Universiteit Leiden

Faculteit der Sociale Wetenschappen

Blind regression analysis to counter p-hacking in psychology

An illustration and evaluation of blinding methods

Veronique Verhees

Master’s Thesis Psychology,

Methodology and Statistics Unit, Institute of Psychology

Faculty of Social and Behavioral Sciences, Leiden University

Date: 31-12-2019

Student number: s2311895

Supervisor: Dr. A.E. van ‘t Veer

Acknowledgements

I would like to give a special thanks to my thesis supervisor, Dr. Anna E. van ‘t Veer. This master thesis would not have been possible without her valuable feedback and guidance through each stage of the process. From our first meeting onwards, I truly enjoyed the enthusiasm in which she talked about my research topic and all the other open science initiatives she is actively involved in. Her genuine interest in publishing some of the discoveries of my master’s thesis, was for me an extra motivational support to get the best out of this research.

In addition, I would like to thank my second reader Sjoerd M. H. Huisman for the intellectual meeting in which we brainstormed about some difficulties I encountered during the project. Together we came up with some insightful data blinding ideas, which I doubt I would have come up with without his guidance.

Last but not least, I would like to express my gratitude to my lovely parents Frans and Helmie, brother Teun and boyfriend Lars. Words cannot express the importance of their unconditional support and encouragement which they gave me during my master’s thesis but also during the rest of my bachelor’s and master’s. Where I sometimes forgot to believe in myself, they always believed in me and cheered me up. Thanks!

Abstract

Manipulating the analysing process in order to achieve statistical significance—also known as p-hacking—is a familiar practice among psychological researchers. This needs to be countered in order to enhance the reliability and robustness of the psychological research field. Pre-registration has recently attracted widespread attention as a measure to counter p-hacking. However, a drawback of this practice is that it can be challenging to make all analytic decisions before having a chance to explore the dataset. A measure called blind analysis may bypass this drawback and may therefore serve as a decent alternative to prevent p-hacking. In blind analysis, some alterations, unknown to the data analyst, are made to the data values and/or data labels of the dataset. This allows the analyst to explore (at least some of) the characteristics of the data and thereby make analytic decisions, while not seeing any of the main-outcomes and thereby not having a chance to p-hack. Although blind analysis is a common practice among physicists to cope with p-hacking, it has barely been introduced in psychology yet. In order to facilitate the use of blind analysis among psychologists, this master's thesis proposes five data blinding methods which are tailor-made to the specifics of a regression research design. We illustrate the blinding methods by applying them to simulated data from a hypothetical regression research design and by presenting how the data as well as the regression results have changed due to the blinding. In addition, we evaluate the blinding methods based on two criteria: (a) are the methods able to hide the main results, so that p-hacking can be countered and, (b) do the methods ensure that the major assumptions of a linear regression are still checkable, so that the analytic decisions based on these assumptions can still be made.

Keywords: p-hacking, blind analysis, data blinding methods, linear regression

Table of Contents

1. Introduction

The analysis part of scientific research is often seen as an objective process. However, datasets seldom can be analysed in only one way (Silberzahn et al., 2018). As a data analyst, there are many semi-subjective decisions to make: How to deal with outliers? How to handle missing data? Which covariates to use? And so on. For typical experiments in psychology, Wicherts et al. (2016) identified 15 of such analytic decisions, also referred to as researchers degrees of freedom (Simmons, Nelson, & Simonsohn, 2011). Almost without exception, psychological researchers do not make all these decisions prior to the analysis. This freedom of analysis allows researchers to explore various analytic alternatives (e.g., they might delete outliers in various ways, try various inclusion and exclusion criteria or apply various data transformations), with the goal to obtain a combination of analytic decisions that yield significant or otherwise desirable results. In the context of significance testing, Simonsohn, Nelson, and Simmons (2014) call this practice 'p-hacking'[1]. They define p-hacking as exploiting—perhaps unconsciously—researchers degrees of freedom in order to achieve statistically significance. In this master's thesis, I will propose, illustrate and evaluate methods that researchers can use to protect themselves against p-hacking during regression analyses: methods of 'data blinding'.

A measure against p-hacking is needed, as it seems to be a common questionable research practice (QRP) among psychologists. A frequently cited study by John, Loewenstein, and Prelec (2012), in which more than 2,000 US psychologists were surveyed, showed that 58% of the psychologists admitted that they had decided whether to collect more data after looking to see whether the results were significant, 22.5% stopped collecting data earlier than planned because one found the result that one had been looking for and 42.4% decided

[1] Also sometimes referred to as *bias* (Ioannidis, 2005), *significance chasing* (Ware & Munafò, 2015), *data snooping* (White, 2000) or *publication bias in situ* (Phillips, 2004).

whether to exclude data after looking at the impact of doing so on the results. Recently, the survey used by John et al. (2012) has been criticized as it would consist of ambiguous questions and misleading response patterns. Also, the participants were not given the opportunity to explain when and why they used a certain practice. A revised version of the survey, made by Fiedler and Schwarz (2016), revealed radically lower levels of self-reported QRP use than originally reported by John et al. (2012). But nonetheless, the results still show that the prevalence of QRPs, such as p-hacking, among psychologists is something to worry about.

This high prevalence of p-hacking among psychological researchers is probably due to today’s academic environment where the pressure to publish-or-perish is immense. Papers are more likely to be published if they report positive results (i.e., results that support the tested hypotheses) (Fanelli, 2012), also referred to as publication bias. As p-hacking can maximize the probability of positive results, researchers can—either consciously or unconsciously—be attracted to this data-manipulation strategy (Agnoli, Wicherts, Veldkamp, Albiero, & Cubelli, 2017; Bakker, Van Dijk, & Wicherts, 2012; Fanelli, 2010; John et al., 2012). For individual researchers, p-hacking therefore seems rational in the sense that it maximizes their chances of academic survival. However, when each researcher acts this way, the cumulative effect of p-hacking can be very problematic.

The primary problem is that p-hacking contributes to an enormous amount of papers with false positive results in the literature. A false positive is a result that incorrectly supports the tested hypothesis, also referred to as Type I error in the language of null hypothesis significance testing. The fact that psychological science suffers from a lot of false positives can be seen by the alarmingly low replication rate in this research field. The Open Science Collaboration (2015) showed for instance that out of 100 replication studies, only 39 supported the conclusions drawn in the original study. That p-hacking is one of the contributing factors to this, is emphasized by several studies (DeCoster, Sparks, Sparks, Sparks, & Sparks, 2015; Ioannidis, 2005; Simmons et al., 2011). Simmons et al., (2011) illustrated for instance how p-hacking, by exploiting four common research degrees of freedom, influences the probability of a false positive result. The flexibility in reporting subsets of experimental conditions leads to the highest probability of obtaining a false positive finding, followed by respectively the flexibility in using covariates, choosing among dependent variables and choosing sample size. When a researcher uses a combination of these four degrees of freedom, the probability to falsely detect a significant effect becomes even more stunning: it reaches over 60%.

Several studies showed that the false positive findings resulting from p-hacking lead to a ‘bump’ just below .05 in the distribution of p-values (also referred to as the ‘p-curve’ by Simonsohn et al. (2014)) (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Leggett, Thomas, Loetscher, & Nicholls, 2013; Masicampo & Lalande, 2012). However, more recent analyses by Bishop and Thompson (2016) and Hartgerink, Van Aert, Nuijten, Wicherts, and Van Assen (2016) demonstrated that the ‘bump’ is neither necessary nor sufficient to indicate false positives resulting from p-hacking. A possible explanation for this might be that there are different ‘types’ of p-hacking behaviours (MacCoun, 2019): someone can p-hack until the .05 threshold is crossed, which indeed lead to a ‘bump’ just below .05, but someone can also p-hack harder and push the p-value even closer to zero, which does not lead to the ‘bump’ just below .05. Whether the false positive findings resulting from p-hacking have a p-value just below .05 or even closer to zero, false positive findings will always be bad for science.

False positives are especially detrimental to science as they are very persistent once they have entered the literature. The main reason for this is that there is, especially in psychological research, little incentive to replicate research (Makel, Plucker, & Hegarty, 2012) and even when research is replicated, early positive studies often receive more attention than later negative ones (Asendorpf et al., 2013). All those false positives that remain in the literature make the psychological research field less reliable, hinder scientific progress and can inspire investment in fruitless research programs. To reduce the amount of false positive results, solutions that can counter p-hacking must be implemented.

One potential solution to counter p-hacking that has recently received an increasing amount of attention is pre-registration (Wagenmakers, Wetzels, Borsboom, Van der Maas, & Kievit, 2012). In pre-registration, researchers specify their analysis plan a priori. This way, the researchers degrees of freedom decreases and the researcher is less able to p-hack. Recently, pre-registration of studies has become more accessible for researchers and is stimulated by an increasing number of journals (Center for Open Science, n.d.). This facilitated a rapid increase in the practice; nowadays, there are already more than 8,000 preregistrations on the Open Science Framework for studies across all research areas (Nosek, Ebersole, DeHaven, & Mellor, 2018). Although pre-registration of studies is becoming routine in many fields, a lot of researchers still shy away from pre-register practices. The main reason for this is that making decisions before having a chance to explore the dataset can be very challenging, especially the first time. It takes practice to make analytic decisions in advance and it takes experience to learn what contingencies are most important to anticipate (Nosek et al., 2019). As a result, it can seem infeasible for researchers to perform a pre-registration, especially for young researchers for whom the pressure to publish-or-perish is already larger than ever. Therefore, a measure that can counter p-hacking, while still having some freedom to explore the characteristics of the dataset, would be most ideal.

A less well-known and less frequently used method, which might be able to serve this goal is blind analysis (Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun & Perlmutter, 2015; 2017). To perform a blind analysis, some alterations, made by a *data manager* (and therefore unknown to the data analyst), are made to the data values and/or data labels of the original dataset. Despite the alterations, most properties of the original data stay intact, which enables the data analyst to explore and handle the 'perturbed' data as usual (e.g., data is cleaned, outliers are handled, transformations are made et cetera). However, the perturbations in the dataset ensure that the effects of all the analytic decisions on the main outcomes of the study cannot be seen (e.g., it is unknown if a transformation has yielded towards a significant or otherwise desirable result) until the blind is 'lifted'. This way, the researcher is not able to p-hack any of the main-results.

Blind analysis was first introduced in the physical sciences. In a particle physics experiment at Stanford University, a group of physicists searched for fractional charges in ordinary matter[2]. The group claimed that there was an existence of fractional charges of $\frac{1}{3}e$, as several tests yielded this result (LaRue, Phillips, & Fairbank, 1981). However, there were some concerns about how this result was obtained; large corrections had to be applied to the raw data by the experimenters themselves and it was suspected that maybe the experimenters were (unconsciously) applying corrections to the raw data until the value turned out to be $\frac{1}{3}e$ (Heinrich, 2003; Lyons, 2008). To circumvent this potential problem, a 'blind test' was incorporated by subsequent studies of the Stanford group. During this blind test a random value was added to the raw data by an independent person. The experimentalists subsequently made all appropriate corrections to this perturbed dataset. After the experimentalists were confident that they have made all appropriate corrections, the random noise was removed and the result of the corrections to the raw data were revealed. It turns out that the studies that incorporated the blind test, did not confirm the original 'discovery' of fractional charges equal to $\frac{1}{3}e$. (Phillips, Fairbank, & Navarro, 1988).

[2] As it is beyond the scope of this master's thesis, we will not discuss the theoretical details of fractional charges. See Goldhaber (2003) for an elaborate description and definition of fractional charges.

After this, blind analyses were increasingly used in areas of particle and nuclear physics (e.g., Heinrich, 2003; Klein & Roodman, 2005; Lyons, 2008). In recent years, a new kind of analytic culture has even emerged in those fields: blind analyses are often considered as the only way to trust results. In addition, blind analyses are becoming more and more widely used in some clinical-trial protocols. The term 'triple-blinding' is often used in this field (Miller & Stewart, 2011), referring to the procedure in which the participants, the experimenters as well as the data analyst are blind to the experimental manipulations. However, triple blinding has not yet become as widespread as single and double blinding. In other research fields, including psychology, blind analyses are remarkably rare. A literature search revealed only three psychological studies reporting blind analyses (Dutilh et al., 2017; Moher, Lakshmanan, Egeth, & Ewen, 2014; van Dongen-Boomsma, Vollebregt, Slaats-Willemse, & Buitelaar, 2013). This while blind analyses seem potentially useful for psychological science as well.

The fact that psychologists do not often blind their data is probably in part due to the scarcity of studies that clarify the importance of blinding and that provide a clear explanation on how to apply blinding techniques. MacCoun and Perlmutter (2015; 2017) were the first that drew attention for performing blind analyses among psychologists. In addition, their article and book chapter, described four methods of data blinding:

- *Adding noise to outcome scores.* In this blinding method, a random number (drawn from an appropriate statistical distribution) is sampled for each subject. Subsequently, a subject's outcome score is averaged with the corresponding random number.
- *Scrambling outcome scores.* In this blinding method, the outcome scores are randomly shuffled over all the subject numbers. Consequently, a subject's outcome score no longer corresponds to its real outcome score, except by chance. This blinding method

is also referred to as item scrambling or row scrambling in MacCoun and Perlmutter (2015; 2017).

- *Adding bias to cells*. In this blinding method, a random number (drawn from an appropriate statistical distribution) is sampled for each experimental condition/cell. Subsequently, a subject's outcome score is averaged with the random number belonging to its experimental condition.
- *Scrambling cells*. In this blinding method, the labels of the experimental conditions/cells are randomly shuffled.

Recently, Dutilh et al., (2019) discussed the same four blinding techniques.

These three existing studies regarding data blinding in psychology (i.e., Dutilh et al., 2019; MacCoun & Perlmutter, 2015 2017) are mainly focused on how to blind data of an experimental research design. A specific feature of such 'experimental data' is that it consists of only categorical predictor variables (i.e., groups or experimental conditions). T-tests and analyses of variance (ANOVAs) are therefore often used to predict the outcome variable. The studies correctly remark that, due to this specific feature, it is not straightforward that the four methods are also applicable to blind data from other research designs. The list of blinding methods is thus definitely not exhaustive and other blinding methods, tailored to the features of other research design, have to be explored in the future. Since a regression design is, in addition to an experimental design, the most commonly used research design in psychological research (Kashy, Donnellan, Ackerman, & Russell, 2009), it seems a logical next step to explore techniques that can blind data of a regression design as well. A specific feature of such 'regression' data is that it consists of at least one continuous predictor variable (in addition, there may also be some categorical (i.e., dummy) predictor variables). Therefore, regression analyses (including factor analyses, path analyses and structural equation modelling), instead of t-tests and ANOVAs, are usually used to predict the outcome variable.

To the best of our knowledge, this study is the first (at least in psychological science) that proposes methods to blind data of a regression research design. In short, the proposed blinding methods work as follows:

- *Adding bias to coefficients.* In this blinding method, a regression model is fit to the original data. Each estimated regression coefficient is multiplied by a unique random number (drawn from an appropriate statistical distribution). Subsequently, new scores on the outcome variable are simulated according to a regression model with the biased regression coefficients (but the original predictor scores and residuals). Note that this blinding method somewhat resembles the 'adding bias to cells'-method proposed by Dutilh et al., (2019) and MacCoun and Perlmutter (2015; 2017). However, as data of a regression design does not consist of cells, bias is now added to the regression coefficients instead of to the cells.
- *Creating new coefficients.* In this blinding method, a regression model is fit to the original data. Each estimated coefficient is replaced by a unique random value (drawn from an appropriate statistical distribution), which represents the 'new' regression coefficient. Subsequently, new scores on the outcome variable are simulated according to a regression model with the new regression coefficients (but the original predictor scores and residuals).
- *Scrambling predictor variables.* In this blinding method, a regression model is fit to the original data. Subsequently, the labels of the predictor variables entered into the model are randomly shuffled, so that there is a possibility that the coefficients no longer belong to the correct predictor variable. Note that this blinding method somewhat resembles the 'scrambling cells'-method proposed by Dutilh et al., (2019) and MacCoun and Perlmutter (2015; 2017). However, as data of a regression design

does not consist of cells, predictor variables (and therefore the regression coefficients) are now scrambled instead of the cells.

In addition to these three blinding methods, the ‘adding noise to outcome scores’-method and the ‘scrambling outcome scores’-method proposed by Dutilh et al., (2019) and MacCoun and Perlmutter (2015; 2017) seem to be applicable to blind data of a regression research design as well (at least when the outcome variable is continuous and not binary).

The goal of my master’s thesis is twofold. First of all, we want to illustrate the five proposed blinding methods—tailored to the specific characteristics of a regression research design—so that researchers will get familiar with the blinding methods and performing a blind regression analyses will be facilitated for researchers. To reach this goal, we will apply each blinding method to simulated data from a hypothetical regression research design and we will present how the data as well as the regression results change due to the blinding.

A second goal of this master’s thesis is to evaluate the proposed blinding methods, so that researchers are informed about which blinding method is best in use. To reach this goal, we will evaluate the blinding methods based on two criteria. First, we will evaluate for each blinding method if (the p-values of) the effects are completely hidden and untraceable, so that p-hacking can be countered. Second, we will evaluate for each blinding method if the major assumptions of a linear regression, which are (a) no multicollinearity, (b) no unusual observations, (c) normality, (d) linearity, and (e) constant error variance (see Part III of Fox, 2016), are still checkable even though the data is blinded, so that the analytic decisions based on these assumptions (i.e., whether, and if so, which variable(s) to remove, which observation(s) to remove, which variable(s) to transform et cetera) can still be made. A blinding method that meets both criteria is most ideal, as it gives the researcher some explorative freedom without being able to p-hack any of the main-results.

2. Method

To illustrate and evaluate the five blinding techniques, we simulated data from a hypothetical regression research design. The simulation of the data as well as the illustration and evaluation of the blinding techniques were performed in R, version 3.6.1 (R Core Team, 2019). The R-script can be downloaded online: https://osf.io/4ay8w/.

2.1. Simulating data from a hypothetical regression research design

Data was simulated based on the following regression research design, which is hypothetical, but not unlike many studies in social psychology. A work and organisational psychologist wants to predict the sick leave of employees (i.e., the number of days a year that they stay at home due to illness) with four predictor variables: (a) the gender; (b) the general health; (c) the experience of stress at work; and (d) the variety of work activities of the employee. *Gender* was dummy variable coded with 0 for males and 1 for females. The other three predictors were all measured on a 7-point Likert scale. The psychologist hypothesizes that *gender* and *stress at work* have a positive effect on *sick leave* (i.e., female employees stay more days at home due to illness than male employees and the more stress the employee experiences at work, the more days a year the employee stays at home due to illness) and that *general health* and *variety of work activities* have a negative effect on *sick leave* (i.e., the better the employee's general health and/or the more varied the employee's work activities are, the fewer days a year the employee stays at home due to illness).

As is common with regression, the values on the predictor variables are considered fixed and known quantities, so we simulated those first. Since *gender* was considered as a dummy variable, we simulated this variable as such that it was binominal distributed with a probability of .5 for males and a probability of .5 for females. Since we assumed that the other three predictor variables were measured on a 7-point Likert scale, we simulated those variables as such that they were all truncated normally distributed with a mean of 4, a standard deviation of 2, a lower truncation point of 1 and an upper truncation point of 7. In addition, the predictor-variable values of those three predictors were rounded to the nearest integer (i.e., no decimals) to reflect the interval scale.

After simulating the predictor variables, we simulated the outcome variable (i.e., *sick leave*) according to the following linear model,

$$\text{Sick leave} = 12 + 3 * \text{gender} - 2 * \text{general health} + 1.5 * \text{stress at work} -$$
$$0.5 * \text{variety of work activities}, \text{where } \varepsilon \sim N(0, \sigma = 4)$$

which is, in terms of the direction of the regression weights, consistent with the abovementioned hypotheses of the psychologist. The intercept of 12 was chosen as this ensures that the values on the outcome variable have a minimum of zero (which makes most sense as the minimum days a year an employee can stay at home due to illness equals zero).

We simulated a sample size of *N* = 85 as this sample size seems reasonably representative to the ‘median’ sample size of psychological studies (Marszalek, Barber, Kohlhart, & Holmes, 2011). We aimed to apply the blinding methods on simulated data that is quite representative to real psychological data, as this will facilitate researchers to apply a blinding method to their own dataset. A power analysis, performed with the ‘pwr’-package (Champely, 2018), revealed that the statistical power for a study with a sample size of 85 and four predictor variables was less than adequate for detecting a small effect (i.e., power of .14) at an alpha level of .05, whereas the power was more than adequate for detecting a medium (i.e., power of .80) and large effect (i.e., power of .99) at an alpha level of .05. The benchmarks for ‘small’, ‘medium’ and ‘large’ effects were f^2 = .02, f^2 = .15, and f^2 = .35, respectively (Cohen, 1988).

For present purposes, we limit ourselves to a single run of the simulation and do not consider asymptotic properties or sensitivity analyses of the various parameters of the simulation. The dataset resulting from the single simulation is hereinafter referred to as the ‘raw dataset’.

2.2. Illustrating the blinding methods

To illustrate the five proposed blinding methods—tailored to the specific characteristics of a regression research design—we blinded the raw dataset according to each blinding method (numbering is arbitrary):

- *Blinding method 1: Adding noise to outcome scores.* In order to blind the raw data, we averaged each of the (non-missing) 85 raw outcome scores (i.e., scores on *sick leave*) together with one of 85 random scores. The random scores were sampled from a normal distribution, in which the mean and standard deviation were equal to those of the raw outcome variable[3]. We deliberately chose to add noise to the data points of the outcome variable instead of the data points of the predictor variable(s), as the latter would make it impossible to check for multicollinearity and high-leverage observations (see paragraph 2.4).
- *Blinding method 2: Scrambling outcome scores.* In order to blind the data, we left the raw outcome scores intact, but we ‘re-randomized’ (or ‘scrambled’) the (non-missing) outcome scores over all the subjects. Consequently, the outcome score for any given subject no longer matches with its real outcome score, except by chance. Again, we deliberately chose to scramble the data points of the outcome variable instead of the data points of the predictor variable(s), as the latter would make it impossible to check

[3] We decided to sample values from a normal distribution in which the mean and standard deviation were equal to those of the raw outcome variable, as our raw outcome score is normally distributed and this way the blind outcome score will be normally distributed as well with approximately the same mean and standard deviation. However, if preferable, it is also possible to sample the random values from another statistical distribution (i.e., a statistical distribution which fits better to the distribution of the outcome variable of the data).

for multicollinearity and high-leverage observations (see paragraph 2.4).

- *Blinding method 3: Adding bias to coefficients.* In order to blind the data, we started with composing the linear regression model we are planning to fit to the data: a model with *sick leave* as outcome variable and *gender*, *general health*, *stress at work* and *variety of work activities* as predictor variables. We fitted this linear regression model to the raw data using a least squares approach and extracted the residuals and the estimated regression coefficients from this (raw) model. The extracted residuals were kept unchanged. However, we did change the estimated coefficients: we added random noise to the coefficients by multiplying each coefficient with a unique random value, drawn from a uniform distribution with a minimum of -2 and a maximum of 2[4]. Subsequently, the raw data was blinded by simulating new values for (the non-missing scores of) the outcome variable (indicated by 'sick leave*') according to the following model:

$$\text{Sick leave}^{*} = \text{B1}^{*} * \text{gender} + \text{B2}^{*} * \text{general health} + \text{B3}^{*} * \text{stress at work} + \\ \text{B4}^{*} * \text{variety of work activities} + \varepsilon,$$

where ε represents the extracted residuals from the raw model and B1^{*}, B2^{*}, B3^{*} and B4^{*} represent the new, biased coefficients (i.e., the extracted coefficients from the raw model multiplied by a random value between -2 and 2). The four predictor variables are equal to the original predictor variables of the raw data.

[4] We decided to sample values between -2 and 2, as this implies that the coefficients can become larger (if the absolute value is bigger than 1) or smaller (if the absolute value is smaller than 1) and that the direction of the coefficients can change from positive to negative or vice versa. However, if preferable it is also possible to sample the random values from a uniform distribution with another minimum and/or maximum or to sample the random values from another statistical distribution.

- *Blinding method 4: Creating new coefficients.* This blinding method is in some way similar to the previous blinding method. In order to blind the data, we started again with fitting the composed linear regression model (i.e., a model with *sick leave* as outcome variable and *gender*, *general health*, *stress at work* and *variety of work activities* as predictor variables) to the raw data, using a least squares approach. However, a difference of this method with respect to the previous one is that this method only extracts the residuals from the model, not the estimated regression coefficients. The new coefficients are namely simulated without reference to the original estimated coefficients: we just 'create' new coefficients instead of adding bias to the original coefficients. We created the new, yet standardized, coefficients by simulating a unique random value, drawn from a uniform distribution with a minimum of -.5 and a maximum of .5[5], for each coefficient. To un-standardize these new coefficients, we multiplied each coefficient by the ratio of the standard deviations of the outcome variable and the (corresponding) predictor variable. Subsequently, the raw data was blinded by simulating new values for (the non-missing scores of) the outcome variable (indicated by 'sick leave*') according to the following model:

$$\text{Sick leave}^* = \text{B1}^* * \text{gender} + \text{B2}^* * \text{general health} + \text{B3}^* * \text{stress at work} + \text{B4}^* * \text{variety of work activities} + \varepsilon,$$

where ε represents the extracted residuals from the raw model and B1*, B2*, B3* and B4* represent the new, unstandardized coefficients. The four predictor variables are

[5] We decided to sample standardized coefficients between -.5 and .5, as this implies that the effects can become both positive and negative with a range from a small to medium effect size (a standardized coefficient is in many ways equivalent to the Cohen's d and a Cohen's d of ± .5 equals a medium effect size (Cohen, 1988)). However, if preferable it is also possible to sample the random values from a uniform distribution with another minimum and/or maximum or to sample the random values from another statistical distribution.

equal to the original predictor variables of the raw data.

- *Blinding method 5: Scrambling predictor labels.* In order to blind the data, we again started with composing the linear regression model we are planning to fit to the data: a model with *sick leave* as outcome variable and *gender, general health, stress at work* and *variety of work activities* as predictor variables. Subsequently, we 're-randomized' (or 'scrambled') the labels of the predictor variables that are entered in the model, so that there is a possibility that the coefficients no longer correspond to the correct predictor variable. For our composed linear regression model with four entered predictor variables, there are 4! = 24 possible ways to scramble their labels. As one way will be equivalent to the labelling of the raw dataset, there will be a 1 in 24th chance that the raw dataset appears as 'blinded' dataset.

After creating the five blinded datasets, a linear regression model with *sick leave* as outcome variable and *gender, general health, stress at work* and *variety of work activities* as predictor variables was fitted to the raw dataset and the blinded datasets, using a least squares approach.

As a result, we were able to present how (a) the data, as well as (b) the regression results, have changed due to the blinding. For presenting the changes in regression results, the t-statistics of the four raw effects were compared with the t-statistics of the four blinded effects. We have chosen to compare the t-statistics, because based on the t-statistics we can indirectly conclude something about a possible change in (a) the direction of the effects, i.e., a t-statistic below zero indicates a negative effect and a t-statistic above zero indicates a positive effect; (b) the strength of the effects, i.e., given our fixed sample size, the higher the absolute t-statistic the stronger the effect; and (c) the significance of the effects, i.e., a t-statistic below -1.990 or above 1.990 (i.e., the critical t-statistics) indicates significance when we test two-tailed with 80 degrees of freedom (i.e., $n - k - 1$, where n equals the sample size of 85 and k equals the amount of predictors, which is 4) and an alpha level of .05.

2.3. Evaluating whether the blinding methods counter p-hacking

In this master's thesis we assume that p-hacking is countered when a researcher is not able to infer (an estimate of) the p-values of the raw effects based on the blind effects he sees. This will be the case when the p-values of the blind effects are totally *independent* on the raw effects. In contrast, we assume that p-hacking is not countered by a blinding method when a researcher is able to infer (an estimate of) the p-values of the raw effects based on the blind effects he sees. This will be the case when the p-values of the blind effects are in a certain way *dependent* on the raw effects (given that the researcher is familiar with this dependency).

To gain more insight into this (in)dependency between the raw effects and blinded effects and therefore to evaluate whether or not p-hacking can be countered, we followed the following procedure for each blinding method: we applied the blinding method 1000 times to the raw dataset, which resulted in 1000 different blinded datasets (i.e., the particular sequence of random noise that has been added, the particular way the outcome scores have been scrambled, the particular sequence of bias terms that has been added to the coefficients, the particular sequence of new coefficients that have been created or the particular way the predictor-variable labels have been scrambled provides each time a (slightly) different blinded dataset). Subsequently, the linear regression model (with *sick leave* as outcome variable and *gender*, *general health*, *stress at work* and *variety of work activities* as predictor variables) was fitted to the 1000 blinded datasets, using a least squares approach. The resulting 1000 t-statistics of each of the four blind effects were then extracted. Based on these 1000 t-statistics, we calculated for each of the four effects the probability that their blind effect is significant at an alpha level of .05 (i.e., t-statistic below -1.990 or above 1.990: the critical t-statistics when we test two-tailed with 80 degrees of freedom).

Subsequently, we compared those probabilities with the raw effects. If it appears that the strongest and weakest raw effect have about the same probability that their blind effect will be significant, this will imply that the p-values of the blind effects are totally *independent* on the raw effects. Consequently, a researcher is not able to infer an estimate of the p-values of the raw effects based on the blind effects he sees and therefore we have to conclude that p-hacking is fully countered by the blinding method. In contrast to this, if it appears that the strongest raw effect has clearly the highest probability that their blind effect will be significant and the weakest raw effect has clearly the lowest probability that their blind effect will be significant, this will imply that the p-values of the blind effects are *dependent* on the raw effects. Given that the researcher is familiar with this dependency, he is able to infer an estimate of the p-values of the raw effect based on the blind effects he sees and therefore we have to conclude that p-hacking is not fully countered by the blinding method.

2.4. Evaluating whether the blinding methods allow to check assumptions

The major assumptions of a linear regression, taken from Part III of Fox (2016) are (a) no multicollinearity, (b) no unusual observations, (c) normality, (d) linearity, and (e) constant error variance. Below, the five major assumptions are briefly summarized and is explained how we evaluated whether the assumptions can still be checked (and therefore whether the analytic decisions based on these assumptions can still be made) even though the data is blinded.

2.4.1. No multicollinearity

Multicollinearity is the occurrence of high intercorrelations among predictor variables in a multiple regression model. According to Fox (2016) the most basic and simplest way to test multicollinearity is to calculate a variance-inflation factor (VIF). A VIF quantifies how much the variance (i.e., the standard error) of an estimated regression coefficient is inflated, as a result of the correlations between the predictor variables in the model. A VIF above 10 is a commonly used cut-off point for determining the presence of multicollinearity (Pallant, 2013). If it appears that the results exceed the recommended cut-off point, it may be a logical (analytical) decision to combine the intercorrelated predictor variables or to remove one of them from the model. Other options might be to perform a variable selection method (e.g., stepwise regression), to perform a biased estimation method (e.g., ridge regression) or to add additional prior information external to the data at hand (Fox, 2016).

As VIFs are a common way to measure multicollinearity and our goal is to decide if multicollinearity can still be checked even though the data is blinded, we examined for each blinded model if their VIFs are the same as the VIFs of the raw model. If so, it was concluded that multicollinearity can still be checked and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, which variable(s) to combine or to exclude from the model) can still be made.

2.4.2. No unusual observations

When talking about unusual observations, it is useful to distinguish among (a) high-leverage observations, (b) regression outliers and (c) influential observations.

In least squares regression, high-leverage observations are observations with unusual combinations of predictor-variable values. The so-called ‘hat-value’ is a common measure of leverage in regression (Fox, 2016). A hat-value measures the distance of the predictor-variable values of an observation from the centroid (point of means) of the predictor variables, taking into account the correlational and variational structure of the predictor variables. Fox (2016) suggests that hat-values exceeding about twice the average hat-value ($\bar{h} = (k + 1)/n$, where k equals the number of predictor variables and n equals the sample size) are noteworthy. In small samples, $3 * \bar{h}$ can be used as cut-off. If it appears that an observation exceeds the recommended cut-off point, it may be a logical (analytical) decision to remove the high-leverage observation from the model.

A regression outlier can be defined as an observation with an unusual outcome-variable value given its combination of predictor-variable values. To identify outliers, Fox

(2016) recommend to investigate the studentized residuals (sometimes referred to as deleted studentized residual or externally studentized residual) of each observation. A studentized residual is a deleted residual (i.e., a residual for an observation when that observation is excluded from the calculation of the regression coefficients) divided by its estimated standard deviation. Observations that have a studentized residual outside the ±2 range are considered statistically significant at the .05 alfa level (Fox, 2016). If it appears that an observation exceeds the recommended cut-off point, it may be a logical (analytical) decision to remove this outlier from the model.

Lastly, influential observations are observations that have, as the name suggests, a substantial influence on the regression coefficients. An observation is influential when it has a high-leverage and is an outlier. The Cook's distance statistic is a common measure to identify influential observations (Fox, 2016): it summarizes directly how much all of the fitted values change when an observation is deleted. Both the hat-values (to measure the leverage of the observations) and the residuals (to measure the outlyingness of the observations) are present in the formula of the Cook's distance. A cut-off rule of thumb is that an observation is influential when Cook's distance is larger than $4/(n - k - 1)$, where n equals the sample size and k equals the number of predictor variables (Fox, 2016). Again, if it appears that an observation exceeds the recommended cut-off point, it may be a logical (analytical) decision to remove this influential observation from the model.

As hat-values, studentized residuals and Cook's distances are a common way to detect respectively high-leverage observations, outliers and influential observations, and our goal is to decide if we can still check for such unusual observations even though the data is blinded, we examined for each blinded model if their index plot of hat-values (i.e., a scatterplot of hat-values versus the observation indices), their index plot of studentized residuals (i.e., a scatterplot of studentized residuals versus the observations indices) and their index plot of

Cook's distances (i.e., a scatterplot of Cook's distances versus the observations indies) are the same as the index plots of the raw model. If so, it was concluded that high-leverage observations, outliers and influential observations can still be (visually) checked and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, which observation(s) to exclude from the model) can still be made.

2.4.3. Normality

The assumption of normality states that the residuals should be normally distributed. A quantile-comparison plot, which compares the sample distribution of the studentized residuals with the quantiles of the t-distribution for $n - k - 2$ degrees of freedom (where *n* equals the sample size and *k* equals the number of predictor variables), is according to Fox (2016) an effective method to graphically examine the distribution of the residuals. There is normality if the plotted points in the quantile-comparison plot are reasonably linear. If it appears that the residuals are not normally distributed, it may be a logical (analytical) decision to transform the data (Fox, 2016).

As a quantile-comparison plot is a common way to examine the normality and our goal is to decide if normality can still be checked even though the data is blinded, we examined for each blinded model if their quantile-comparison plot is the same as the quantile-comparison plot of the raw model. If so, it was concluded that normality of residuals can still be (visually) checked and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable(s)) can still be made.

2.4.4. Linearity

The linearity assumption states that there must be a linear relationship between the outcome variable and the predictor variables. Component-plus-residual plots, also called partial residual plots, are often effective in determining the linearity of a multiple regression model (Fox, 2016). In a component-plus-residual plot, the partial residuals of a predictor variable (i.e., the linear component from the partial regression plus the least-squares residuals) are plotted against the corresponding predictor-variable values and a line of best fit is added. There is linearity between the outcome variable and the predictor variables if the plotted points in the component-plus-residual plots are reasonably linear. If it appears that there is no linearity, it may be a logical (analytical) decision to transform the data. Another option might be to alter the form of the model (e.g., include a quadratic term in a predictor variable) (Fox, 2016).

As component-plus-residual plots are a common way to examine the linearity and our goal is to decide if linearity can still be checked even though the data is blinded, we examined for each blinded model if the component-plus-residual plot of *stress at work* is the same as the component-plus-residual plot of *stress at work* of the raw model[6]. If so, it was concluded that linearity of residuals can still be (visually) checked and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable(s)) can still be made.

2.4.5. Constant error variance

The constant error variance (also called homoscedasticity) assumption states that the variance of the residuals around the predicted outcome scores should be the same for all predicted scores. According to Fox (2016), a pattern of changing error variance can be easily detected in a plot of studentized residuals against fitted values. There is constant error variance when the studentized residuals are roughly rectangular distributed around the 0 line. If it appears that there is no constant error variance, it may be a logical (analytical) decision to transform the data. Other options might be to substitute the ordinary least squares estimation

[6] For the sake of example, we only compared the component-plus-residual plots of *stress at work,* as the same conclusions will be drawn when the component-plus-residual plots of the other predictor variables are compared.

with a weighted-least squares estimation or to correct the coefficient standard errors (Fox, 2016).

As a plot of studentized residuals against fitted values is a common way to examine constant error variance and our goals is to decide if constant error variance can still be checked even though the data is blinded, we examined for each blinded model if their plot of studentized residuals against fitted values is the same as the plot of studentized residuals against fitted values of the raw model. If so, it was concluded that constant error variance can still be (visually) checked and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable(s)) can still be made.

3. Results

3.1. Blinding method 1: Adding noise to outcome scores

3.1.1. Illustration of blinding method

Table 1 presents the changes in the outcome variable when each of the (non-missing) 85 raw outcome scores (i.e., scores on *sick leave*) were averaged with one of 85 random scores, sampled from a normal distribution in which the mean and standard deviation were equal to those of the raw outcome variable. As can be seen, the predictor-variable values remain unchanged when applying this blinding technique.

Table 2 shows the t-statistics of the effects of *gender*, *general health*, *stress at work* and *variety of work activities* when the linear regression model was fitted to the raw dataset (first row) and to the dataset where random noise was added to the outcome scores (second row). Comparing these t-statistics, we can conclude that the direction of all effects has remained unchanged: *gender* and *stress at work* still have a positive effect on *sick leave* and *general health* and *variety of work activities* still have a negative effect on *sick leave*. In addition, the three strongest effects (i.e., *gender*, *general health* and *stress at work*) have become weaker due to the blinding (i.e., they are driven closer to zero), while the weakest effect (i.e., *variety of work activities*) has become a bit stronger due to the blinding. Even though some effects have weakened, all effects are still significant at an alpha level of .05. In addition to the changes in effects, the residuals and fitted values of the linear regression model have changed as well due to the blinding.

Table 1

Changes in dataset when noise is added to the outcome scores

ID	**Gender**	**General health**	**Stress at work**	**Variety of work activities**	**Sick leave**	**Sick leave (blinded)**
1	0	4	5	4	10	(10 + 6) ÷ 2
2	1	6	4	5	4	(4 + 8) ÷ 2
3	0	2	5	5	14	(14 + 18) ÷ 2
4	1	5	2	6	2	(2 + 10) ÷ 2
5	1	5	2	5	5	(5 + 10) ÷ 2
6	0	2	3	2	14	(14 + 18) ÷ 2
…	…	…	…	…	…	…
85	0	3	2	3	2	(2 + 8) ÷ 2

Note. The red-coloured column represents the outcome-variable values of the raw dataset while the green-coloured column represents the outcome-variable values of the blinded dataset.

3.1.2. Evaluation of p-hacking

As can be seen in Figure 1 (blue curves), the 1000 blind t-statistics belonging to an effect are normally distributed. The standard deviation of the blind t-statistics is approximately the same for each of the four effects. However, the average blind t-statistic is clearly different for each of the four effects: the average blind t-statistics is higher when the t-statistic of the raw data is higher. This implies that the stronger the raw effect (i.e., the higher the t-statistic of the raw effect), the higher the probability that the blind effect will be significant at an alpha level of .05 (see Figure 1). To illustrate, the raw effect of *general*

Table 2

T-statistics, p-values and VIFs of raw model and blinded models

	Gender			General health			Stress at work			Variety of work activities		
Model	*t*	*p*	*VIF*	*t*	*p*	*VIF*	*t*	*p*	*VIF*	*t*	*p*	*VIF*
Raw	3.986^{***}	<.001	1.021	-6.554^{***}	<.001	1.048	4.996^{***}	<.001	1.057	-2.227^{*}	.029	1.058
Add noise to outcome scores	2.223^{*}	.029	1.021	-5.394^{***}	<.001	1.048	3.233^{**}	.002	1.057	-2.410^{*}	.018	1.058
Scramble outcome scores	.651	.517	1.021	-1.118	.267	1.048	1.199	.234	1.057	-.042	.967	1.058
Add bias to coefficients	-3.387^{**}	.001	1.021	-7.558^{***}	<.001	1.048	-1.819^{**}	.073	1.057	-3.412^{**}	.001	1.058
Create new coefficients	-2.660^{**}	.009	1.021	3.563^{***}	.001	1.048	-1.120	.266	1.057	4.711^{***}	<.001	1.058
Scramble predictor labels	4.996^{***}	<.001	1.057	-2.227^{*}	.029	1.058	3.986^{***}	<.001	1.021	-6.554^{***}	<.001	1.048

Note. *VIF* = Variance-inflation factor.

**p* < .05 (two-tailed), ***p* < .01 (two-tailed) ****p* < .001 (two-tailed).

health is the strongest (i.e., has the largest absolute t-statistic equal to -6.554) and the probability that the blind effect of *general health* is—like the raw effect—significant is equal to .99. The raw effect of *variety of work activities* is the weakest (i.e., has the smallest absolute t-statistic equal to -2.227) and the probability that the blind effect of *variety of work activities* is—like the raw effect—significant is only .21. Theoretically, when a raw t-statistic is equal to zero, it will be very unlikely that the blind effect is significant.

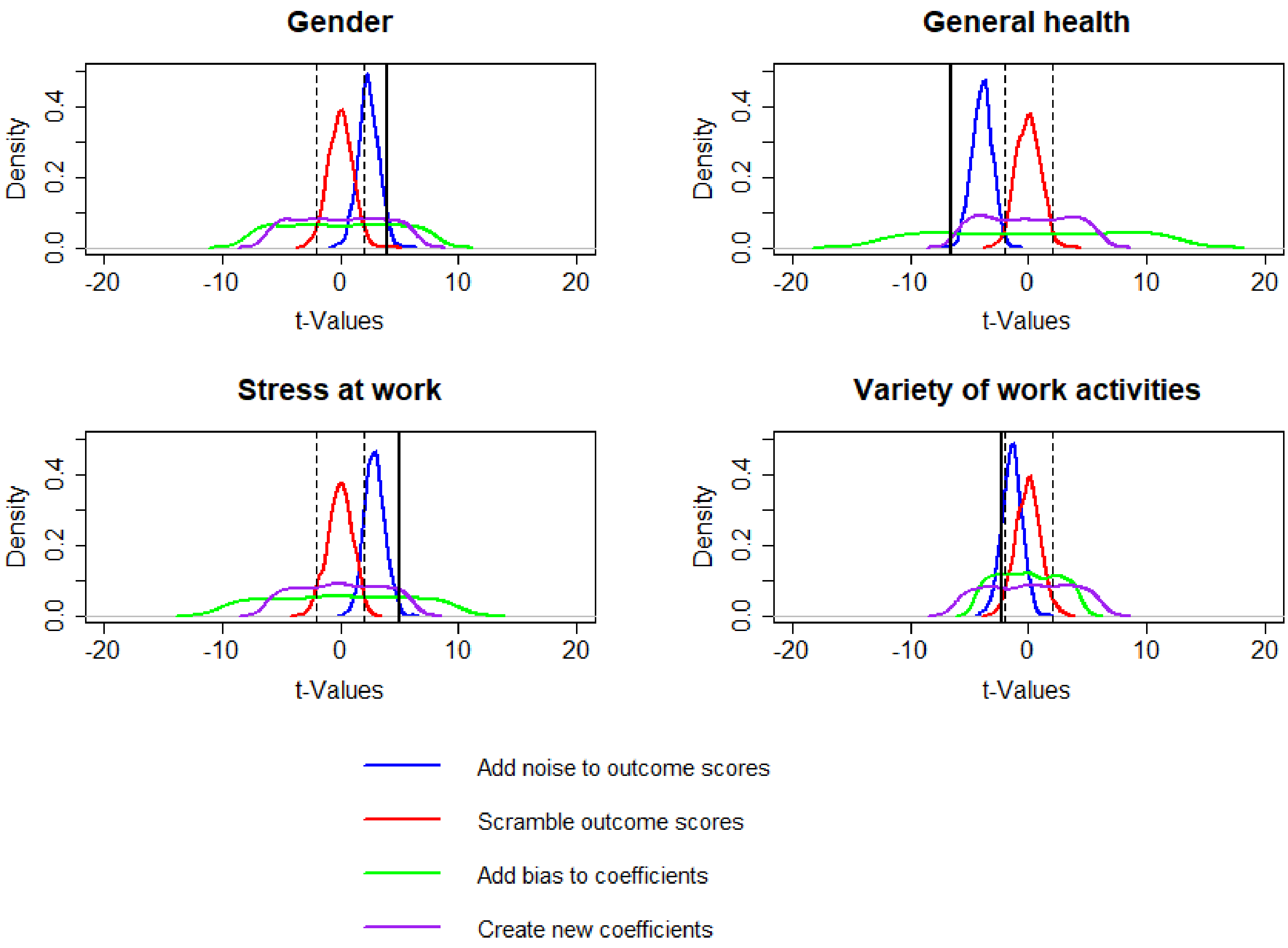


Figure 1. For each blinded model, the density distributions of the 1000 t-statistics of each of the four blind effects are displayed. The bold vertical line represents the t-statistic of the raw effect. The broken vertical lines are drawn at -1.990 and 1.990 and represent the critical t-statistics for a two-tailed test with 80 degrees of freedom (i.e., $n - k - 1$, where n equals the sample size of 85 and k equals the amount of predictors, which is 4) and an alpha level of .05.

There is thus a certain dependency between the p-values of the blind effects and the raw effects. When a researcher is familiar with this dependency, he is able to infer an estimate of the p-values of the raw effects based on the blind effects he sees. Therefore, we must conclude that p-hacking is not fully countered by this blinding method. What makes p-hacking even easier, is the fact that the direction of the (non-zero) raw effects can be inferred by the blind effects as well (since blind effects are more likely to be in the same direction as their raw effects, see Figure 1). Because of this, a researcher can easily figure out how to exploit his researchers degrees of freedom to make effects stronger and/or (even more) significant.

3.1.3. Evaluation of checking assumptions

As can be seen in Table 2, the VIFs of the blinded model (second row) correspond to the VIFs of the raw model (first row). This makes sense as only the predictor-variable values are involved in calculating the VIFs and these values have not changed due to the blinding. The correspondence in VIFs implies that it is still possible to check for multicollinearity and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, which variable(s) to combine or to exclude from the model) can still be made.

By comparing Figure 2.a with Figure 2.b, one will notice that the hat-values of the observations do not differ between the two models either. This is because the hat-values are, like the VIFs, calculated based on the predictor-variable values only (and those have not changed due to the blinding). The resemblance in hat-values implies that it is still possible to check for high-leverage observations and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, which observation(s) to exclude from the model) can still be made. However, when comparing Figure 3.b with 3.a and Figure 4.b with 4.a, it will become clear that respectively the studentized residuals and Cook’s distances of the observations of the blinded model do not correspond to those of the raw model. This makes sense as both the studentized residuals and the Cook’s distances depend on the residuals, and there is a difference between the residuals of the blinded model compared with the residuals of the raw model. The differences in plots indicate that it is no longer possible to check for regression outliers and influential observations anymore when noise is added to the outcome scores.

Also, the quantile-comparison plot, the component-plus-residual plot of *stress at work* and the studentized residuals against fitted values plot of the blinded model (respectively Figure 5.b, 6.b and 7.b) do not match with those of the raw model (respectively Figure 5.a, 6.a and 7.a), enabling us to conclude that it is not possible to check for respectively normality, linearity and constant error variance anymore. Again, this can be explained by the fact that all plots are based on the residuals (and those have changed due to the blinding).

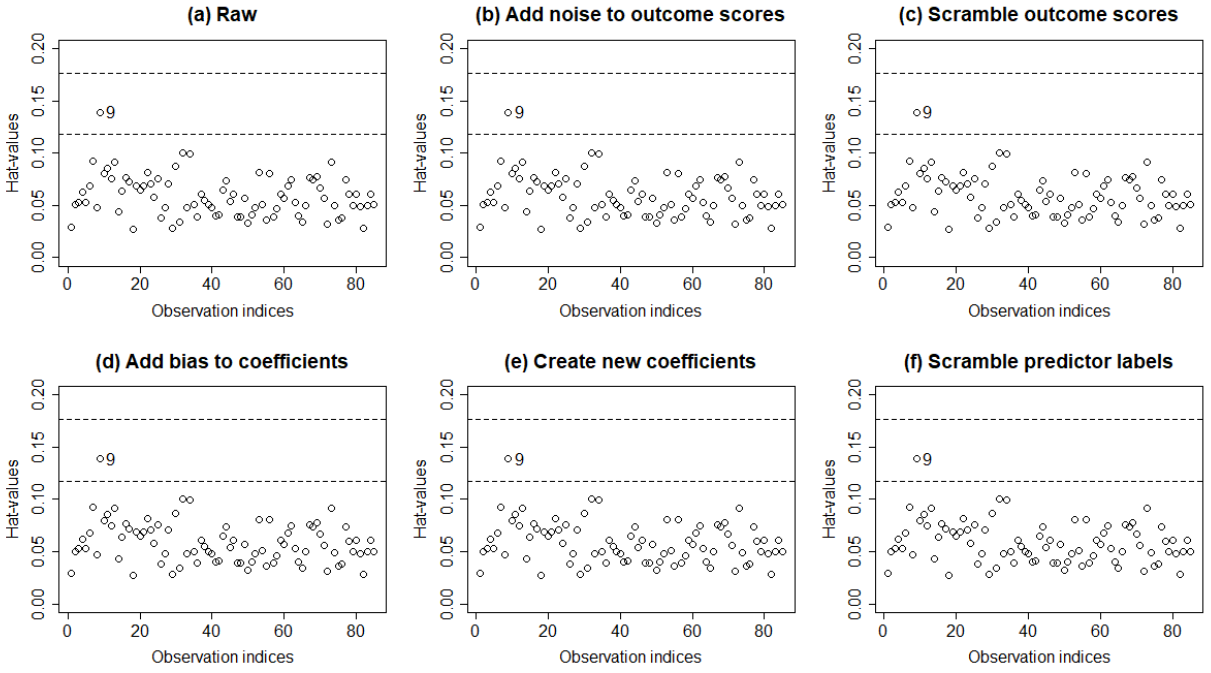


Figure 2. Index plot of hat-values of the raw model and the blinded models, with broken lines at the cut-off scores of $2{*}\bar{h}$ and $3{*}\bar{h}$. The observations that exceed the cut-off scores are identified.

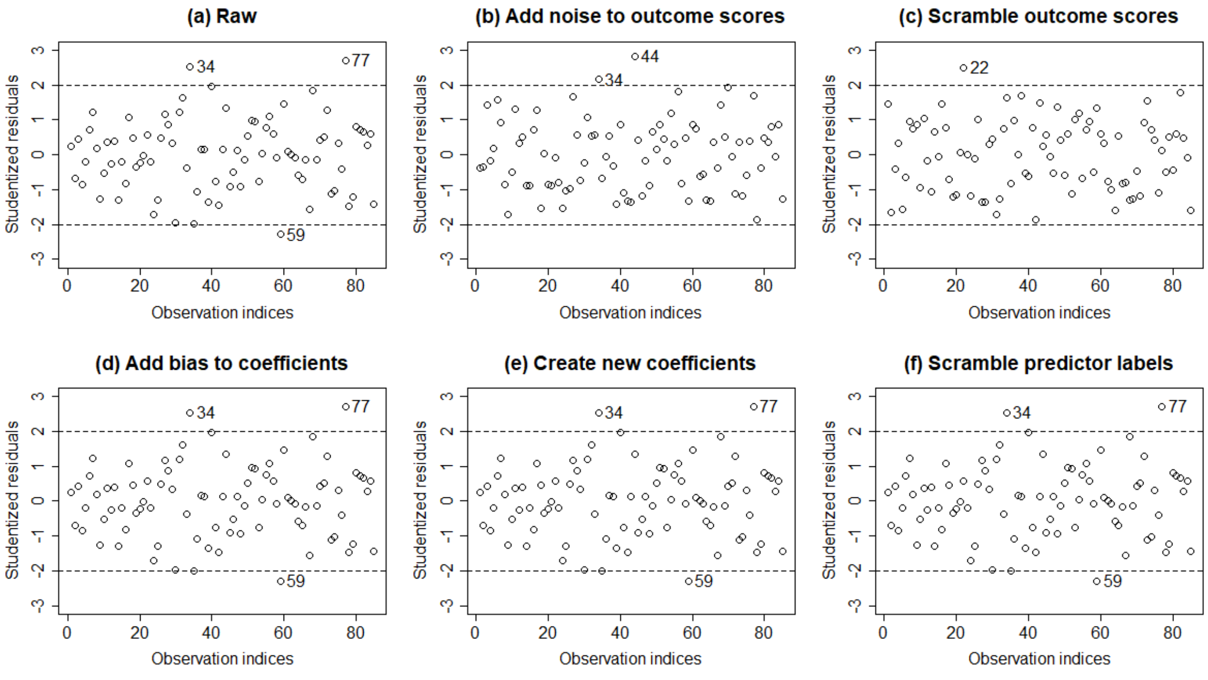


Figure 3. Index plot of studentized residuals of the raw model and the blinded models, with broken lines at the cut-off scores of ±2. The observations that exceed the cut-off scores are identified.

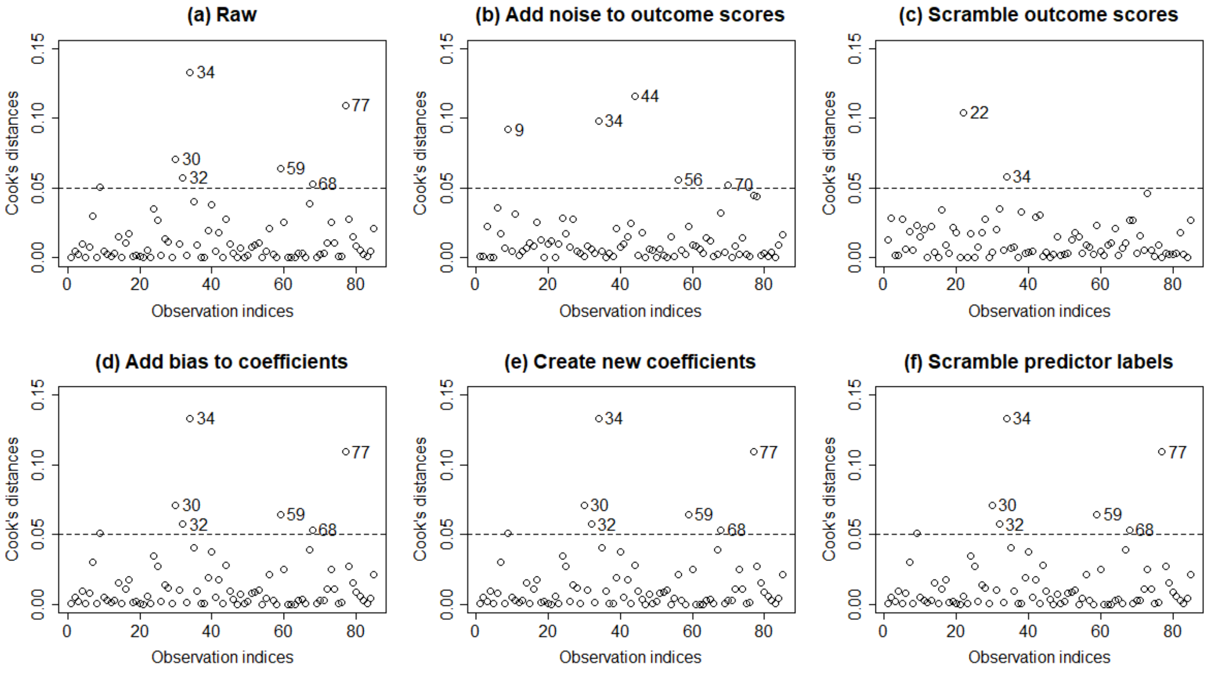


Figure 4. Index plot of Cook's distances of the raw model and the blinded models, with a broken line at the cut-off score of $4/(n - k - 1)$, where n equals the sample size and k the number of predictor variables. The observations that exceed the cut-off score are identified.

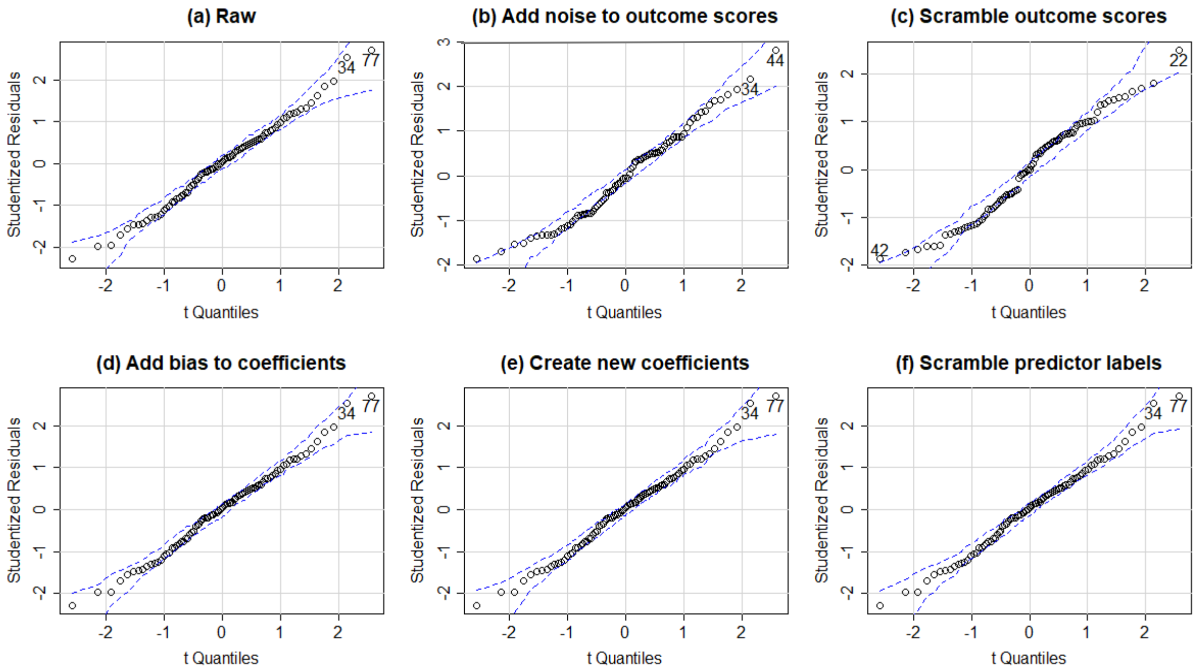


Figure 5. Quantile-comparison plot of the raw model and the blinded models. The broken lines in the plots represent a pointwise 95% simulated confidence envelope. The two observations with the most extreme absolute studentized residuals are identified.

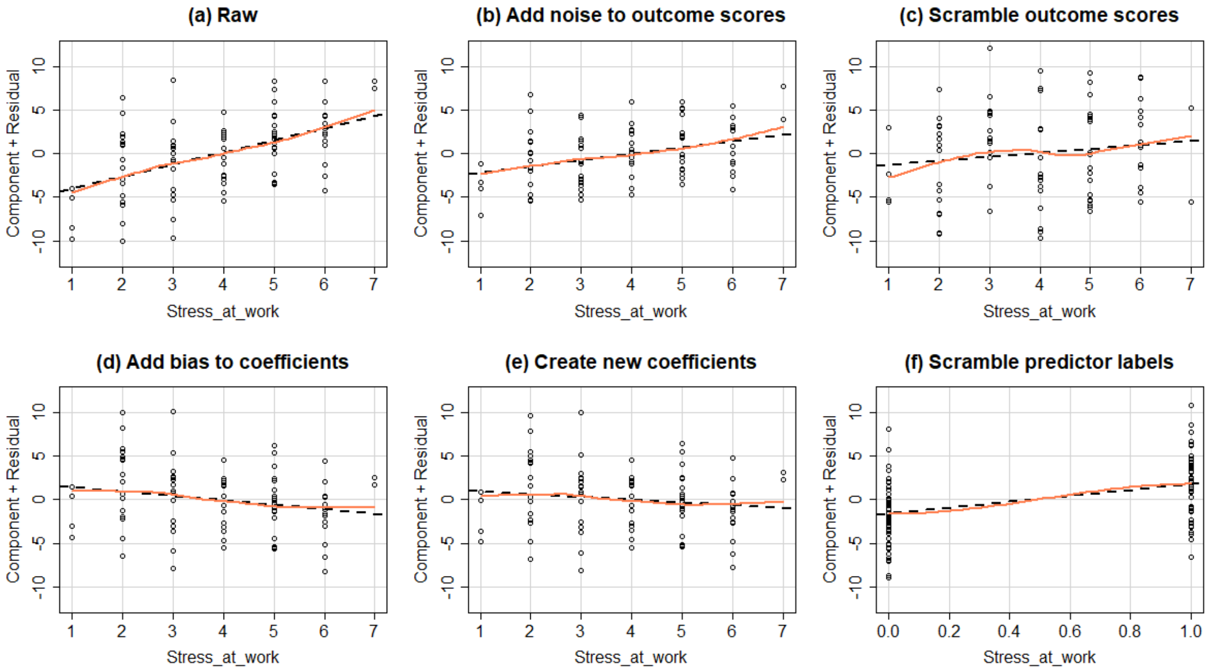


Figure 6. Component-plus-residual plot of *stress at work* of the raw model and the blinded models. The broken line represents the linear least squares line and the orange line represents the lowess line.

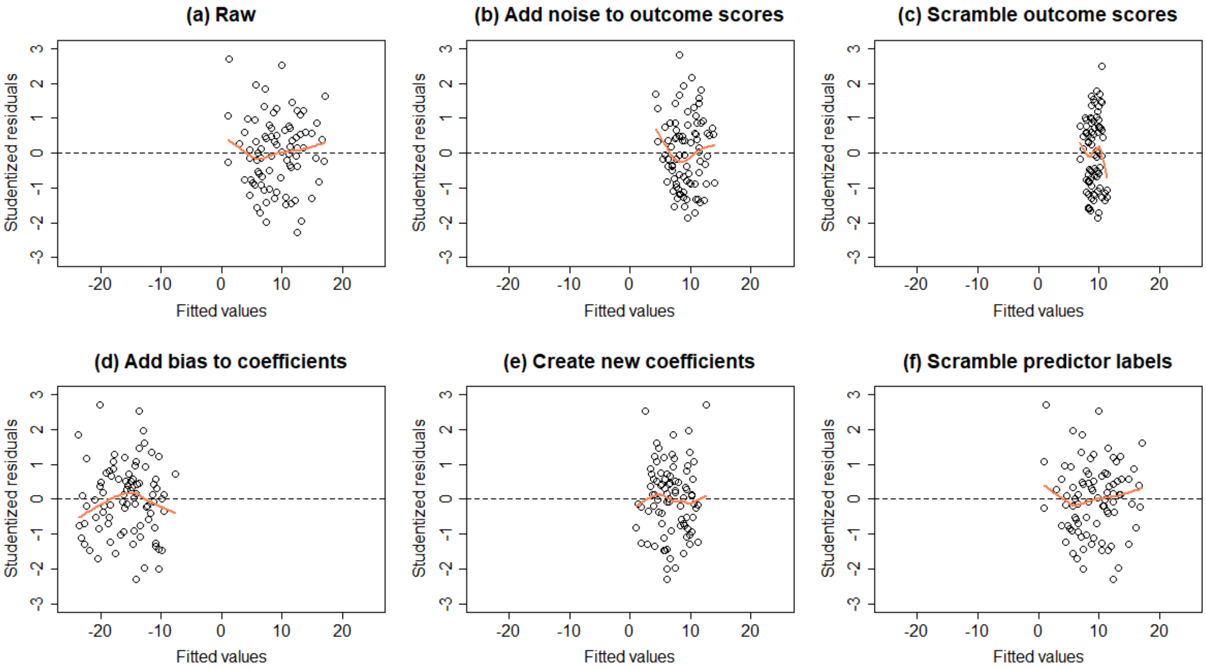


Figure 7. Plot of studentized residuals versus fitted values of the raw model and the blinded models. The broken line runs through the zero y-axis and the orange line represents the lowess line.

3.2. Blinding method 2: Scrambling outcome scores

3.2.1. Illustration of blinding method

Table 3 presents the changes in the outcome variable when each of the (non-missing) 85 raw outcome scores is randomly scrambled over all the subjects. As can be seen, the raw outcome score for any given subject no longer matches with its real outcome score, except by chance (see for example ID 85). The subject's predictor-variable values remain unchanged when applying this blinding technique.

Table 2 shows the t-statistics of the effects of *gender*, *general health*, *stress at work* and *variety of work activities* when the linear regression model was fitted to the raw dataset (first row) and to the dataset where the outcome scores were randomly scrambled (third row). Comparing these t-statistics, we can conclude that the direction of all effects has remained unchanged: *gender* and *stress at work* still have a positive effect on *sick leave* and *general health* and *variety of work activities* still have a negative effect on *sick leave*. In addition, all effects have become weaker due to the blinding. All effects have even been driven so close to zero that they are no longer significant at an alpha level of .05. In addition to the changes in effects, the residuals and fitted values of the linear regression model have changed as well due to the blinding.

Table 3

Changes in dataset when outcome scores are scrambled

ID	**Gender**	**General health**	**Stress at work**	**Variety of work activities**	**Sick leave**	**Sick leave (blinded)**
1	0	4	5	4	10	2
2	1	6	4	5	4	14
3	0	2	5	5	14	10
4	1	5	2	6	2	5
5	1	5	2	5	5	14
6	0	2	3	2	14	4
…	…	…	…	…	…	…
85	0	3	2	3	2	2

Note. The red-coloured column represents the outcome-variable values of the raw dataset while the green-coloured column represents the outcome-variable values of the blinded dataset.

3.2.2. Evaluation of p-hacking

As can be seen in Figure 1 (red curves), the 1000 blind t-statistics belonging to an effect are normally distributed. The mean (which is equal to zero) as well as the standard deviation of the blind t-statistics is approximately the same for each of the four effects. This implies that the strongest and weakest raw effect (i.e., raw effects with high and low t-statistics) have about the same probability that their blind effect will be significant at an alpha level of .05 (i.e., in our scenario always around .05, see Figure 1). To illustrate, the raw effect of *general health* is the strongest (i.e., has the largest absolute t-statistic equal to -6.554) and the probability that the blind effect of *general health* is—like the raw effect—significant is equal to .04. The raw effect of *variety of work activities* is the weakest (i.e., has the smallest absolute t-statistic equal to -2.227) and the probability that the blind effect of *variety of work activities* is—like the raw effect—significant is equal to .06.

There is thus an independency between the p-values of the blind effects and the raw effects. Consequently, a researcher is not able to infer (an estimate of) the p-values of the raw effects based on the blind effects he sees. Therefore, we must conclude that p-hacking is fully countered by this blinding method.

3.2.3. Evaluation of checking assumptions

For blinding method 2, the exact same conclusions are made regarding checking the assumptions as for blinding method 1. Again, the VIFs and hat-values of the blinded model (see respectively Table 2 (third row) and Figure 2.c) correspond to the VIFs and hat-values of the raw model (see respectively Table 2 (first row) and Figure 2.a), which makes sense as only the predictor-variable values are involved in calculating the VIFs and hat-values and these values have not changed due to the blinding. The resemblances imply that it is still possible to check for multicollinearity and high-leverage observations and that therefore the analytic decisions based on these assumptions can still be made.

In addition, the index plot of studentized residuals, the index plot of Cook's distances, the quantile-comparison plot, the component-plus residual plot of *stress at work* and the studentized residuals against fitted values plot of the blinded model (respectively Figure 3.c, 4.c, 5.c, 6.c and 7.c) do not match with those of the raw model (respectively Figure 3.a, 4.a, 5.a, 6.a and 7.a), enabling us to conclude that it is no longer possible to check for respectively regression outliers, influential observations, normality, linearity and constant error variance anymore when the data is blinded. This can be explained by the fact that all plots are based on the residuals and those have changed due to the blinding.

3.3. Blinding method 3: Adding bias to coefficients

3.3.1. Illustration of blinding method

In this blinding method, new values for (the non-missing scores of) the outcome variable (i.e., *sick leave*) are simulated according to a linear regression model with the raw predictor variables and the extracted residuals of the raw model, but with the new, biased coefficients (i.e., the extracted coefficients from the raw model multiplied by a random value between -2 and 2). All the predictor-variable values remain unchanged when applying this blinding technique.

Table 2 shows the t-statistics of the effects of *gender, general health, stress at work* and *variety of work activities* when the linear regression model was fitted to the raw dataset (first row) and to the dataset where bias was added to the coefficients (fourth row). Comparing these t-statistics, we can conclude that the direction of the effects of *gender* and *stress at work* has changed from positive to negative (i.e., the original coefficients are multiplied by a negative random value). As a result, all blind effects are negative. In addition, the effects of *general health* and *variety of work activities* have been strengthened (i.e., the original coefficients are multiplied by an absolute value bigger than 1), while the effects of *gender* and *stress at work* have been weakened by the blind (i.e., the original coefficients are multiplied by an absolute value smaller than 1). The effect of *gender* is still significant despite the weakening. However, the effect of *stress at work* has been driven so close to zero that it is no longer significant at an alpha level of .05. In addition to the changes in effects, the fitted values of the linear regression model have changed as well due to the blinding. However, contrary to blinding method 1 and 2, the blinded dataset is created as such that fitting the linear regression model to the blinded data will lead to the exact same residuals as fitting the regression model to the raw data.

3.3.2. Evaluation of p-hacking

As can be seen in Figure 1 (green curves), the 1000 blind t-statistics belonging to an effect are normally distributed. The average blind t-statistic (which is equal to zero) is the same for each of the four effects. However, the standard deviation of the blind t-statistics is clearly different for each of the four effects: the standard deviation is higher when the t-statistic of the raw data is higher. This implies that the stronger the raw effect (i.e., the higher the t-statistic of the raw effect), the higher the probability that the blind effect will be significant at an alpha level of .05 (see Figure 1). To illustrate, the raw effect of *general health* is the strongest (i.e., has the largest absolute t-statistic equal to -6.554) and the probability that the blind effect of *general health* is—like the raw effect—significant is equal to .85. The raw effect of *variety of work activities* is the weakest (i.e., has the smallest absolute t-statistic equal to -2.227) and the probability that the blind effect of *variety of work activities* is—like the raw effect—significant is only .55. Theoretically, when a raw t-statistic is equal to zero, it will be very unlikely that the blind effect is significant.

There is thus a certain dependency between the p-values of the blind effects and the raw effects. When a researcher is familiar with this dependency, he is able to infer an estimate of the p-values of the raw effects based on the blind effects he sees. Therefore, we must conclude that p-hacking is not fully countered by this blinding method. One thing that makes p-hacking a bit more challenging, is the fact that the direction of the raw effects cannot be inferred by the blind effects (since blind effects have a 50% chance of being positive and a 50% chance of being negative, see Figure 1). Because of this it is more challenging for a researcher to figure out how to exploit his researchers degrees of freedom to make effects stronger and/or (even more) significant. But by trial and error, a researcher will nevertheless find out, making p-hacking still possible.

3.3.3. Evaluation of checking assumptions

Like in blinding method 1 and 2, the VIFs and hat-values of the blinded model (see respectively Table 2 (fourth row) and Figure 2.d) correspond to the VIFs and hat-values of the raw model (see respectively Table 2 (first row) and Figure 2.a), which makes sense as only the predictor-variable values are involved in calculating the VIFs and hat-values and these values have not changed due to the blinding. The resemblances imply that it is still possible to check for multicollinearity and high-leverage observations and that therefore the analytic decisions based on these assumptions can still be made.

However, contrary to blinding methods 1 and 2, the studentized residuals and Cook's distances of the observations of the blinded model (respectively Figure 3.d and 4.d) correspond to those of the raw model (respectively Figure 3.a and 4.a) as well. This makes sense as both the studentized residuals and the Cook's distances depend on the residuals of the model and the blinded dataset is created as such that fitting the linear regression model to the blinded data will lead to the exact same residuals as fitting the regression model to the raw data. The resemblances in plots indicate that it is still possible to check for regression outliers and influential observations even though bias is added to the coefficients. The analytic decisions based on these assumptions (e.g., whether, and if so, which observation(s) to exclude from the model) can therefore still be made.

For the same reason (i.e., similarity in residuals), the quantile-comparison plot of the blinded model (Figure 5.d) matches with that of the raw model (Figure 5.a). This enables us to conclude that it is still possible to check for normality and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable(s)) can be made as well.

The component-plus-residual plot of *stress at work* of the blinded model (Figure 6.d) is partly the same and partly different than the component-plus-residual plot of *stress at work* of the raw model (Figure 6.a). The ‘residual’-part of the plot is the same: i.e., the residuals of the partial regression of *stress at work* have not changed due to the blinding. Therefore, the way the linear least squares lines (broken black line) and the line of best fit (orange line) runs through the points is the same between the two models. The resemblance in this ‘residual’-part of the plot enables to check whether there are deviations from linearity. The analytic decision regarding *whether* to transform a variable can therefore still be made. However, the ‘component’-part of the plot is different: i.e., the linear component from the partial regression of *stress at work* has become a bit weaker and has switched from positive to negative due to the bias that is added to the coefficient. Consequently, the strength and the direction of the linear least squares line and the line of best fit of the component-plus-residual plot differ between the two models. Because of this difference in the ‘component’-part of the plot, it is not possible to decide *how* to transform in case of non-linearity. Reason for this is that the direction of the bulge (i.e., the deviation of the line of best fit from the least squares line) indicates the direction of the power transformation of the outcome variable and/or the predictor variable(s) to straighten the relationship between them (Mosteller & Tukey, 1977). Because the direction of the bulge can differ between the models, a suggested transformation for the blinded data will not per definition straighten the relationship between the outcome variable and predictor variable of the raw data.

Unfortunately, it is no longer possible to check for constant error variance when bias is added to the coefficients. Although there is no difference between the residuals of the blinded model compared with the residuals of the raw model, there is a clear difference between the fitted values of the two models. Therefore, the studentized residuals against fitted values plot of the blinded model (Figure 7.d) does not correspond to the studentized residuals against fitted values plot of the raw model (Figure 7.a).

3.4. Blinding method 4: Creating new coefficients

3.4.1. Illustration of blinding method

In this blinding method, new values for (the non-missing scores of) the outcome variable (i.e., *sick leave*) are simulated according to a linear regression model with the raw predictor variables and the extracted residuals of the raw model, but with new, created coefficients (which are unstandardized equal to random values drawn from a uniform distribution with a minimum of -.5 and a maximum of .5). All the predictor-variable values remain unchanged when applying this blinding technique.

Table 2 shows the t-statistics of the effects of *gender, general health, stress at work* and *variety of work activities* when the linear regression model was fitted to the raw dataset (first row) and to the dataset where new coefficients were created (fifth row). Comparing these t-statistics, we can conclude that the direction of all effects has flipped: the direction of the effects of *gender* and *stress at work* has changed from positive to negative (i.e., the new created coefficients are negative) and the direction of the effects of *general health* and *variety of work activities* has changed from negative to positive (i.e., the new created coefficients are positive). In addition, the effects of *gender, general health* and *stress at work* have been weakened by the blind (i.e., the new created coefficients are smaller than the raw coefficients), while the effect of *variety of work activities* has been strengthened by the blind (i.e., the new created coefficient is larger than the raw coefficient). The effects of *gender* and *general health* are still significant despite the weakening. However, the effect of *stress at work* has been driven so close to zero that it is no longer significant at an alpha level of .05. In addition to the changes in effects, the fitted values of the linear regression model have changed as well due to the blinding. However, like in blinding method 3, the blinded dataset is created as such that fitting the linear regression model to the blinded data will lead to the exact same residuals as fitting the regression model to the raw data.

3.4.2. Evaluation of p-hacking

As can be seen in Figure 1 (purple curves), the 1000 blind t-statistics belonging to an effect are normally distributed. The mean (which is equal to zero) as well as the standard deviation of the blind t-statistics is approximately the same for each of the four effects. This implies that the strongest and weakest raw effect (i.e., raw effects with high and low t-statistics) have about the same probability that their blind effect will be significant at an alpha level of .05 (i.e., in our scenario always around .68, see Figure 1). To illustrate, the raw effect of *general health* is the strongest (i.e., has the largest absolute t-statistic equal to -6.554) and the probability that the blind effect of *general health* is—like the raw effect—significant is equal to .70. The raw effect of *variety of work activities* is the weakest (i.e., has the smallest absolute t-statistic equal to -2.227) and the probability that the blind effect of *variety of work activities* is—like the raw effect—significant is equal to .68.

There is thus an independency between the p-values of the blind effects and the raw effects. Consequently, a researcher is not able to infer (an estimate of) the p-values of the raw effects based on the blind effects he sees. Therefore, we must conclude that p-hacking is fully countered by this blinding method.

3.4.3. Evaluation of checking assumptions

For blinding method 4, the exact same conclusions are made regarding checking the assumptions as for blinding method 3. Again, the VIFs and hat-values of the blinded model (see respectively Table 2 (fifth row) and Figure 2.e) correspond to the VIFs and hat-values of the raw model (see respectively Table 2 (first row) and Figure 2.a), which makes sense as only the predictor-variable values are involved in calculating the VIFs and hat-values and these values have not changed due to the blinding. The resemblances imply that it is still possible to check for multicollinearity and high-leverage observations and that therefore the analytic decisions based on these assumptions can still be made.

In addition, the index plot of studentized residuals, the index plot of Cook's distances and the quantile-comparison plot of the blinded model (respectively Figure 3.e, 4.e, 5.e) correspond to those of the raw model (respectively Figure 3.a, 4.a, 5.a), enabling us to conclude that it is also possible to check for respectively outliers, influential observations and normality even though the data is blinded. The analytic decisions based on these assumptions can therefore still be made. The resemblances in plots can be explained by the fact that all plots are based on the residuals, and the blinded dataset is created as such that fitting the linear regression model to the blinded data will lead to the exact same residuals as fitting the regression model to the raw data.

Like in blinding method 3, the 'residual'-part of the component-plus-residual plot of *stress at work* of the blinded model (Figure 6.e) is the same as the 'residual'-part of the component-plus-residual plot of *stress at work* of the raw model (Figure 6.a), which enables to check whether there are deviations from linearity. The analytic decision regarding *whether* to transform a variable can therefore still be made. However, the 'component'-part of the component-plus-residual plot differs between the raw and blinded model. Consequently, the strength and the direction of the linear least squares line and the line of best fit of the component-plus-residual plot differ between the two models. Because of this, it is impossible to decide *how* to transform the variable in case of non-linearity.

Unfortunately, it is no longer possible to check for constant error variance when new coefficients are created. Although there is no difference between the residuals of the blinded model compared with the residuals of the raw model, there is a clear difference between the fitted values of the two models. Therefore, the studentized residuals against fitted values plot of the blinded model (Figure 7.e) does not correspond to the studentized residuals against fitted values plot of the raw model (Figure 7.a).

3.5. Blinding method 5: Scrambling predictor labels

3.5.1. Illustration of blinding method

Table 4 presents the changes in the dataset when the labels of the four predictor variables are randomly scrambled. As can be seen, the labels of the predictor variables do no longer match with their real predictor-variable labels: the predictor labels of *gender* and *stress at work* have been swapped and the predictor labels of *general health* and *variety of work activities* have been swapped. The values on the predictor variables as well as on the outcome variable remain unchanged when applying this blinding technique.

Table 4

Changes in dataset when predictor labels are scrambled

ID	**Gender**	**General health**	**Stress at work**	**Variety of work activities**	**Sick leave**
ID	**Stress at work (blinded)**	**Variety of work activities (blinded)**	**Gender (blinded)**	**General health (blinded)**	**Sick leave**
1	0	4	5	4	10
2	1	6	4	5	4
3	0	2	5	5	14
4	1	5	2	6	2
5	1	5	2	5	5
6	0	2	3	2	14
…	…	…	…	…	…
85	0	3	2	3	2

Note. The red-coloured row represents the labels of the predictor variables of the raw dataset while the green-coloured row represents the labels of the predictor variables of the blinded dataset.

Table 2 shows the t-statistics of the effects of *gender, general health, stress at work* and *variety of work activities* when the linear regression model was fitted to the raw dataset (first row) and to the datasets where the predictor-variable labels were randomly scrambled (sixth row). Comparing these t-statistics, we can conclude that the 'scrambling predictor labels'-method is undoubtedly able to change the direction and strength of the raw effects, but it does so in a different manner than the previous discussed blinding methods. As this method keeps the values of the outcome variable as well as the values of the predictor variables intact, fitting the linear model to the blinded dataset gives the exact same four t-statistics as fitting the model to the raw dataset. However, because the names of the predictor variables were scrambled, the t-statistics may no longer belong to the correct effect. To illustrate, the t-statistics of *gender* and *stress at work* have been swapped (as their variable names have been swapped) and the t-statistics of *general health* and *variety of work activities* have been swapped (as their variable names have been swapped). As no changes are made to the data-values, the residuals and fitted values of the linear regression model remain unchanged.

The purpose of this blinding method is of course that a researcher should not know which predictor variable belongs to which effect until the blind is lifted. However, it is important to note that this can only be hidden when predictor variables are measured on the same scale. When not, it is possible to discover what the raw effect of a predictor variable is by simply looking at the scale of the variable. To illustrate, a researcher analysing our blinded datasets will always be able to figure out what the raw effect of *gender* is, since *gender* is measured on a different scale than the other three variables. When looking at the predictor labels of the blinded dataset (see Table 4), it will immediately be clear that *stress at work* only has zeros and ones, while the variable is from origin measured on a 1-7 point Likert scale. This will reveal that the blind t-statistic of the effect of *stress at work* is in fact the raw t-statistic of the effect of *gender*. Standardizing the variables so they are all measured on the same scale does not solve the problem, as it is still possible to see that one variable has only two unique values (namely *gender*) while the others have seven unique variables (namely *general health, stress at work* and *variety of work activities*). As *general health*, *stress at work* and *variety of work activities* are, however, measured on the same scale, the 'scrambling predictor labels'-method is able to blind those three effects.

So, it is important to realize that the 'scrambling predictor labels'-method can only be used as blinding method when all predictors of interested are measured on the same scale. By illustrating and evaluating this blinding method, we therefore assume for the sake of example that the researcher is not interested in the effect of *gender* on *sick leave* (i.e., *gender* is a control variable in the model) and that it is therefore not a problem for now that the effect of *gender* is revealed.

3.5.2. Evaluation of p-hacking

The blinding method ensures that the blind effects are by definition the same as the raw effects. This enables the researcher to know how many effects are (in)significant at a certain alpha level. However, as the predictor labels are scrambled, the researcher cannot know which effect belongs to which variable. The researcher can therefore not define which predictor has the strongest effect and which predictors has the weakest effect. To illustrate, the blinded dataset consists of three blinded effects: one positive significant effect (t = 4.996) and two negative significant effects (t = -6.554 and t = -2.227). (Remember that the effect of gender is already revealed and therefore not 'blinded'). Despite the blinding, it will be immediately clear that all three effects are significant. However, it remains unclear which variable belongs to which effect. Based on hypotheses, a researcher could expect that the positive effect belongs to *stress at work*, that the strongest negative effect belongs to *general health* and that the weakest negative effect belongs to *variety of work activities*. However, this are still just speculations.

There is thus clearly a dependency between the p-values of the blind effects and the raw effects. Therefore, we must conclude that p-hacking is not fully countered by this blinding method. The extent to which this method cannot counter p-hacking depends on the ‘p-hack-aim’ of the researcher. When a researcher aims to achieve statistical significance for a *specific number of effects*—no matter which effects—this method is not able to counter p-hacking at all (as the researcher can exactly see the number of significant effects). However, when a researcher aims to achieve statistical significance for a *specific effect*, this method makes p-hacking at least more difficult (as the researcher cannot know for sure which effect belongs to which variable). P-hacking a specific effect becomes even more challenging when a lot of predictor variables are entered in the linear model. The more predictor variables, the lower the probability that a certain effect belongs to a certain variable and thus the harder it will be for a researcher to make expectations about the way the names of the predictors were scrambled.

3.5.3. Evaluation of checking assumptions

As can be seen in Table 2, the four VIFs of the blinded model (sixth row) correspond to the four VIFs of the raw model (first row). This makes sense as only the predictor-variable values are involved in calculating the VIFs and these values have not changed due to the blinding. However, as can be seen, the four VIFs do no longer belong to the correct predictor-variable name anymore. This implies that we can still check for multicollinearity and that the analytic decisions based on this assumption (e.g., whether, and if so, which variable(s) to combine or to exclude from the model) can still be made, but that the names of the involved variable(s) will not be clear for the researcher until the blind is lifted.

By comparing Figure 2.a with Figure 2.f, one will notice that the hat-values of the observations do not differ between the two models either. This is because the hat-values are, like the VIFs, calculated based on the predictor-variable values only (and those have not changed due to the blinding). The resemblance in hat-values implies that it is still possible to check for high-leverage observations and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, which observation(s) to exclude from the model) can still be made. Furthermore, the studentized residuals and Cook's distances of the observations of the blinded model (respectively Figure 3.f and 4.f) correspond to those of the raw model (respectively Figure 3.a and 4.a) as well. This makes sense as both the studentized residuals and the Cook's distances depend on the residuals of the model and, since no changes are made to the data-values, the residuals of the blinded and raw regression model are the same. The resemblances in plots indicate that it is also possible to check for regression outliers and influential observations even though the predictor-variable names are scrambled. The analytic decisions based on these assumptions (e.g., whether, and if so, which observation(s) to exclude from the model) can therefore still be made.

For the same reason (i.e., similarity in residuals), the quantile-comparison plot of the blinded model (Figure 5.f) matches with that of the raw model (Figure 5.a). This enables us to conclude that it is still possible to check for normality and that therefore the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable(s)) can still be made as well.

As can be seen in Figure 6, the component-plus-residual plot of *stress at work* of the blinded model (Figure 6.f) corresponds to the component-plus-residual plot of *gender* (see the 0/1 values) of the raw model (Figure 6.a). More generally, while the four component-plus-residual plots of the blinded model are the same as the four component-plus-residual plots of the raw model, the plots do no longer belong to the correct predictor-variable name anymore. This implies that we can still check for linearity and that the analytic decisions based on this assumption (e.g., whether, and if so, how to transform the variable) can still be made, but that the names of the involved variable(s) will not be clear for the researcher until the blind is lifted.

Finally, the studentized residuals versus fitted values plot of the blinded model (Figure 7.f) matches with that of the raw model (Figure 7.a), enabling us to conclude that it is still possible to check for constant error variance. This makes sense as both the residuals and the fitted values do not change due to the blinding.

4. Discussion

The high prevalence of p-hacking among psychological researchers is one of the causes of the abundance of false positive results in the literature. Because of this, measures against p-hacking seem desperately needed. Although pre-registration is a potential measure that raises in popularity nowadays, some researchers still shy away from this practice as it is quite challenging to register the entire analysis plan in advance, especially the first time. In physics, there exists another commonly used measure to counter p-hacking, which has barely been introduced in psychology to date: blind analysis. Only a few studies have clarified the importance of blind analyses in psychology and have discussed methods to blind data of an experimental research design (Dutilh, Sarafoglou, & Wagenmakers, 2019; MacCoun & Perlmutter, 2015; 2017). To the best of our knowledge, this master’s thesis is the first to propose methods to blind particularly data of a regression research design: (a) adding noise to the outcome scores, (b) scrambling the outcome scores, (c) adding bias to the coefficients, (d) creating new coefficients and, (e) scrambling the predictor labels. After illustrating the five blinding methods, we have evaluated them based on two criteria: Are the blinding methods able to counter p-hacking? And are the major assumptions of a linear regression still checkable even though the data is blinded?

4.1. Discussion regarding evaluation of p-hacking

In this master's thesis we assume that p-hacking is countered when a researcher is not able to infer (an estimate of) the p-values of the raw effects based on the blind effects he sees. Our evaluation showed that only the 'scrambling outcome scores'-method and 'creating new coefficients'-method met this requirement. This implies that the majority of the methods (i.e., adding noise to outcome scores, adding bias to coefficients and scrambling predictor labels) are unable to fully prevent p-hacking. For these methods there exists a certain dependency between (the p-values of) the blind effects and (the p-values of) the raw effects. However, it seems unfair to characterize the three blinding methods entirely as fruitless, since they can at least make p-hacking more difficult. Three reasons why the methods of adding noise to outcome scores, adding bias to coefficients and scrambling predictor labels can still counter p-hacking to some extent are listed below.

In order to p-hack, the researcher must be in first instance familiar with the fact that there exists a certain dependency between the raw effects and blinded effects. Given that a researcher has not read this master's thesis and has not tested the blinding methods himself, it can be hard to figure out this dependency; especially the fact that adding noise to the outcome scores and adding bias to the coefficients ensures that the strongest blind effects are most likely to have the strongest raw effects. When a researcher is not familiar with the existence of the dependency between the raw and blinded effects, this means that he has no clue about the p-values of the raw effects and p-hacking is still fully countered.

Furthermore, when a researcher is familiar with the existence of the dependency, it is still impossible for him to be 100% sure about what (the p-values of) the raw effects will look like exactly. By scrambling the outcome scores or by adding bias to the coefficients the researcher can only get a *suspicion* of the strength of the raw effects. Similarly, by scrambling the predictor labels the researcher can only get a *suspicion* of which effect belongs to which predictor variable. And depending on the topic and predictors at hand, it can be quite hard to make these suspicions about the raw effects based on the blind effects. To make it even more difficult and time-consuming, it might be an idea to sample multiple (e.g., six) blinded dataset, rather than one. Because with each sampled dataset the researcher is put on a (slightly) different track and therefore gets a (slightly) different suspicion about the raw effects. MacCoun and Perlmutter (2017) proposed this procedure, which Dutilh et al. (2019) call ‘decoy of data analysis’, only for the ‘scrambling predictor labels’-method. However, we think it can be an overarching method that has potential to make p-hacking extra hard and time-consuming for the methods of adding noise to outcome scores and adding bias to coefficients as well.

Third, when blinding data with the ‘adding noise to outcome scores’-method, ‘adding bias to coefficients’-method or ‘scrambling predictor labels’-method it takes some effort to p-hack the raw effects, especially when multiple blinded datasets are created. When a researcher makes all this effort just to manipulate the results, one would expect the researcher to be very aware of his p-hacking behaviour. P-hacking in such a conscious way seems equivalent to research fraud. The three methods may indeed not be able to prevent this *conscious,* fraudulent p-hacking behaviour for 100%. However, an important addition to this evaluation is that it seems that they are well able to prevent *unconscious,* innocent p-hacking behaviour. Since we expect (and mainly hope) that researchers p-hack most of the time unconsciously, this again indicates that the three methods seem moderately effective in countering p-hacking.

However, a nuance in this adjusted evaluation has to be made for the ‘scrambling predictor labels’-method: that this method appears to be moderately effective in countering p-hacking holds only when the researcher aims to achieve statistical significance for a *specific effect*. When he aims to achieve statistical significance for a *specific number of effects* (i.e., no matter which effects), the ‘scrambling predictor labels’-method is not able to counter p-hacking at all (also not when multiple datasets are created), as a researcher can still get all the ‘marginally’ significant effects below the alpha level. To summarize, while the ‘scrambling outcome scores’-method and the ‘creating new coefficients’-method succeed best in countering p-hacking, scrambling the predictor labels (with the aim to p-hack a specific effect), adding noise to the outcome scores and adding bias to the coefficients are blinding methods that should not be thrown immediately overboard either.

That all five blinding methods are thus—to some extent—capable of countering p-hacking is of course only applicable when researchers do not perform any additional analysis after the blind is lifted and the actual results are revealed. Sometimes, however, additional analyses are needed: think of a simple coding error that is only discovered after the blind is lifted. We therefore believe that such ‘post-(blind) analyses’ should be permitted, but only if any post-(blind) analysis is distinguished from the blind analyses in the write-up of the manuscript. A parallel can be drawn to the fact that it is allowed to report any post-hoc test separately from the main hypothesis tests and to report any non-registered analysis separately from the pre-registered analyses. This way, it will be clear to the reader in which analyses the researcher took care to protect himself against p-hacking and in which analyses he did not.

In addition to this, p-hacking is of course only countered if the blind analyses are conducted honestly. Claiming that you performed blind analyses although you peeked at the data is of course highly questionable. In theory, one might expect that the smaller the research team, the less likely the team members will object to any effort to cheat the blinding procedure (Faia, 2000). Logically, this could mean that the chance of performing a blind analysis dishonestly is generally higher in psychology, where research is often done by a single person or a very small team, compared to for example physics, where large interdisciplinary research teams are often required. To prevent cheating, Dutilh et al. (2019) proposed to perform the blinding according to an online protocol in which both parties (the data manager as well as the data analyst) have to sign a ‘contract’. By signing the contract, the data manager declares that he did not share the raw data with the data analyst and the data analyst declares that he did not have access to the raw data during the analyses. In principle, researchers can of course sign the contract while still performing questionable practices. However, the idea is that signing the contract increases at least extra awareness of sticking to the rules. In addition, when the researchers follow the blinding protocol, they receive a blinding certificate which they can include to their manuscript. Thanks to the blinding certificate, the readers, editors, reviewers, and colleagues can also trust that the blinding was conducted honestly.

4.2. Discussion regarding evaluation of checking assumptions

That all blinding methods are—to some extent—able to counter p-hacking when performed honestly, is however not sufficient to conclude that the methods act as good blinding techniques. The blinding also has to ensure that particular properties of the data, needed to check the major assumptions of a regression model, remain intact. Otherwise these assumptions can no longer be checked by the researcher and the analytic decisions based on these assumptions can no longer be made.

Our evaluation showed that the first (adding noise to outcome scores) and second (scrambling outcome scores) discussed blinding methods do not seem to be very attractive in terms of checking the assumptions. Both blinding methods cause particular properties of the data (especially the residuals) to change, making it impossible to check the majority of the assumptions. In fact, only multicollinearity and high-leverage can be checked with these blindings, as the predictor scores do not change. To account for the remaining assumptions while still countering p-hacking, it might be an idea to combine the blind analyses with a pre-registration of the remaining assumptions (i.e., every assumption except multicollinearity and high-leverage). The pre-registration then specifies the actions to be taken if it appears that the raw data does not meet one or more of these assumptions. However, this means that pre-registration is still needed and as stated before this can be a challenging practice.

When adding bias to the coefficients (blinding method 3) and creating new coefficients (blinding method 4) more properties of the data (especially the residuals) remain unchanged, making it possible to check, in addition to multicollinearity and high-leverage observations, also for outliers, influential observations, normality and linearity. All the analytic decisions based on these assumptions can therefore be made. However, as mentioned earlier, an additional remark has to be made for the linearity assumption-check: although we can, based on the visually check (i.e., the component-plus-residual plot), decide whether, and if so, which variable to transform, the visually check will not reveal how to transform the variable. In addition, constant error variance cannot be checked due to the changes in fitted values and therefore the analytic decisions based on this assumption cannot be made either. Again, to account for the constant error variance assumption and the complication of the linearity assumption while still countering p-hacking, the researcher might pre-register those parts of the analysis. Although the third and fourth blinding methods still have to be used in combination with pre-registration to counter p-hacking to the fullest, they seem more attractive than blinding method 1 and 2 when it comes to checking the assumptions, since less pre-registration is required.

While the two blinding methods are able to check the ‘residual’-based assumptions (i.e., outliers, unusable observations, normality and linearity), it is important to realize that a problem with checking these assumptions occurs when a researcher decides based on an assumption-violation to change the original model (e.g., to remove a variable from the model due to multicollinearity or to transform a variable in the model due to non-normality and/or non-linearity). The blinded dataset is created as such that fitting the original regression model to the blinded data will lead to the exact same residuals as fitting the original regression model to the raw data. Fitting a new model (i.e., with a variable removed or with a transformed variable) to this blinded data will therefore lead to different residuals than fitting this new model to the original/raw data. As a result, it will be not possible to check outliers, influential observations, normality and linearity anymore after removing a variable due to multicollinearity or after transforming a variable due to non-normality and/or non-linearity (note that it is therefore also not possible to check if a transformation has corrected non-normality and/or non-linearity).

To address this complication, we propose the following. It might be an idea to create a new blinded dataset each time a change in the linear model is proposed. This means that the data manager has to blind the data again according to the new regression model (and therefore the new residuals). This blinded dataset is passed back to the data analyst. The data analyst can then fit the new model to this dataset, which allows him to check the residual-based assumptions again. We are aware that the transfer of the data between data analyst and data manager makes this process not easy and ideal. However, it might be possible to implement algorithms in standard analysis software programs, which can ensure that the switch between data analyst and data manager is no longer necessary.

Lastly, while the ‘adding bias to the coefficients’-method and ‘creating new coefficients’-method seem quite promising in reaching the goal of checking the assumptions, our evaluation showed that the ‘scrambling predictor labels’-method succeeds even better when it comes to this goal. The data points do not change when the predictor labels are scrambled and therefore it is still possible to check and make analytic decisions for all the five major assumptions. Hence, no preregistration of how to handle a violated assumption is required at all when using this blinding method. In addition, when changing the original model due to a violated assumption (e.g., removing a variable from the model due to multicollinearity or transforming a variable in the model due to non-normality and/or non-linearity and/or non-constant error variance), all assumptions can still be checked.

4.3. Remaining advantages and disadvantages of blinding methods

Besides the evaluation of the blinding methods in terms of p-hacking and checking the major assumptions, we came across some remaining pros and cons of the blinding methods which are important to consider as well when deciding which blinding method is best in use. As discussed before, a major drawback of the fifth blinding method (i.e., scrambling predictor labels) is that this method can only work effectively when all variables of interest are measured on the same scale, which is rarely the case. Although this method seems thus a good and easy (i.e., just ask a colleague down the hall to scramble the labels of the predictor variables) way to counter p-hacking of specific effects while still checking all the main assumptions, it can unfortunately hardly be used in practice. The first four blinding methods do not encounter this problem: they work effectively when data is measured on the same as well as on different scales.

Second, while the first four blinding methods can be used when the researcher is planning to fit a simple linear regression model as well as a multiple linear regression model, it appears that the fifth blinding method (i.e., scrambling predictor labels) can only be used when the researcher is planning to fit a multiple linear regression model (i.e., a linear regression with more than one predictor variable). When the model consists of only one predictor variable, there simply are no predictor variables to scramble. This drawback is luckily not very disturbing, as multiple regression is more frequently used than simple regression (Kashy et al., 2009).

A third point to consider when deciding upon a blinding method is the following. In order to blind the data according to blinding method 3 (adding bias to coefficients), 4 (creating new coefficients) and 5 (scrambling predictor labels), the researcher first has to decide upon the linear regression model he is planning to fit to the data. This means that the researcher has to decide which predictor variables, covariates, interaction terms, polynomial terms et cetera he wants to include in the model, without being able to check whether this is appropriate in the light of the data. This advance-planning can be challenging for researchers and therefore they can shy away from these three blinding methods, like some researchers tend to shy away from pre-registration for the same reason. When a data analyst decides after seeing the blinded data that he wants to enter one or more extra predictor variables, covariates, interaction terms, polynomial terms et cetera in the regression model, the results of these additions can only be analysed blind when the data manager blinds the data again based on the ‘new’ model. The first two blinding methods (adding noise to outcome scores and scrambling outcome scores) encounter this problem to a much lesser extent: it is possible to blind the data using these two methods by only defining the outcome variable of the model in advance.

But the blinding methods have also some advantages. While this master’s thesis focuses on the use of blind analysis on its own, an additional advantage of blinding in general, and therefore applicable to all five blinding methods, is that blind analysis can also be used in combination with pre-registration. Remember that some people shy away from pre-registration as it is a major challenge to make all analytic decisions before having a chance to explore the dataset. When combining pre-registration with blind analysis, this challenge can fade away as the researcher can already explore (at least a part of) the data, while not having a chance to p-hack. To illustrate, during a blind analysis the researcher can observe an extreme skew of the data and can explore several transformations to fix this, while not seeing any of the main results. The best transformation can then be written down in the analysis plan, something that would have been difficult when no blind analysis was performed.

Another additional advantage of blinding is that it can really bring back the excitement in research. Imagine the moment when the blind is lifted and the raw results are revealed. This is surely much more exciting than today’s research where so much time is spent on iterating between analysis and outcome that the eventual results are not surprising at all.

4.4. Future research

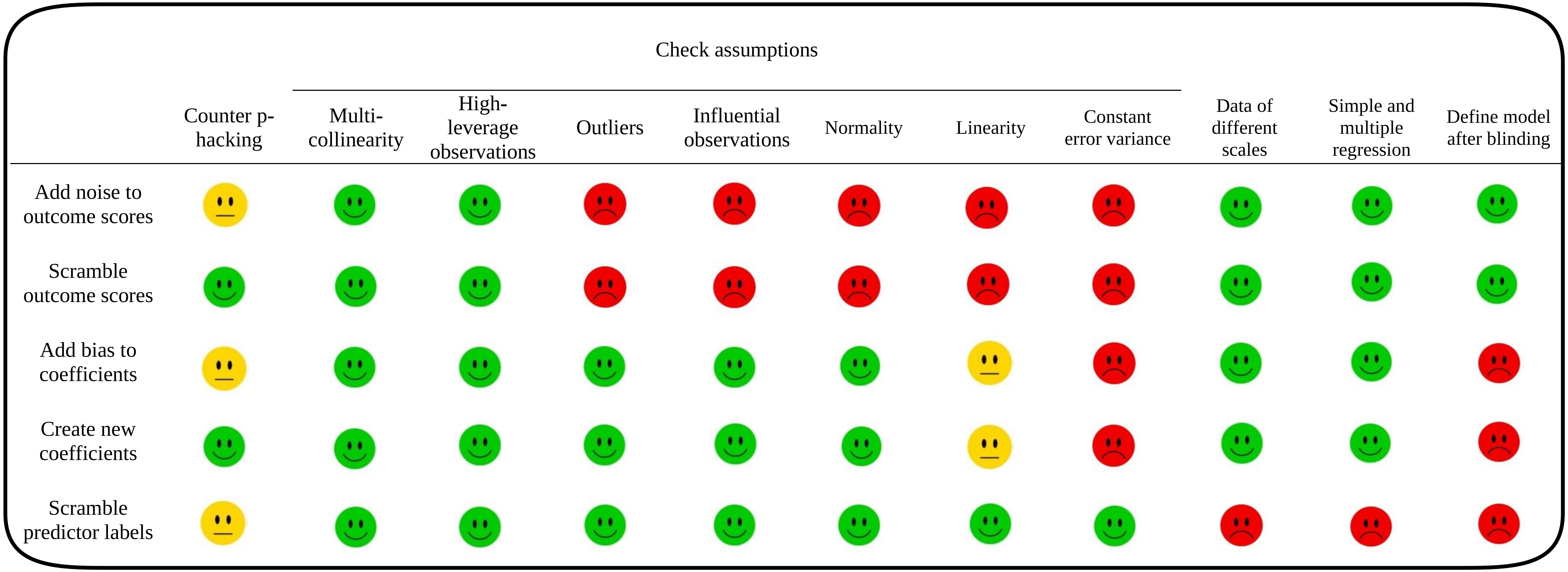
A summary of the advantages and disadvantages of the blinding methods is displayed in Figure 8. Although this master’s thesis suggests some promising techniques to blind data of a regression design, there is, as can be seen, not a method that satisfies all conditions. We therefore hope that future research will combine the discussed methods, try to make improvements on the discussed methods and explore and test additional methods of data blinding. This way, our list of blinding methods will hopefully be expanded with other (perhaps more) effective blinding techniques. To realize this, it might be an idea to look (even more) at what is being done in other research fields. The field of physics might be a good starting point, since blind analyses are originated from this field and physics is therefore at the forefront regarding research in this area (see for example Klein & Roodman, 2005). Also, since blind analyses are already widely used in physics, we can probably learn what methods (or what combination of methods) are most appropriate by, for example, a review of best practices (in addition to or instead of evaluating the methods based on mathematical analysis, Monte Carlo simulations and empirical testing).

In addition, it is important to keep in mind that blinding is not a one-size-fits all approach. As already explained, it is not self-evident that the blinding methods proposed by Dutilh et al., (2019) and MacCoun and Perlmutter (2015; 2017) to blind data of an experimental research design can also be (appropriately) used to blind data of a regression research design. In the same way, it is not self-evident that the blinding methods that are applied on our simulated dataset can also be (appropriately) used to blind data of a regression design with other characteristics. It might for example be that hierarchical or longitudinal 'regression' data, or 'regression' data with binary, categorical or count data as outcome variable (and therefore another link function) require other, tailor-made blinding methods. This again emphasizes that our list of blinding methods is definitely not exhaustive and really meant as a starting point. We hope therefore that future research will explore techniques tailored to the specific features of other research designs as well.

Lastly, although this research makes a good first step in facilitating the blinding of data, many more steps in this direction must and can be taken in the future. A potential next step might be to introduce algorithms in standard analysis software that make it possible to apply a certain blinding method with much ease to a dataset. The majority of the blinding methods ask more sophisticated practices than simply asking a colleague down the hall to scramble the labels of the predictor variables. The extra effort and difficulties that researchers can encounter when applying a blind can discourage them from using this practice, which is why ease of implementation is key.

4.5. Concluding remarks

To conclude, the current master's thesis can be seen as a starting point to familiarize researchers with a handful of blinding methods that can be applied to data of a regression research design, including their advantages and disadvantages. Once additional and tailor-made blinding methods are explored and tested and when it is facilitated for researchers to apply the blinding methods through off-the-shelf algorithms, hopefully more and more researcher will perform blind (regression) analyses in the future. Blind analysis is one of the few ways to really trust research results, as it prevents researchers from fooling themselves by p-hacking their results. And as Feynman (1985) puts it: "The first principle [of science] is that you must not fool yourself —and you are the easiest person to fool."



		Check assumptions									
	Counter p-hacking	Multi-collinearity	High-leverage observations	Outliers	Influential observations	Normality	Linearity	Constant error variance	Data of different scales	Simple and multiple regression	Define model after blinding
Add noise to outcome scores	😐	🙂	🙂	🙁	🙁	🙁	🙁	🙁	🙂	🙂	🙂
Scramble outcome scores	🙂	🙂	🙂	🙁	🙁	🙁	🙁	🙁	🙂	🙂	🙂
Add bias to coefficients	😐	🙂	🙂	🙂	🙂	🙂	😐	🙁	🙂	🙂	🙁
Create new coefficients	🙂	🙂	🙂	🙂	🙂	🙂	😐	🙁	🙂	🙂	🙁
Scramble predictor labels	😐	🙂	🙂	🙂	🙂	🙂	🙂	🙂	🙁	🙁	🙁

Figure 8. Overview of the pros and cons of the different blinding methods.

5. References

Agnoli, F., Wicherts, J. M., Veldkamp, C. L. S., Albiero, P., & Cubelli, R. (2017). Questionable research practices among Italian research psychologists. *PLoS ONE*, *12*(3), e0172792. https://doi.org/10.1371/journal.pone.0172792

Asendorpf, J. B., Conner, M., De Fruyt, F., De Houwer, J., Denissen, J. J. A., Fiedler, K., … Wicherts, J. M. (2013). Recommendations for increasing replicability in psychology. *European Journal of Personality*, *27*(2), 108–119. https://doi.org/10.1002/per.1919

Bakker, M., Van Dijk, A., & Wicherts, J. M. (2012). The rules of the game called psychological science. *Perspectives on Psychological Science*, *7*(6), 543–554. https://doi.org/10.1177/1745691612459060

Bishop, D. V. M., & Thompson, P. A. (2016). Problems in using p-curve analysis and text-mining to detect rate of p-hacking and evidential value. *PeerJ*, *4*, e1715. https://doi.org/10.7717/peerj.1715

Center for Open Science. (n.d.). *Registered Reports: Peer review before results are known to align scientific values and practices*. Retrieved from https://cos.io/rr/?_ga=2.146759412.1994345811.1567513924-1882659382.1565427361#journals

Champely, S. (2018). pwr: Basic functions for power analysis. R package version 1.2-2. https://CRAN.R-project.org/package=pwr

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). New York, NY: Lawrence Erlbaum Associates.

DeCoster, J., Sparks, E. A., Sparks, J. C., Sparks, G. G., & Sparks, C. W. (2015). Opportunistic biases: Their origins, effects, and an integraded solution. *American Psychologist*, *70*(6), 499–514. https://doi.org/10.1037/a0039191

Dutilh, G., Sarafoglou, A., & Wagenmakers, E.-J. (2019). Flexible yet fair: Blinding analyses in experimental psychology. *Synthese*. Advance online publication. https://doi.org/10.31234/osf.io/d79r8

Dutilh, G., Vandekerckhove, J., Ly, A., Matzke, D., Pedroni, A., Frey, R., … Wagenmakers, E. J. (2017). A test of the diffusion model explanation for the worst performance rule using preregistration and blinding. *Attention, Perception, and Psychophysics*, *79*(3), 713–725. https://doi.org/10.3758/s13414-017-1304-y

Faia, M. A. (2000). “Three can keep a secret if two are dead” (Lavigne, 1996): Weak ties as infiltration routes. *Quality and Quantity*, *34*(2), 193–216. https://doi.org/10.1023/A:1004785122594

Fanelli, D. (2010). Do pressures to publish increase scientists’ bias? An empirical support from US states data. *PLoS ONE*, *5*(4), e10271. https://doi.org/10.1371/journal.pone.0010271

Fanelli, D. (2012). Negative results are disappearing from most disciplines and countries. *Scientometrics*, *90*(3), 891–904. https://doi.org/10.1007/s11192-011-0494-7

Feynman, R. P. (1985). *Surely you’re joking, Mr. Feynman!* New York, NY: W. W. Norton.

Fiedler, K., & Schwarz, N. (2016). Questionable research practices revisited. *Social Psychological and Personality Science*, *7*(1), 45–52. https://doi.org/10.1177/1948550615612150

Fox, J. (2016). *Applied regression analysis and generalized linear models* (3th ed.). Los Angeles, CA: Sage.

Hartgerink, C. H. J., Van Aert, R. C. M., Nuijten, M. B., Wicherts, J. M., & Van Assen, M. A. L. M. (2016). Distributions of p-values smaller than .05 in psychology: What is going on? *PeerJ*, *4*, e1935. https://doi.org/10.7717/peerj.1935

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS Biology*, *13*(3), e1002106.

https://doi.org/10.1371/journal.pbio.1002106

Heinrich, J. G. (2003). Benefits of blind analysis techniques. *Unpublished manuscript.* Retrieved from https://www-cdf.fnal.gov/physics/statistics/notes/cdf6576_blind.pdf

Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine, 2*(8), 696–701. https://doi.org/10.1371/journal.pmed.0020124

John, L. K., Loewenstein, G., & Prelec, D. (2012). Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological Science*, *23*(5), 524–532. https://doi.org/10.1177/0956797611430953

Kashy, D. A., Donnellan, M. B., Ackerman, R. A., & Russell, D. W. (2009). Reporting and interpreting research in PSPB: Practices, principles, and pragmatics. *Personality and Social Psychology Bulletin*, *35*(9), 1131–1142. https://doi.org/10.1177/0146167208331253

Klein, J. R., & Roodman, A. (2005). Blind analysis in nuclear and particle physics. *Annual Review of Nuclear and Particle Science*, *55*, 141–163. https://doi.org/10.1146/annurev.nucl.55.090704.151521

LaRue, G. S., Phillips, J. D., & Fairbank, W. M. (1981). Observation of fractional charge of (1/3)e on matter. *Physical Review Letters*, *46*(15), 967–970. https://doi.org/10.1103/PhysRevLett.46.967

Leggett, N. C., Thomas, N. A., Loetscher, T., & Nicholls, M. E. R. (2013). The life of p: “Just significant” results are on the rise. *The Quarterly Journal of Experimental Psychology*, *66*(12), 2303–2309. https://doi.org/10.1080/17470218.2013.863371

Lyons, L. (2008). Open statistical issues in particle physics. *The Annals of Applied Statistics*, *2*(3), 887–915. https://doi.org/10.1214/08-AOAS163

MacCoun, R. J. (in press). P-hacking: A strategic analysis. *SSRN Electronic Journal*. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3433221

MacCoun, R. J., & Perlmutter, S. (2015). Hide results to seek the truth. *Nature, 526*(7572), 187–189. https://doi.org/10.1038/526187a

MacCoun, R. J., & Perlmutter, S. (2017). Blind analysis as a correction for confirmatory bias in physics and in psychology. In S. O. Lilienfeld & I. D. Waldman (Eds.), *Psychological science under scrutiny: Recent challenges and proposed solutions* (pp. 297–322). https://doi.org/10.1002/9781119095910.ch15

Makel, M. C., Plucker, J. A., & Hegarty, B. (2012). Replications in psychology research: How often do they really occur? *Perspectives on Psychological Science*, *7*(6), 537–542. https://doi.org/10.1177/1745691612460688

Marszalek, J. M., Barber, C., Kohlhart, J., & Holmes, C. B. (2011). Sample size in psychological research over the past 30 years. *Perceptual and Motor Skills*, *112*(2), 331–348. https://doi.org/10.2466/03.11.PMS.112.2.331-348

Masicampo, E. J., & Lalande, D. R. (2012). A peculiar prevalence of p values just below .05. *The Quarterly Journal of Experimental Psychology*, *65*(11), 2271–2279. https://doi.org/10.1080/17470218.2012.711335

Miller, L. E., & Stewart, M. E. (2011). The blind leading the blind: Use and misuse of blinding in randomized controlled trials. *Contemporary Clinical Trials*, *32*(2), 240–243. https://doi.org/10.1016/j.cct.2010.11.004

Moher, J., Lakshmanan, B. M., Egeth, H. E., & Ewen, J. B. (2014). Inhibition drives early feature-based attention. *Psychological Science*, *25*(2), 315–324. https://doi.org/10.1177/0956797613511257

Mosteller, F., & Tukey, J. W. (1977). *Data analysis and regression: A second course in statistics*. Reading, MA: Addison-Wesley.

Nosek, B. A., Beck, E. D., Campbell, L., Flake, J. K., Hardwicke, T. E., Mellor, D. T., … Vazire, S. (2019). Preregistration is hard, and worthwhile. *Trends in Cognitive Sciences*, *23*(10), 815–818. https://doi.org/10.1016/j.tics.2019.07.009

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science, 349*(6251), e4716. https://doi.org/10.1126/science.aac4716

Pallant, J. (2013). *SPSS survival manual: A step by step guide to data analysis using IBM SPSS* (5th ed.). Maidenhead, England: Open University Press. https://doi.org/10.1111/1753-6405.12166

Phillips, J. D., Fairbank, W. M., & Navarro, J. (1988). Recent results in the search for fractional charge at Stanford. *Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment*, *264*(1), 125–130. https://doi.org/10.1016/0168-9002(88)91113-8

Phillips, C. V. (2004). Publication bias in situ. *BMC Medical Research Methodology, 4*(20). https://doi.org/10.1186/1471-2288-4-20

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from https://R-project.org/

Silberzahn, R., Uhlmann, E. L., Martin, D. P., Anselmi, P., Aust, F., Awtrey, E., … 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., Nelson, L. D., & Simmons, J. P. (2014). P-curve: A key to the file-drawer. *Journal of Experimental Psychology: General*, *143*(2), 534–547. https://doi.org/10.1037/a0033242

van Dongen-Boomsma, M., Vollebregt, M. A., Slaats-Willemse, D., & Buitelaar, J. K. (2013). A randomized placebo-controlled trial of electroencephalographic (EEG) neurofeedback in children with attention-deficit/hyperactivity disorder. *Journal of Clinical Psychiatry*, *74*(8), 821–827. https://doi.org/10.4088/JCP.12m08321

Wagenmakers, E. J., Wetzels, R., Borsboom, D., Van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*(6), 632–638. https://doi.org/10.1177/1745691612463078

Ware, J. J., & Munafò, M. R. (2015). Significance chasing in research practice: Causes, consequences and possible solutions. *Addiction*, *110*(1), 4–8. https://doi.org/10.1111/add.12673

White, H. (2000). A reality check for data snooping. *Econometrica*, *68*(5), 1097–1126. https://doi.org/10.1111/1468-0262.00152

Wicherts, J. M., Veldkamp, C. L. S., Augusteijn, H. E. M., Bakker, M., van Aert, R. C. M., & van Assen, M. A. L. M. (2016). Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-hacking. *Frontiers in Psychology*, *7*, e1832. https://doi.org/10.3389/fpsyg.2016.01832