club and league, player’s position) were measured at the time of data collection and not at the specific times the red cards were received. This strongly limited the analytic potential of the data, which was also criticized by many teams in the CSI.

<sup>3</sup>Alternatively, we could have run regressions on the dyad level and control for the number of games. However, controlling for the number of games might introduce a collider bias (players with a red card are suspended also for the following game, which makes the number of games endogenous to the number of red cards in the dyad). Therefore, we used this alternative strategy only for robustness checks (available on request). This did not change any of our conclusions. Some teams in the CSI used the proportion of red cards per game as outcome. Note, however, that this is an inferior control for the number of games because this value is zero for all dyads with no red cards that might, however, consist of different numbers of games.

following analyses. So, the analyses are based on information from 1,551 players in 371,813 player-referee-game combinations.

The outcome variable for these expanded data is a 0/1 indicator of whether the player received a red card in a respective game or not. (Soccer players can receive at maximum one red card per game.) Because there was only the aggregate information on the overall number ( $q_j$ ) of red cards on the dyad level, we randomly assigned the red cards on the game level within the dyads. Because there was not any information that was measured on the level of games, the random assignment of red cards was innocuous. The overall probability of a player receiving a red card in a game is .43 percent.

For the treatment variable, skin tone, we used the mean value of the two 5-point ratings that were provided in the data. These ratings ranged between 0, very light skin, and 1, very dark skin (with interrater reliability  $r = .92$ ).

Now, to get an upper bound for model uncertainty, we had to decide on all model ingredients that seem reasonable for the research question at hand. Model ingredients are the set of potential control variables (adjustment set), the set of different functional forms, and the set of different operationalizations. The combination of all model ingredients defines the model space.

First, we decided on the variables in the adjustment set (an overview and more detailed arguments on the variables in the adjustment set can be found in Table A2 in the Supplemental Material). Remember the general rule: Include all productivity-relevant mediators and composition variables; do not include colliders. Thus, we included the composition variable *age* in the adjustment set for reasons specified already before. We also included four mediating variables: player's *height* and *weight* (players that are larger/heavier built might more likely receive a red card), player's *position* (e.g., red cards are more likely received by defenders), and the proportion of *victories per game* (indicator for the level of frustration, which may have increased the risk of red cards due to more aggressive playing).

In addition, to provide a conservative test of our hypothesis, we also included two variables in the adjustment set for which it is unclear whether they are mediators or colliders: *club* and *country*. Club may be a mediator because clubs might teach different playing styles (e.g., aggressiveness). However, there are also arguments that speak for club as being a collider: The admission to top clubs may depend on both discrimination and the player's ability to avoid red cards (being issued a red card puts the player's team in disadvantage). Similar arguments apply for country (see Table A2 in the Supplemental Material). Note that club and country are collinear, and therefore, only one of the two variables can be included in a particular model.

Furthermore, we did not include number of goals per game in the adjustment set because we were very confident

that it would induce a backward causal link. Players who receive a red card are sent off for the rest of that game, shortening the possible time in which players can score goals.

Finally, for various reasons, we excluded two variables that potentially could be used in the models. We excluded referee's IDs because there is no reason to assume that players with different skin tones could self-select into games monitored by specific referees.<sup>4</sup> We also did not use a variable informing on the referee's country of origin. There might be a country-specific level of prejudice against dark-skin persons, and referees might have internalized these prejudices during their socialization.<sup>5</sup> However, from the perspective of causal models, it only makes sense to include this variable in form of an interaction term with skin tone to explore effect heterogeneity (moderator effects), which is out of the scope of our research goal.

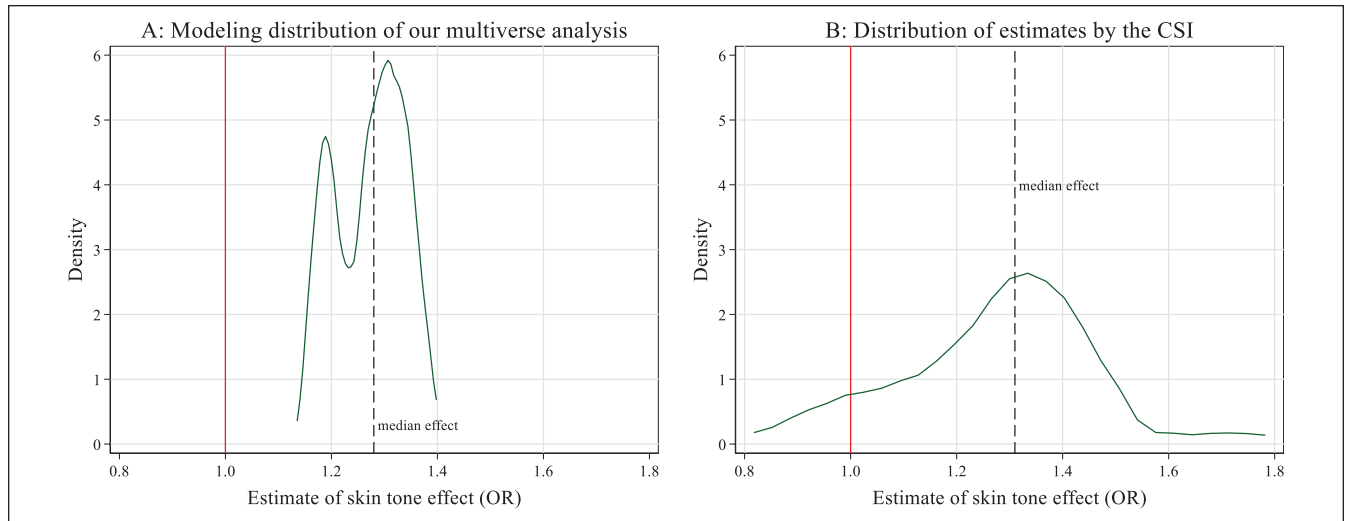
We ran logistic regressions to estimate ORs as they were reported in the CSI.<sup>6</sup> Regressions were run with cluster-robust standard errors to account for the nested data structure with several games being observed for the same player (Rogers 1993).

Because there is uncertainty concerning the functional form and operationalization of some controls, we increased the model space by allowing for variation in this respect. For metric covariates (age, height, weight), we allowed for linear or quadratic specifications ( $Z$  or  $Z + Z^2$ ). We also allowed for

<sup>4</sup>In professional leagues, it is always a board that selects the referees, and we see no reason why these decisions should be associated with players' skin tone. If this assumption holds, including referee fixed effects (as some teams in the CSI did) would only help to reduce some random noise caused by variation in referee's decision behavior (some referees might in general tend to award a red card more frequently than other referees). However, the drawback is that 429 cases of referees that took responsibility for only one game would be dropped. For these reasons, we decided to run only some robustness checks with referee fixed effects, which did not change any major conclusions (available on request).

<sup>5</sup>There was also one variable in the data set to measure this level of prejudice with Implicit Association Test scores. However, there was much discussion in the CSI that this variable was not a good proxy and therefore better not used. Some inconsistent results in the CSI may have been caused by the fact that some teams nevertheless stuck to this variable, which may have led to extreme effect heterogeneity, especially in the exploratory research.

<sup>6</sup>When using logistic or probit regressions, the estimated parameters may also change due to the inclusion of controls that are not correlated with  $X$  (Mood 2010). Therefore, these models are fundamentally problematic for causal analyses. Nevertheless, we used logistic regression to compare our results with the CSI results. As a robustness check, we also performed multiverse analysis with linear probability models (available on request). The general pattern of results did not change.



**Figure 2.** Distribution of skin-tone effects estimated by multiverse analysis and the crowdsourcing initiative (CSI).

Note: Density graphs of  $N = 486$  estimates (Panel a) and  $N = 27$  estimates (Panel b). The red line marks the odds ratio (OR) of 1 (zero effect). The kernel density in Panel a is based on multiverse analysis with the Stata ado *mrobust*, using the soccer data provided in the CSI ( $N = 371,813$  games). The kernel density in Panel b was produced with the effect sizes reported in Silberzahn et al. (2018a), omitting the two most extreme outliers (which were ORs of 2.88 and 2.93).

two alternative operationalizations of the player's position (using the original categorical variable with 12 levels or, alternatively, only 5 levels).

The combination of all these model ingredients generates a model space of 486 different regression models. The most parsimonious models include no controls<sup>7</sup>—thereby in fact answering Research Question 1; others include the complete adjustment set, and most models are in between. Multiverse analysis runs the whole set of regressions in the model space and thus provides 486 different estimates of the skin-tone effect.<sup>8</sup> We used the Stata ado *mrobust* provided by Cristobal Young and Katherine Holsteen (2017; see also Muñoz and Young 2018; Young 2018).

First, in Figure 2b, we show a graphical representation of the CSI findings: Median effect size was 1.31, but the variation was quite high, as can be seen by the wide range of the distribution (even though we dropped the two most extreme ORs of 2.9 from the graph). The standard deviation (including the outliers) was .45.

Figure 2a shows the modeling distribution of our reanalysis: The median effect size was 1.28, which is very close to the CSI median. However, despite allowing a very large model uncertainty, the estimates for the skin-tone effect

were in a much narrower range than the ones reported in the CSI ( $SD = .06$ ).<sup>9</sup>

Our multiverse analysis shows that all (reasonable) model specifications provide substantively very similar answers: All effect sizes were in a narrow range; throughout, all estimates pointed in the same direction ( $OR > 1$ ); and two thirds (68%) of the estimates were statistically significant at the 5 percent level (for these and further model statistics provided by the *mrobust* algorithm, see Table A3 in the Supplemental Material).

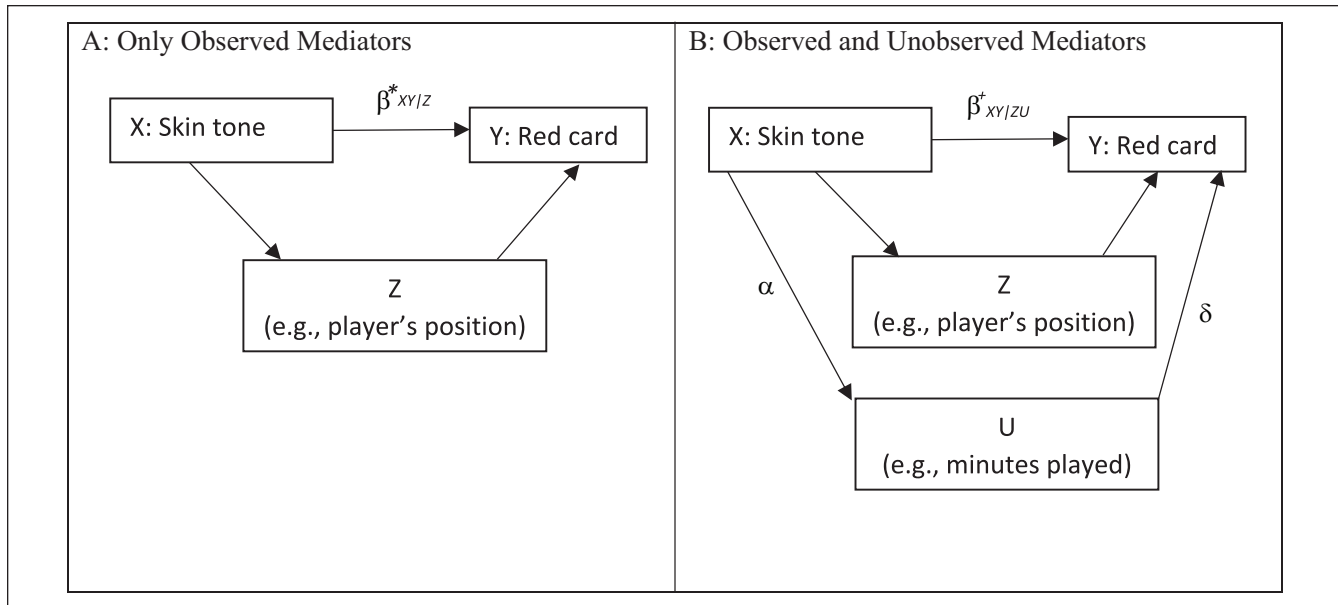
Also noteworthy is that the distribution of estimates from the multiverse analysis was bimodal, meaning that there were two “clusters” in the model space. These two clusters were close to each other ( $OR = 1.2$  and  $OR = 1.3$ ). In other applications, the differences in estimates may be more substantial, and in these cases in particular, one may want to learn which covariates are causing the variation in results. The *mrobust* algorithm also allows one to identify the model ingredients of most influence. In our case, it was the variable *club*. The mean effect size was  $-.12$  smaller when the models included the club variable. As mentioned previously, one might argue that the club variable is a collider and therefore should not be included in the adjustment set. This would narrow the range of the results even further. However, to provide a conservative test of our hypothesis, we did not exclude this variable from the adjustment set.

To sum up so far, the results of our multiverse analysis support our basic hypothesis. By rigorous social research

<sup>7</sup>Strictly speaking, number of games is always controlled by design.

<sup>8</sup>Age, height, and weight all come in three variants (not in the model, linear, or linear and quadratic). Position comes also in three variants (not, 5-category, and 12-category). Club/country comes in three variants (not, club, and country). Victories per game comes in two variants. This gives a model space with  $3 \times 3 \times 3 \times 3 \times 3 \times 2 = 486$  elements.

<sup>9</sup>A very similar modeling distribution emerges when we restrict the CSI results to the 10 teams where we are certain that they followed Research Question 2: Median effect size was 1.33, and the standard deviation was .10.



**Figure 3.** Causal diagram: direct effect estimated with observed and unobserved mediators.

Note: The parameter  $\alpha$  denotes the coefficient of a bivariate regression of  $U$  on  $X$ , and  $\delta$  is the coefficient from a multivariable regression of  $Y$  on  $U$  and all other covariates in the model. The product of these two parameters quantifies the size of the “omitted variable bias.”

(using a precisely defined research question and applying theoretically informed causal reasoning), it is possible to arrive at a consistent answer: The soccer data at hand support the conclusion that there is a (moderate) racial bias in the likelihood of receiving a red card.<sup>10</sup>

### Consistent Results Still Can Be Wrong: Sensitivity to Omitted Variables

Unfortunately, this is not the end of the story. Consistency of results is not the only criterion for credible results. Consistent results obtained with observational data easily can be wrong if there are unobservables. With observational data, there always is the potential of bias due to omitted variables. Thus, we admit that our reanalysis of the CSI data only may have shown that social research can produce consistent results but not necessarily valid results. We are aware of the fact that our result of a (moderate) racial bias may be invalid because important productivity-relevant mediators are missing in the CSI data. However, and this is the good lesson of this section, there are tools that at least allow one to estimate the sensitivity of results to bias caused by omitted variables.

<sup>10</sup>One reviewer argued “that there is inherent researcher variability that goes along with any given researcher” and therefore our results as reported in Figure 2a may be very subjective. Fortunately, there is a second, completely independent reanalysis of the CSI data that corroborates our main finding. Young and Stewart (2021) also reanalyzed the CSI data by multiverse analysis. They allowed for even more modeling uncertainty by including many more functional forms than we did. Nevertheless, their modeling distribution was very similar to our modeling distribution that we report in Figure 2a.

In the following, we want to make scholars aware of the additional insights that can be gained by such sensitivity analyses. Multiverse analysis is helpful in testing the robustness of results against combinations of *observed* variables (which helps, for instance, to see whether results hinge on “knife edge” model specifications; Muñoz and Young 2018). But multiverse analysis does not help to test the sensitivity in regard to *unobserved* variables. For this, one has to use different algorithms that give indication toward the sensitivity of results against unobservables.

According to Research Question 2, we are interested in the direct causal effect of skin tone net of productivity relevant mediators. We estimate the effect  $\beta^*$  shown in Equation 2: the skin-tone effect conditional on all mediators that are available in the CSI data (Figure 3a). Given the theoretical structure from Figure 3a, causal inference is threatened by two main problems with unobservables. We have to assume (1) that there is no unobserved mediator-outcome confounding<sup>11</sup> and (2) that there are no unobserved productivity-relevant mediators. Methods for checking the sensitivity of results concerning the first assumption are too complex to discuss here (Lundberg et al. 2021 provide an introductory discussion to the subtleties of mediation analysis). Therefore, we focus only on the second assumption in the following.

A violation of our second assumption is illustrated in Figure 3b. There is an unobserved productivity-relevant

<sup>11</sup>An example might be a player’s unobserved competitiveness, which might affect his position and also affect the probability of receiving a red card. Then player’s position is a collider, and controlling for it will bias the estimate of the direct effect. We are grateful to an anonymous reviewer who brought this to our attention.



mediator denoted  $U$ .  $\alpha$  denotes the effect of  $X$  on  $U$ , measured by a bivariate regression, and  $\delta$  is the effect of  $U$  on  $Y$ , measured by a multivariable regression.  $U$  could be, for instance, unobserved differences in the exact time “at risk” on the playing field (caused by varying numbers of minutes played per game or by varying numbers of overtimes played). In case meaningful unobserved mediators exist ( $\alpha \neq 0$  and  $\delta \neq 0$ ), our estimate  $\beta^*$  would be biased ( $\Delta = \beta^*_{XY|Z} - \beta^+_{XY|ZU} \neq 0$ ).

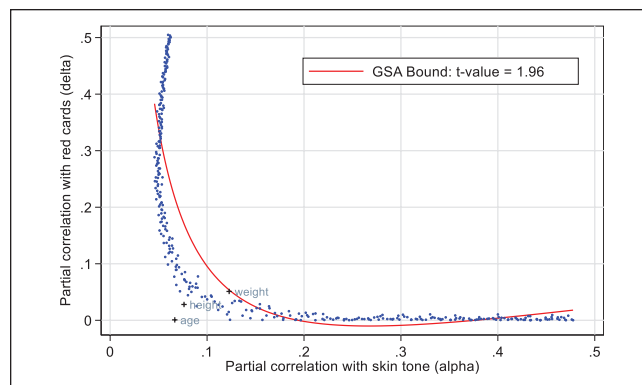
In the example at hand, the question of whether we can give the skin-tone effect  $\beta^*$  a causal interpretation boils down to the question of whether there could be sizable bias due to the omission of meaningful mediators. Technically, the bias that is caused by the omission of a mediator is an “omitted variable bias” (OVB). The size of this OVB is the product of two terms, often called *sensitivity parameters*: (1) the effect  $\alpha$  of the treatment  $X$  on the mediator  $U$  and (2) the partial effect  $\delta$  of the mediator  $U$  on the outcome  $Y$ .<sup>12</sup>

We do not know these two parameters. But we can do a kind of thought experiment: How large would these parameters have to be so that the parameter of interest is brought down to a negligible level? To answer this question, one can use general sensitivity analysis (GSA) provided by Harada (2013), which builds on the approach developed by Imbens (2003). The GSA algorithm (provided with the Stata ado *gsa*) uses the unexplained variance (residuals) to estimate combinations of the two sensitivity parameters ( $\alpha$  and  $\delta$ ) that would change the treatment effect (or its test statistic) to a target criterion defined by the user. The size of these estimated sensitivity parameters allows one, then, to judge the likelihood that there might exist an omitted mediator that could “explain away” the treatment effect (here, the skin-tone effect).

We illustrate this with the CSI data. We base the sensitivity test on a logistic regression model that includes the covariates age, weight, height, player’s position (five categories), proportion victories, and country. Using this regression, the estimate of  $\beta^*_{XY|Z}$  was 1.29 (OR)—which matches the median estimate found in CSI and also the multiverse analysis very well (see “Data and Results” section). The effect was statistically significant ( $p = .019$ ). As target criterion for the GSA, we chose a  $t$  value of 1.96, which corresponds to a marginally insignificant treatment effect (using the common 5 percent level for statistical significance).

The results of this sensitivity analysis are summarized in the “contour curve” in Figure 4. To ease interpretation, the sensitivity parameters  $\alpha$  and  $\delta$  are converted to partial correlations. The correlation with skin tone ( $\alpha$ ) is on the  $x$ -axis, and the partial correlation with red cards ( $\delta$ ) is on the  $y$ -axis. The dots show combinations of these two parameters (simulated unobservables) that would change the skin-tone effect to an insignificant effect ( $t = 1.96$ ). The curve termed *GSA bound* is the fitted curve to these dots. For instance,

<sup>12</sup>The same is true for an omitted variable bias caused by a missing confounder. In regression analyses, the effects of both types of variables, mediators and confounders, are conceptually the same (see Wooldridge 2013).



**Figure 4.** Contour plot of general sensitivity analysis for the skin-tone effect.

Note: This figure was produced with the Stata ado *gsa*. The target criterion for this sensitivity analyses is  $t = 1.96$ . The dots show combinations of the two partial correlations where the test statistic of the skin-tone effect would change to  $t = 1.96$ , meaning that the skin-tone effect would become insignificant. GSA bound is a curve fitted to the dots (fractional polynomial). For three observed variables, the combination of the partial correlations is plotted.  $N = 371,813$  games.

one can see that in case an unobserved mediator would be modestly correlated with skin tone ( $\alpha = .1$ ), already a small partial correlation of this mediator with the outcome (red cards) of around  $\delta = .05$  would be sufficient to turn the skin-tone effect insignificant. In general, one can learn from these analyses that small to modest associations of unobserved mediators with the treatment and outcome variable would explain the (direct) skin-tone effect away.

Is it likely that such unobserved mediators exist? We already mentioned the exact minutes played. Another candidate might be season effects (the likelihood of calling fouls might have changed over time). Finally, there might be a baseline rate at which individual players commit fouls (there was already evidence that such variables can significantly vary across athletes with different skin tone; see Price and Wolfers 2010). Given this and also the fact that observed variables, such as player’s weight, showed partial correlations that were close to the GSA bound (see Figure 4), we think that it is fairly likely that the direct skin-tone effect found in the CSI data is not causal but would vanish if these mediators would be controlled.<sup>13</sup>

<sup>13</sup>This conclusion that there is likely no discrimination is additionally corroborated by the fact that the findings are not robust to using an alternative discrimination outcome, which is the likelihood of receiving a yellow card. This likelihood is significantly *negatively* correlated with skin tone, meaning that the darker players’ skin tone, the *less* likely they received yellow cards (for another reanalysis focusing on this effect, see Berrar, Lopes, and Dubitzky 2017). Again, one would need more sophisticated analyses to see whether this effect is causal. Repeating analyses as the ones shown in this article, the yellow card effect seemed to be of similar robustness as the red card effect: It was in a narrow range when estimated in a multiverse analysis, but there likely exist unobserved mediators that explain the effect away (results on request).

In sum, the result obtained with the CSI data that there is a modest racial bias in receiving red cards is plausibly due to the fact that the data are very limited in measured productivity-relevant mediators. It is very likely that more informative soccer data would show no racial bias in receiving red cards. The more general message here is that poor data and relying only on observed variables may produce very consistent results that are nevertheless wrong. Sensitivity analysis provides a very useful tool to evaluate this type of uncertainty in social research with observational data.

## Summary

The CSI organized by Silberzahn et al (2018a) seemed to have demonstrated that social science is not able to provide consistent answers. Their main result was that the answer for a typical sociological research task largely depends on substantive analytical choices made by single researchers; therefore, the social sciences seem to have a credibility problem.

We argued that the CSI underestimated the credibility of social science findings. This was mainly due to the fact that the CSI did not start with a clear research question (and no precisely defined parameter of interest). A high variation in estimates would be problematic only as long as the estimates relate to the same parameter of interest. Teams in the CSI focused, however, on four different research tasks, reaching from descriptive to *X*-centered causal to *Y*-centered and to explorative research. Each of these research questions defines a different parameter of interest and requires a different research design to identify this parameter.

The two main findings of our empirical analyses are as follows. First, we reanalyzed the CSI data to demonstrate that one can indeed achieve much more consistent results when one specifies only *one* concrete research question and uses theory-guided causal reasoning to derive reasonable model specifications. We focused only on *X*-centered, causal discrimination research that is common in sociological research and used multiverse analysis to identify the full range of model uncertainty in social research with observational data. Remarkably, our reanalysis very well replicated the median estimate in the CSI, which was a moderate skin-tone effect. However, although we tested a much higher number of alternative model specifications by systematically varying plausible model ingredients in the multiverse analysis (that included hundreds of different model specifications), our results were within a much narrower range (i.e., much more consistent). Our conclusion is therefore that the CSI showed that nonrigorous social research that does not start with a clear research question provides divergent results. However, rigorous social science research is able to provide a more consistent answer.

Second, we argued and demonstrated that consistent results might still be biased. Even when results are very robust to numerous (manually) chosen model specifications, they might not catch the “true” causal effect as long as there is omitted variable bias. This is because results obtained with

observational data are always contingent on the information content of the data. If the data do not contain information on important controls, results may be consistent but wrong. Therefore, to be credible, social science research also should be transparent concerning the sensitivity of the results against unobservables. Applying sensitivity analyses to the result obtained with the CSI data, we demonstrated that the result of a modest racial bias in the likelihood of receiving a red card is quite sensitive to unobserved mediators. This indicates that this estimate, although consistent across many different model specifications, is probably not a true causal effect in itself.

## Discussion

All in all, we paint a relatively optimistic picture of social research: Only “bad” social research that is not pursuing a clear research task has a strong credibility problem; “good” social research can provide more definitive answers. Nevertheless, one might argue that the CSI mirrored standard flaws in social science research settings and thus depicts a “realistic” picture of social research. We discuss the argument in the following section. Some suggest that crowdsourcing could be a way to enhance the credibility of social research, which we also discuss in the following. Finally, we present our own suggestions for improving social research derived from our reanalysis of the CSI.

### *Did the CSI Provide a Realistic Picture of Social Science Research?*

Skeptics might argue that our optimistic picture is an ideal and not from this world. The CSI used real researchers and thereby showed the current reality of social research. And the current reality is nonrigorous social research of low credibility. We (partly) agree. In fact, in many published articles, there is often only a vague specification of the research question. The parameter of interest is not precisely defined and can be inferred only implicitly by the reader (see Lundberg et al. 2021). There is no clear causal reasoning to justify model specification (Kohler et al. 2018). Furthermore, it is common practice to not only interpret the estimate for the parameter of interest but also to give the effects of the control variables a causal interpretation (Keele, Stevenson, and Elwert 2020).

But on the other side, the CSI overstressed this problem because real social science research would at least specify the broad type of the research question (e.g., descriptive, causal, or exploratory). In addition, it is standard that articles include a theory section that at least implicitly gives a specification of the research question. Thereby, the largest source of variability in the CSI likely did not mirror real research practice.

Furthermore, the CSI did not implement standards of quality control as is typical in the real world of social science research: There was no strict peer review. The CSI implemented only a loose review among the participants, but

teams were at their discretion to follow the suggestions of other teams or not. Given that many teams in the CSI were very inexperienced (e.g., consisted of bachelor's students), this might have added additional variance to the results. Also, because of the exotic methods used by some teams and the many indications of misspecified models (e.g., strongly inflated standard errors), we suspect that most of this research would not have stood serious peer review.<sup>14</sup> Thus, many weird results entered the CSI end report that very likely would have been filtered out by a strict peer-review process. Consequently, in the real world, the variability of results would have been lower.

Finally, crowdsourcing initiatives (and metaresearch more general) might have a systematic bias toward showing that research is *not* credible.<sup>15</sup> Some teams might be motivated to stand out from the crowd: Particularly creative scientists might not follow the crowd and estimate boring standard regressions but might instead be motivated by the crowd-source setting to use weird methods (i.e., to follow Research Question 4). Another motivation for doing so might be to increase the body of findings and thereby promote evolutionary scientific progress. We argue that in real social research, such motivations exist to a much lesser extent and that weird articles are often screened out by peer review.

Our design, however, also comes with limitations. The most serious limitation is probably that our multiverse approach focused only on model uncertainty, including different categorizations of key variables, but not on uncertainty that can be caused by coding errors or flaws in data preparation. A new crowdsourcing exercise (Breznau et al. 2021) argues that such hidden sources of variation are very common. Thus, our approach might underestimate the variability of real research. Therefore, further developments of automatic robustness analysis that also uncover such hidden sources of model uncertainty would be very helpful.

One might conclude that in the CSI, an overly vague research task and some flawed research that was not filtered out by peer review may have produced an overly pessimistic portrayal of the credibility of real social research. Nevertheless,

we recognize that real social research also does not come close to the optimistic picture we have painted. Most likely, current social science research practice lies somewhere in between. More metaresearch on standard research practices is needed to come to more firm conclusions.<sup>16</sup>

### *Crowdsourcing as the Future Mode of Social Research?*

How could we improve the credibility of social research? So far, crowdsourcing exercises have not been very successful in finding the reasons for the high variability of results. Therefore, some argue that variability is unavoidable: "Discrepant results and variability in research findings . . . are perhaps unavoidable, and might best be embraced as a normal aspect of the scientific process" (Landy et al. 2020:469). The recent study by Breznau et al. (2021) found "a vast universe of research design variability normally hidden from view in the presentation, consumption, and perhaps even creation of scientific results." This finding is even more pessimistic: not only that results vary but also that, even more, this variation is for unknown reasons ("a hidden universe of analytical flexibility"; similar results and arguments can be found in Huntington-Klein et al. 2021).

Therefore, some crowdsourcing enthusiasts conclude that "taking any single analysis seriously could be a mistake" (Silberzahn and Uhlmann 2015:191). Consequently, Landy et al. (2020) argue that we should change the mode of the scientific enterprise: The prevailing mode should no longer be that single teams investigate a research question, but rather, a crowd of teams should investigate a research question, and the unavoidably diverging results should then be averaged somehow (i.e., through some sort of meta-analysis; see Landy et al. 2020).

Although this method of "crowdsourcing hypothesis tests" might be helpful with experimental research (the context of the Landy et al. 2020 study), we do not think that crowdsourcing currently should become the standard for observational data analysis. The majority of observational studies may not be very rigorous. Averaging over these could be counterproductive. For instance, if a field is full of misspecified models, these will dominate the result (as is well known in meta-analysis: "garbage in, garbage out"). Instead, we would suggest

<sup>14</sup>Many teams stated themselves that they would never had submitted their findings to a journal. Many teams had not any experience in multilevel analyses and/or discrimination research, and some teams also decided for pragmatic reasons to apply methods that they themselves considered flawed (e.g., because they lacked time or because they struggled with technical issues such as "overheated computers").

<sup>15</sup>There is evidence that replication audits tend to draw an overly pessimistic picture of replicability (e.g., by "null-hacking" or ignoring boundary conditions for effects that were specified in the original research; see Bryan, Yeager, and O'Brien 2019). There seems to be a publication bias in the opposite direction than in original research: Findings of spectacular low reproducibility are more easily published. To advance science, we certainly need a more balanced picture.

<sup>16</sup>There are more crowdsourcing studies around (for an overview, see Uhlmann et al. 2019). Two studies (Breznau et al. 2021; Huntington-Klein et al. 2021) are directly relevant to our discussion because they crowdsourced sociological research questions with observational data. Both studies report much variation in the results obtained by different teams and argue that this is due to hidden decisions by the researchers. Although we concur with the general finding of these studies, we suspect that they also tend to overstate the uncertainty of social science research due to the mechanisms that we discussed in this section. Therefore, our discussion might be helpful for making future crowdsourcing exercises more "realistic."



that to increase the credibility of social research, it is more helpful to first increase the quality of each single study (for suggestions on this, see the following section). Admittedly, also with rigorous social research, some uncertainty will remain (as shown in the “Data and Results” section). Then multiverse analyses or crowdsourcing might be helpful to uncover the remaining amount of uncertainty.

Finally, we offer a few notes on promising avenues for future crowdsourcing exercises. (1) The model space in crowdsourcing initiatives is naturally limited by the number of teams. Thus, a promising avenue for future replication initiatives could be an innovative combination of crowdsourcing and multiverse analysis. (2) Instead of crowdsourcing in the form of competing teams, the forces of all participating researchers could be joined to deliberate on the “optimal” analysis. This could also reduce the personal harm caused by conflicting results between teams. (3) In addition, a promising avenue for future meta-research could be crowdsourcing initiatives that explicitly incorporate elements of causal reasoning into their design. For example, one could test different approaches in split-half samples, such as teams working independently or collaboratively on the best causal modeling.

### ***An Alternative First Step: Improving Social Research Practice***

In a nutshell, we have argued that it is more productive to increase the quality of any single study rather than crowdsourcing (and then simply averaging) many studies of lower quality. Some practical conclusions can be drawn from our reanalysis that could form a kind of blueprint for better social research. These recommendations are detailed in the following (a very similar list of recommendations was proposed for psychology by Grosz, Rohrer, and Thoemmes 2020; see also Box A1 in the Supplemental Material).

First, good social research should always start from a clearly defined research question and give a precise definition of the parameter of interest. Ideally, this parameter would be derived from a formal, theoretical model. At least, researchers should clearly specify which parameter in their statistical model provides information for their research question. In this regard, research designs should always be optimized for only one parameter of interest (Keele et al. 2020; Kohler et al. 2018). It is generally not possible to answer many different questions with one design. Recently, similar points were made much more forcefully by Lundberg et al. (2021). These authors argue that productive social research must start with precise definitions of the theoretical (research question) and empirical (parameter of interest) estimands. We refer the reader to this article for a much more nuanced treatise of these issues.

Second, good social research should use theory-guided causal reasoning to justify an appropriate model specification for identifying the parameter of interest. It is insufficient to simply throw the “usual suspects” as controls in a regression model. Unfortunately, this is common practice, as

shown by Kohler et al. (2018). Instead, the identifying assumptions should be made explicit, preferably through visualization in the form of graphical causal models, such as directed acyclic graphs (see e.g., Elwert and Winship 2014).

Third, robustness analysis of the results should become common practice. Many articles report robustness checks. However, there is evidence that robustness analyses are selectively reported that support the main results at 100 percent (Young and Holsteen 2017). Therefore, we need more serious robustness checks. In this vein, it should become standard to present the full distribution of estimates that can be obtained based on all reasonable specifications. Multiverse analysis, as used in this article, seems particularly promising in this regard. In addition, one can complement summary tables (e.g., descriptive statistics and regression tables) with visualizations that disclose the full variance in raw data and results (Cumming 2014; Healy and Moody 2014). If the robustness analysis shows that results vary widely over the model space, then researchers need to make explicit why they chose their specific model specification. This will increase transparency and credibility of social research.

Fourth, sensitivity analysis should also become standard practice. As we have argued, very consistent answers obtained with observational data may well be wrong. The estimates could be sensitive to unobserved variables. This is particularly relevant when the data contain only a few variables (as is the case with the CSI data). Then the results will be very consistent due to the limited number of controls available, but they are likely to be wrong because important controls are unobserved. Therefore, to achieve full credibility of social science results, one must demonstrate that they are not sensitive to OVB. For this, one could use tools such as those we have used in this article (for more tools for robustness and sensitivity analyses, see Christensen, Freese, and Miguel 2019; Ding and VanderWeele 2016).

Altogether, precise specification of the parameter of interest, transparent reasoning about the assumptions necessary for its identification, and transparency about the robustness and sensitivity of the results to other reasonable model choices are probably the most effective measures to increase credibility in the social sciences.

### **Authors' Note**

We used data and materials from the project page of the CSI (Silberzahn et al. 2018b, <https://osf.io/gvm2z/>). Our replication files can be found on the following OSF-project page (Auspurg and Brüderl 2021, <https://osf.io/h57tj/>).

### **Acknowledgement**

We thank the CSI for making their data and materials openly available (Silberzahn et al. 2018b). For preparing the data, we used code provided there by “Crowdstorming Team 01” (Bryson R. Pope and Nolan G. Pope). We appreciate the thoroughness and transparency of data preparation provided by this team. We are grateful for the



comments received by two anonymous reviewers. For helpful suggestions we thank participants at the lab meeting of the Meta-Research Innovation Center at Stanford (METRICS) on March, 6, 2020, and the online workshop on Analytical Sociology at the Venice International University in November 2020 and the Nuffield College Sociology Seminar (held online) at the University of Oxford in December 2020. This paper was written while Katrin Auspurg was a visiting scholar at Stanford University.

## ORCID iD

Katrin Auspurg  <https://orcid.org/0000-0003-4504-0391>

## Supplemental Material

Supplemental material for this article is available online.

## References

- Auspurg, Katrin, and Josef Brüderl. Forthcoming. "How to Increase Reproducibility and Credibility of Sociological Research." In *Handbook of Sociological Science: Contributions to Rigorous Sociology*, edited by Klarita Gërxhani, Nan Dirk de Graaf, and Werner Raub. Cheltenham, UK: Edward Elgar. Available at OSF. June 04, 2021. <https://osf.io/tavc5/>
- Auspurg, Katrin, and Josef Brüderl. 2021. "Replication Files for 'Reanalyzing the "Many Analysts, One Data Set" Project.'" OSF. May 28. [osf.io/h57tj](https://osf.io/h57tj).
- Begley, C. Glenn, and Lee M. Ellis. 2012. "Raise Standards for Preclinical Cancer Research". *Nature* 483(7391):531–33. doi:10.1038/483531a.
- Berrar, Daniel, Philippe Lopes, and Werner Dubitzky. 2017. "Caveats and Pitfalls in Crowdsourcing Research: The Case of Soccer Referee Bias." *International Journal of Data Science and Analytics* 4(2):143–51. doi:10.1007/s41060-017-0057-y.
- Breznau, Nate, Eike Mark Rinke, Alexander Wuttke, Muna Adem, Jule Adriaans, Amalia Alvarez-Benjumea, and Henrik K. Andersen, et al. 2021. "Observing Many Researchers Using the Same Data and Hypothesis Reveals a Hidden Universe of Data Analysis." MetaArXiv. <https://osf.io/preprints/metaarxiv/cd5j9/>.
- Bryan, Christopher J., David S. Yeager, and Joseph M. O'Brien. 2019. "Replicator Degrees of Freedom Allow Publication of Misleading Failures to Replicate." *Proceedings of the National Academy of Sciences* 116(51):25535–45. doi:10.1073/pnas.1910951116.
- Camerer, Colin F., Anna Dreber, Felix Holzmeister, Teck-Hua Ho, Jürgen Huber, Magnus Johannesson, and Michael Kirchler, et al. 2018. "Evaluating the Replicability of Social Science Experiments in Nature and Science between 2010 and 2015." *Nature Human Behaviour* 2(9):637–44. doi:10.1038/s41562-018-0399-z.
- Christensen, Garret, Jeremy Freese, and Edward Miguel. 2019. *Transparent and Reproducible Social Science Research: How to Do Open Science*. Oakland: California University Press.
- Christensen, Garret, and Edward Miguel. 2018. "Transparency, Reproducibility, and the Credibility of Economics Research." *Journal of Economic Literature* 56(3):920–80.
- Cumming, Geoff. 2014. "The New Statistics: Why and How." *Psychological Science* 25(1):7–29. doi:10.1177/0956797613504966.
- Damian, Elena, Bart Meuleman, and Wim van Oorschot. 2019. "Transparency and Replication in Cross-National Survey Research: Identification of Problems and Possible Solutions." *Sociological Methods & Research*. doi:10.1177/0049124119882452.
- Ding, Peng, and Tyler J. VanderWeele. 2016. "Sensitivity Analysis without Assumptions." *Epidemiology* 27(3):368–77.
- Elwert, Felix. 2013. "Graphical Causal Models." Pp. 245–73 in *Handbook of Causal Analysis for Social Research*, edited by S. L. Morgan. Dordrecht, the Netherlands: Springer.
- Elwert, Felix, and Christopher Winship. 2014. "Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable." *Annual Review of Sociology* 40(1):31–53. doi:10.1146/annurev-soc-071913-043455.
- Freese, Jeremy, and David Peterson. 2017. "Replication in Social Science." *Annual Review of Sociology* 43(1):147–65. doi:10.1146/annurev-soc-060116-053450.
- Ganghof, Steffen. 2005. "Kausale Perspektiven in der vergleichenden Politikwissenschaft: X-zentrierte und Y-zentrierte Forschungsdesigns." Pp. 76–93 in *Vergleichen in der Politikwissenschaft*, edited by S. Kropp and M. Minkenberg. Wiesbaden: VS Verlag für Sozialwissenschaften.
- Gelman, Andrew, and Eric Loken. 2014. "The Statistical Crisis in Science." *American Scientist* 102(6):460–65.
- Grosz, Michael P., Julia M. Rohrer, and Felix Thoemmes. 2020. "The Taboo against Explicit Causal Inference in Nonexperimental Psychology." *Perspectives on Psychological Science* 15(5): 1243–55. <https://doi.org/10.1177/1745691620921521>.
- Harada, Masataka. 2013. "Generalized Sensitivity Analysis and Application to Quasi-experiments." Working Paper, National Graduate Institute for Policy Studies, Tokyo, Japan.
- Healy, Kieran, and James Moody. 2014. "Data Visualization in Sociology." *Annual Review of Sociology* 40(1):105–28. doi:10.1146/annurev-soc-071312-145551.
- Huntington-Klein, Nick, Andreu Arenas, Emily Beam, Marco Bertoni, Jeffrey R. Bloem, Pralhad Burli, and Naibin Chen, et al. 2021. "The Influence of Hidden Researcher Decisions in Applied Microeconomics." *Economic Inquiry* 59(3):944–60. doi:10.1111/ecin.12992.
- Imbens, Guido W. 2003. "Sensitivity to Exogeneity Assumptions in Program Evaluation." *American Economic Review* 93(2):126–32.
- Keele, Luke, Randolph T. Stevenson, and Felix Elwert. 2020. "The Causal Interpretation of Estimated Associations in Regression Models." *Political Science Research and Methods* 8(1):1–13. doi:10.1017/psrm.2019.31.
- Kohler, Ulrich, Tim Sawert, and Fabian Class. 2018. "Bring Research Design Back in: How DAGs Help to Identify and Solve Flaws in Covariate Selection." Working Paper, University of Potsdam.
- Landy, Justin F., Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, and Magnus Johannesson, et al. 2020. "Crowdsourcing Hypothesis Tests: Making Transparent How Design Choices Shape Research Results." *Psychological Bulletin* 146(5):451–79. doi:10.1037/bul0000220.
- Lundberg, Ian, Rebecca Johnson, and Brandon M. Stewart. 2021. "What Is Your Estimand? Defining the Target Quantity Connects Statistical Evidence to Theory." *American Sociological Review*. OnlineFirst. <https://doi.org/10.1177/00031224211004187>.
- Merton, Robert K. 1973. *The Sociology of Science: Theoretical and Empirical Investigations*. Chicago: University of Chicago Press.

- Mood, Carina. 2010. "Logistic Regression: Why We Cannot Do What We Think We Can Do, and What We Can Do about It." *European Sociological Review* 26(1):67–82.
- Muñoz, John, and Cristobal Young. 2018. "We Ran 9 Billion Regressions: Eliminating False Positives through Computational Model Robustness." *Sociological Methodology* 48(1):1–33. doi:10.1177/0081175018777988.
- Open Science Collaboration. 2015. "Estimating the Reproducibility of Psychological Science." *Science* 349(6251):aac4716. doi:10.1126/science.aac4716.
- Pager, Devah, and Hana Shepherd. 2008. "The Sociology of Discrimination: Racial Discrimination in Employment, Housing, Credit, and Consumer Markets." *Annual Review of Sociology* 34(1):181–209. doi:10.1146/annurev.soc.33.040406.131740.
- Pope, Bryson, and Nolan G. Pope. 2014. "Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results." OSF. August 24. osf.io/gvm2z.
- Popper, Karl R. 1959. *The Logic of Scientific Discovery*. London: Routledge.
- Price, Joseph, and Justin Wolfers. 2010. "Racial Discrimination among NBA Referees." *The Quarterly Journal of Economics* 125(4):1859–87.
- Reed, Robert W. 2019. "Takeaways from the Special Issue on the Practice of Replication." *Economics: The Open-Access, Open-Assessment E-Journal* 13(2019-13).
- Rogers, William H. 1993. "Regression Standard Errors in Clustered Samples." *Stata Technical Bulletin* 13:19–23.
- Silberzahn, Raphael, and Eric L. Uhlmann. 2015. "Crowdsourced Research: Many Hands Make Tight Work." *Nature* 526(7572):189–91. doi:10.1038/526189a.
- Silberzahn, Raphael, Eric L. Uhlmann, David P. Martin, Pablo Anselmi, Frederik Aust, Eli Awtrey, and Štěpán Bahník, et al. 2018a. "Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results." *Advances in Methods and Practices in Psychological Science* 1(3):337–56. doi:10.1177/2515245917747646.
- Silberzahn, Raphael, Eric L. Uhlmann, David P. Martin, Pablo Anselmi, et al. 2018b. "Many Analysts, One Dataset: Making Transparent How Variations in Analytical Choices Affect Results." OSF. August 24. osf.io/gvm2z.
- Steegen, Sara, Francis Tuerlinckx, Andrew Gelman, and Wolf Vanpaemel. 2016. "Increasing Transparency through a Multiverse Analysis." *Perspectives on Psychological Science* 11(5):702–12.
- Uhlmann, Eric Luis, Charles R. Ebersole, Christopher R. Chartier, Timothy M. Errington, Mallory C. Kidwell, Calvin K. Lai, Randy J. McCarthy, Amy Riegelman, Raphael Silberzahn, and Brian A. Nosek. 2019. "Scientific Utopia III: Crowdsourcing Science." *Perspectives on Psychological Science* 14(5):711–33. doi:10.1177/1745691619850561.
- Wooldridge, Jeffrey M. 2013. *Introductory Econometrics. A Modern Approach*. 5th ed. Mason, OH: South-Western.
- Young, Cristobal. 2018. "Model Uncertainty and the Crisis in Science." *Socius* 4:2378023117737206. doi:10.1177/2378023117737206.
- Young, Cristobal, and Katherine Holsteen. 2017. "Model Uncertainty and Robustness: A Computational Framework for Multimodel Analysis." *Sociological Methods & Research* 46(1):3–40. doi:10.1177/0049124115610347.
- Young, Cristobal, and Sheridan A. Stewart. 2021. "Multiverse Analysis: Advancements in Functional Form Robustness." Cornell University, Ithaca, NY. Unpublished working paper.

### Author Biographies

**Katrin Auspurg** holds a full professorship in sociology at the Department of Sociology at the LMU Munich, Germany. She investigates how inequalities in the labor market and the family affect each other. In addition, her current projects advance innovative experimental (survey) methods that allow the testing of causal mechanisms that explain social inequalities or subtle forms of discrimination. Recent publications examine the fairness of earnings and gender status beliefs (*American Sociological Review*), the gendered division of housework (*Social Science Research*), and ethnic discrimination on housing markets (*Journal of Ethnic and Migration Studies*).

**Josef Brüderl** holds a full professorship in sociology at the Department of Sociology at the LMU Munich, Germany. He works on methods for collecting and analyzing panel data. His substantive interests are in labor markets and family research. Recent publications examine the male marital wage premium (*American Sociological Review*), factors affecting second and third birth rates in West Germany (*Journal of Family Research*), and methods for collecting event history data with panel surveys (mda: *methods, data, analyses*).