Observation oriented modeling revised from a statistical point of view

Sebastian Sauer¹

Published online: 25 August 2017
© Psychonomic Society, Inc. 2017

Abstract Observation Oriented Modeling was proposed to overcome some of the problems in the application of statistical inference methods in the behavioral sciences. In this paper, we refine one part of this approach and show how it is connected to methods that are well known in statistical learning. Specifically, we argue that the Moore-Penrose pseudo inverse is superior to the initial solution from a statistical point of view. With this we also show that Observation Oriented Modeling can indeed be appropriate for some tasks in the analysis of observed data. To provide a practical example, we demonstrate the revised method by analyzing the effect of mindfulness training on attentional processes.

Keywords Statistical inference · Statistical modeling · Observation oriented modeling · Mindfulness

A dominant procedure in applied statistical reasoning is null hypothesis significance testing (NHST). Although it is used frequently, voices of criticism have always been present and have gained prominence recently. One alternative is “Observation Oriented Modeling” (OOM), developed by Grice (2011). This method is suitable in some situations in the behavioral sciences as it circumvents some of the problems with existing inference or modeling techniques. The aim of the

present paper is to further refine OOM, and present a version that is based on well known classification methods. Specifically, we present a matrix algebra computation approach that is different from the one proposed by Grice, and we add an accuracy gain coefficient that is novel to OOM. The paper is structured as follows: First, we describe some problems with current inference practices, and go on to explain why OOM may be able to alleviate some of these problems. Next, we describe the original OOM and our revised version. The final part of the manuscript reports a practical application of OOM in some detail as a reanalysis of published data.

Criticisms on NHST

At least in applied research settings, the dominant paradigm for statistical inference is the technique of NHST, in research practice often understood as a “hybrid” version (Gigerenzer, 2004, p. 591) of the approaches by Fisher (1935) and Neyman and Pearson (1966). Despite its popularity, not only consumers of statistical methods but also statisticians have lamented the pitfalls of the concept (Cohen, 1990; Hubbard & Lindsay, 2008; Meehl, 1978). A thorough yet accessible criticism was brought forward by Lambdin (2012).

The criticism of NHST is twofold. One criticism is that, when applying NSHT it is easy to misinterpret the p -value. Another criticism is that the p -value is ill-suited for science, particularly because it does not address the right questions or is based on unrealistic assumptions (Halsey, Curran-Everett, Vowler, & Drummond, 2015; Wagenmakers, 2007). The American Statistical Association published a statement expressing the view that the p -value as a single number is often “misused”, but “can be a useful statistical measure” (Wasserstein and Lazar, 2016, p. 131).

Electronic supplementary material The online version of this article (<https://doi.org/10.3758/s13428-017-0949-8>) contains supplementary material, which is available to authorized users.

✉ Sebastian Sauer
sebastian.sauer@fom.de

¹ Iwp Institute, FOM University of Applied Sciences, Zeltnerstr. 19, 90443 Nürnberg, Germany

There might be some basis for the criticism that p -values are frequently misunderstood – at least for less statistically trained researchers. What might cause the confusion is the fact that the p -value is based on the tail probability of $p(D|H)$, with D , the data, summarized in an appropriate, i.e. sufficient test statistic and with H , the null hypothesis, and not, as it is often mistakenly understood, as $p(H|D)$. Gigerenzer (2004), for instance, presented empirical evidence that many research psychologists (including those teaching statistics) fail to interpret the p -value correctly.

The second criticism, that the p -value does not adequately address the more important questions of science, can be approached from several directions. Bayesians claim that scientists are interested in $p(H|D)$ instead of $p(D|H)$. Further, Bayesians claim that only Bayesian inference can estimate $p(H|D)$ by including “subjective” prior beliefs, as opposed to Frequentists who insist on considering actual data only for inference (Edwards, Lindman, & Savage, 1963).

Furthermore, there are also some doubts about whether the assumptions on which an inference is based (i.e., $p(D|H)$) hold in typical research scenarios. These assumptions often include the normal distribution or at least uni-modality, and, in bivariate cases, linear or monotone relations, respectively. Quite often it is unclear what the consequences of violations of the assumptions are. An additional problem needs to be considered in some research branches, including most of psychology: Are psychological (i.e., “latent”) attributes additive in structure, that is, are they quantitative (Stevens, Volkman, & Newman, 1937)? Some have expressed serious doubts about whether psychological attributes possess an additive structure (Michell, 1997, 2000).

Also it is well known that even when results achieve statistical significance, real effects are often small. As a result, overconfidence may occur. The p -value does not address relevance: It does not say how big a significant difference is – the problem of “sizeless science” (Ziliak and McCloskey, 2008).

Apart from Bayesian inference, the remedies that have been proposed to address the perceived problems with the p -value within the Frequentist framework include confidence intervals and effect sizes, to name a couple of concepts (Thompson, 1999; Wagenmakers, 2007). Nevertheless both Frequentist and Bayesian approaches may struggle if the aforementioned assumptions are not met.

Psychological theories are about organisms, not parameters

Many theories in the behavioral sciences are not about means or parameters but are about individual organisms. For example, the so-called by-stander effect in social psychology holds that if an individual is not alone, his or her

willingness to help in emergencies will decrease. In other words, victims are more likely to receive assistance when only a single individual witnesses the emergency (Latané & Nida, 1981). Although not uncontested, social inhibition in helping as a stable phenomenon has become common textbook knowledge (Aronson, Wilson, & Sommers, 2015). This theory states that an *individual* will decrease his or her willingness to help others in certain situations; the theory does not make any assertions about populations means. Thus, inferences about population means might not be suitable for such theories although frequently computed (Levine, Cassidy, Brazier, & Reicher, 2002). In order to evaluate a theory one should rather be interested in knowing for how many individuals the model hold than knowing how likely a mean is under the assumption of randomness.

Observation oriented modeling

Observation Oriented Modeling (OOM) includes a data analytic approach that is aimed at overcoming the aforementioned challenges (Grice, 2011; Grice, Barrett, Schlimgen, & Abramson, 2012). OOM is an exploratory analysis of model-data fit. The most prominent feature of OOM is that it focuses only on counts of observations (e.g. individuals) and not on parameters. Analytically, OOM is based on the cross-tabulation of data.

Epistemologically, OOM is founded on moderate realism, which goes to Aristotle and Thomas Aquinas (Grice, 2011). In contrast to other concepts of causality (e.g., Pearl, 2009), causality is analyzed by checking for the number of observations for which the effect conforms to the cause (i. e., counting and cross-tabulation). The observed pattern should therefore closely match the hypothesized pattern, similar to Ordinal Pattern Analysis (Thorngate & Ma, 2016). It should be noted that theory-driven modeling is important as missing factors can lead to biased results, such as in Simpson’s paradox (Simpson, 1951). In sum, OOM may present an approach to circumvent some of the problems attributed to common inference tools as discussed above for some scientific questions. It has been applied and extended in different fields of the behavioral and social sciences (Craig, Varnon, Sokolowski, Wells, & Abramson, 2014; Dinges et al., 2013; Grice, 2015; Grice, Cohn, Ramsey, & Chaney 2015a; Grice, Craig, & Abramson 2015b; Grice, Yopez, Wilson, & Shoda, 2016; Valentine & Buchanan, 2013).

Non-technical overview on OOM

In non-technical terms, OOM is based on matrix algebra approaches that can quantify the degree to which data fit a model by cross-tabulation. Basically, OOM provides a

descriptive and visual data summary plus a measure of association based on the cross-tabulation and classification of data. The generalization (i.e., inference to out-of-sample data) of results relies on replication only in OOM. One important method in OOM, primarily exploratory in nature, is called “Binary Procrustes Rotation” (Grice, 2011). A refinement of this method is the focus of the present work. In this context the goal of the proposed transformation is to gauge the similarity of two sets of data. These two sets are the observed behavior pattern y_i and the predicted behavior $\hat{y}_i$. As a data set can be seen as a collection of vectors (i.e., a matrix), we can project one of the data sets onto the other. To find the closest match between y_i and $\hat{y}_i$, one has to identify the optimal projection. The degree of proximity is taken as an indicator of model fit.

OOM converts all values to indicator (0/1) values. For example, the variable *gender* (with the two distinct values *female* and *male*) would be converted to two binary variables *female* and *male*. Thus, the proximity simplifies to the (relative) number of matches (i.e., the number of “conformed” observations). Grice (2011) referred to matrices of this binary indicator form as “Deep Structure” matrices.

When the measurement scale of the events is continuous or the event space is (very) large, some kind of discretization or binning may be needed (Dougherty, Kohavi, & Sahami, 1995). This transformation into a nominal scale can of course cause information loss – if the data are truly metric. But this preprocessing step can also enhance researchers’ understanding of their data (e.g., in terms of the relevance of the values in the model). Szepannek, Luebke, and Weihs (2005) employed a similar idea in a classification context; they spoke of “relevant regions” for pattern recognition.

By converting a variable to indicator variables the problem of whether a variable possesses metric level is circumvented. In addition, assumptions such as a monotonic or linear relationship, uni-modality, or the absence of outliers are no longer needed. In OOM, statistical reasoning is based on the observations, and their model fit. In contrast, probability distribution and parameters thereof are of no concern in OOM, thus avoiding overconfidence in the explanation or prediction of individual observations by significant parameters.

Of course, model fit plays an important role in OOM. For this purpose, Grice (2011) proposed a number of coefficients. The most important is the “Percent Correct Classification” (PCC) statistic. The PCC is the percentage of correctly assigned target observations (i.e., the relative number of observations where predicted values $\hat{y}_i$ equal observed values y_i) – just like the success rate for accuracy in machine learning (Labatut and Cherifi, 2012):

$$\text{PCC} = \frac{\sum_{i=1}^n I_{\hat{y}_i=y_i}}{n}. \quad (1)$$

Note that in contrast to p -values, the PCC is not a function of sample size n , but it gets more stable with higher sample sizes, e.g. applying resampling techniques or when seeing PCC as a realization of a random variable.

To estimate the probability of encountering a certain PCC by chance alone, permutation or cross-validation statistics can be computed. Grice (2011) reported a “chance value” c , which is the probability of achieving an equally high (or even higher) accuracy (PCC) calculated on the basis of a permutation of the data, comparable to randomization tests (Edgington and Onghena, 2007). In order to avoid overconfidence or over-fitting, cross-validation techniques for calculating the PCC can be applied (Efron & Gong, 1983), especially in the case of small samples or complex models.

Data analysis in the original OOM

In this section, we provide a more formal explanation of what Grice (2011) called “Binary Procrustes Rotation”, as outlined in the previous section. For n observations, let x_i be the value (the event in probability theory) of some effect for observation i , which can take one of L different values, so $x_i \in 1, 2, \dots, L$. The Deep Structure of x is then a $n \times L$ matrix X with

$$x_{ij} = \begin{cases} 1 & : \text{if } x_i = j \\ 0 & : \text{else.} \end{cases}$$

Note that within this framework the effect is endogenous (i.e., it is a dependent variable) but represented by x . Similarly, let y_i be the value of some hypothesized cause (exogenous, independent variable) for observation i . y_i can take one of K different values, so $y_i \in 1, 2, \dots, K$. The Deep Structure of y is then an $n \times K$ indicator matrix Y . In order to make Y “conform” to X , Grice (2011) employed a transformation matrix T such that

$$T = X'Y. \quad (2)$$

This $L \times K$ matrix is a cross-tabulation of effects and causes. Let P be the resulting matrix after the normalization of the transformation matrix T , where

$$P = UTV, \quad (3)$$

with $U \in \mathbb{R}^{L \times L}$ and $V \in \mathbb{R}^{K \times K}$ being diagonal matrices that are necessary for normalization. For both for U and V , the non-zero elements u_{ll} , and v_{kk} , respectively, refer to the inverse of the sum of the squared elements of the l th column and k th row, respectively, of T . Further, let

$$\hat{Y} = XP \quad (4)$$

be the rotated and normalized conforming matrix (i.e., the matrix that is to be compared with Y); Grice (2011) denoted $\hat{Y}$ as Y^C .

Model fit in OOM

Visually, a researcher might check how many observations (i.e., rows) of $\hat{Y}$ match the conforming matrix Y (see example below). More precisely, predicted values $\hat{y}_i$ can be compared with y_i , where $\hat{y}_i = \arg \max \hat{y}_{i1}, \hat{y}_{i2}, \dots, \hat{y}_{iK}$. By cross-tabulating x and the classification results of $\hat{y}$ (either correct, false, or ambiguous) one can eyeball the support for the model (i.e., visualize the number of observations for which the model holds, cf. Fig. 1). This diagram is called *multigram* in OOM (Grice, 2011).

A further statistic in OOM is the Classification Strength Index (CSI), proposed by Grice (2011). The CSI of observation i is designed as the highest value of row i in $\hat{Y}$ (i.e., $\max \hat{y}_i$). The mean CSI can therefore be used as an indicator of the extent to which the procedure is clearly able to assign the observations to an event (i.e., certainty within the model).

Revised version of OOM

We propose that the data analysis in OOM can be seen as a statistical classification approach that uses indicator matrices in a regression framework (Hastie, Tibshirani, & Friedman, 2009). Thus, one aims to minimize

$$\|Y - XP\|^2. \quad (5)$$

Our main contribution to OOM is that we propose that P (in Eq. 5) should be computed by using the Moore-Penrose pseudo inverse, X^+ (Ben-Israel and Greville, 2003). With some matrix calculations, it can be shown that in the case of indicator matrices the calculation of the Moore-Penrose pseudo inverse simplifies to

$$X^+ = \Lambda X^T. \quad (6)$$

Here, $\Lambda \in \mathbb{R}^{L \times L}$ is a diagonal matrix that contains the inverse of the event frequencies of x :

$$\lambda_{ii} = \begin{cases} \frac{1}{\sum_{j=1}^n x_{ji}} & : \text{if } \sum_{j=1}^n x_{ji} > 0 \\ 0 & : \text{else.} \end{cases} \quad (7)$$

Instead of using a one-dimensional vector x , our revised version of OOM uses an L -dimension indicator matrix X , and thus circumvents the problem of masking (Hastie et al., 2009), which can occur in the case of the regression of indicator matrices Y . Thus, in contrast to Grice (2011) method, we propose that Eq. 5 be optimized by

$$P = X^+Y = \Lambda X^T Y \quad (8)$$

in OOM. The results of Eq. 8 are identical to those obtained by using the “normalizing the conforming variable only” function in the OOM software (Grice, 2017). By using the

Moore-Penrose pseudo inverse in our adapted approach to OOM, Eq. 8 differs from the initial normalization procedure in Eq. 3. One advantage of the present approach is that the generalized pseudo inverse, such as the Moore-Penrose pseudo inverse, is the optimal solution to Eq. 5 (Ben-Israel and Greville, 2003). Moreover, P can now be seen as a classifier under the diagnostic paradigm (Dawid, 1976); note that in Eq. 2

$$t_{ij} = \sum_{k=1}^n x_{ik} y_{kj}, \quad (9)$$

(i.e., the number number of co-occurrences of $x = i$ and $y = j$, Grice (2011)).

This is just the cross-tabulation of data. Hence

$$p_{ij} = \begin{cases} \frac{\sum_{k=1}^n x_{ik} y_{kj}}{\sum_{k=1}^n x_{ki}} & : \text{if } \sum_{k=1}^n x_{ki} > 0 \\ 0 & : \text{else.} \end{cases} \quad (10)$$

is the conditional probability $P(y = j | x = i)$, so $\hat{Y} = XP$ contains the conditional probabilities of $y_k = j$ given $x_k = i$ for sample data (i.e., the conditional relative frequencies). So one can see that the revised version of OOM is a naive Bayes classifier, where the conditional probability is non-parametric and based on the empirical events (or binned versions thereof).

We propose that two issues regarding PCC calculation demand attention. First, should observations which fit two or more classes equally well (ambiguous observations) be counted as success or as error? Our PCC statistic reports the overall correct classification rate when ambiguous classification are treated as success. In contrast, our PCC_{amb} statistic counts ambiguous classification as not matching. Second, which class should be picked if two classes match equally well? In other words: How should ties be broken? Two possibilities are to 1) break ties randomly or to 2) match the observation to the most frequent class. Our algorithm allows for both tie breaking methods; in this paper, we demonstrate method 2.

Furthermore, we propose that the accuracy of a model should be judged against some baseline criterion. A useful baseline criterion is a model that always predicts the category for which the relative frequency of the predicted event is highest. The resulting coefficient may be called the (*accuracy gain*), which is comparable to performance measures in machine learning. Thus, the gain of the model is calculated as

$$\text{Gain} = \frac{PCC}{\max P(y = j)}, \quad (11)$$

Thus, the gain is defined as the improvement in the classification of y by including x in the model.

Of course it should be noted that finer data-analytic techniques exist for the analysis of cross-tabulations of events, such as association measures or independence tests (cf.

Fig. 1 Example of a multigram. A multigram is a visual representation of the classification accuracy. For each of two experimental groups (values of the target variable in OOM parlance), the number of correctly, falsely, or ambiguously classified observations are shown in the form of a bargraph-like plot

Agresti, 2003; Stemmler, 2014). The same is true for classifying $y \in \{1, 2, \dots, K\}$ on the basis of x (i.e., supervised classification methods, cf. Hastie et al. (2009)). However, the different epistemological approach of OOM should be remembered. In OOM, the focus of interest lies in the empirical pattern of the frequency of observations. OOM considers the question of how many observations fit, or, as Gosset put it to Pearson (1939, p. 247)

(...) obviously the important thing in such is to have a low real error, not to have a “significant” result (...).

In sum, our refined OOM approach is optimal in terms of the regression of indicator matrices and can be seen as a naive Bayes classifier, methods that are well known in statistical learning (Hastie et al., 2009). Unless stated otherwise, we have used this revised version of OOM throughout the paper.

OOM example in mindfulness research

In the first part of the article, we presented the extended and revised version of OOM, and explained its use. In the remainder, we present an empirical example to demonstrate an OOM analysis. To this end, we reanalyzed a published study (Sauer et al., 2012). The study tested whether mindfulness training would be found to be associated with an extension of the subjective now. Participants engaging in

regular mindfulness training were expected to indicate a longer subjective “now”, according to the model. The initial analysis was based on testing for statistically significant differences and found some evidence for an effect of mindfulness on the extension of the subjective now. However, our OOM analysis presents a more nuanced picture with less evidence for the effect in question.

Mindfulness and time perception

Mindfulness has been defined as focusing attention on present experiences, both internal (e.g., thoughts), and external (e.g., visual impressions) (Kabat-Zinn, 2003). This attentional focus is accompanied by a stance of nonjudgmental awareness with regard to a stream of consciousness perception (Sauer et al., 2013). Mindfulness can be conceptualized as an inherent human capacity, and it can be trained (Sauer et al., 2013). In a similar vein, “informal” mindfulness has been shown to reduce cognitive bias in a randomized trial (Williams, Teasdale, Segal, & Soulsby, 2000). Of course, during “formal” mindfulness practice (e.g. meditation), the effects of mindfulness should be stronger. Some interesting data that support this claim have been published. For example, Carter et al. (2005) found that during concentrative meditation, ocular rivalry was reduced in mindfulness experts.

Research on mindfulness has received a surge in interest in recent years, with a rising number of publications. A growing body of research suggests its effectiveness in alleviating a wide number of health conditions including stress, anxiety, depression, pain, high blood pressure, skin diseases, eating disorder, and substance abuse, to name a few (Chiesa & Serretti, 2014; Godfrey, Gallo, & Afari, 2014; Gu, Strauss, Bond, & Cavanagh, 2015; Hofmann, Sawyer, Witt, & Oh, 2010; Keng, Smoski, & Robins, 2011; Piet & Hougaard, 2011). It is interesting that, from a neurophysiological point of view, it appears that not only does mindfulness affect attentional resources (Slagter et al., 2007), but it has also been shown to be related with altered brain structures. For example, Gard, Hölzel, and Lazar (2014) investigated whether mindfulness could lessen the impact of age-related neurological decay and found some evidence to support this. Further, a study by Hölzel et al. (2011) suggest that mindfulness training may build up grey matter in areas that are associated with attention and emotion regulation.

Another strand of research links mindfulness to time perception. As the practice of mindfulness consists of staying as long as possible in the present moment, it appears plausible to hypothesize that mindfulness training may also prolong the “subjective now” (Wassenhove, Wittmann, Craig, & Paulus, 2011). Investigating human time perception is crucial for fostering the understanding of perceptual and cognitive processes. It should be stressed that individual temporal perception is not isomorphic to physical time duration (Wassenhove et al., 2011). From a practical point of view, time perception may be of interest, as the temporal borders of time perception limit the human capacity to perceive “zeitgestalts”. Thus, extending the borders of temporal perception may have implications for working memory or perhaps enable the perception of previously imperceptible precepts. On this basis, Sauer et al. (2012) argued that regular mindfulness training should be associated with a longer subjective now (as operationalized below) compared with a control group of mindfulness-naïve laypersons.

Method

Participants

The sample consisted of 38 mindfulness meditation experts (21 female, 17 male), and an age-matched group of 38, recruited in the UK, Germany, and Switzerland. Meditators were trained in Buddhist meditation techniques. The meditation group was recruited from two sources, first, a Tibetan Buddhist meditation center in which all meditators practiced various meditation exercises including concentration and compassion meditation (“Tibet group”, $n = 16$), and second, a Zen meditation group focusing on concentration meditation (“Zen group”, $n = 22$). The mean age was 51 years in each group. Inclusion criteria consisted of at least 5 years of daily meditation practice for the meditation group, and no meditation experience for the control group. Thus, high selection standards were applied.

Materials and procedure

Among other measures (see the original publication (Sauer et al., 2012) for details), individuals were instructed to watch a Necker Cube, a basic form of a bistable image (cf. Fig. 2). The most prominent feature of the Necker cube is that perception changes after some amount of time, irrespective to resistance of the individual. After some seconds, the cube “falls” from one visual perspective, say B, to the other (i.e., C).

In block 1, each individual was instructed to watch the computerized image of the Necker cube, and indicate any change in perspective upon occurrence by pressing a button on the keyboard. In block 2, individuals were instructed to hold the perspective for as long as possible. Duration of each block was 3 min (see the additional materials, S2, Fig. 3 for an illustration).

Hypotheses and data analysis

On the basis of the theoretical account on mindfulness and time perception presented above, we posit the following

Fig. 2 The Necker Cube: A simple example of a bistable image. The Necker cube (A), as presented to the participants. Due to its simplicity, it is a paradigmatic example of an bistable image, i.e., an unchanging visual stimulus that is perceived by humans in two (or more) different

ways. The Necker Cube is perceived as either the upper right corner in the front (B) or with its lower left corner in front (C). The perception switches from one perception to the other after some time. This change cannot be voluntarily suspended

Fig. 3 Violin plots and frequency plots (rug plots) comparing the number of reversals per group. In block 1 (*left panel*), participants did not receive any instructions to hold their perception of the Necker cube constant. In block 2 (*right panel*), participants were instructed to hold their perception constant for as long as possible. Individuals with a smaller number of reversals occurred most often in the Zen

group, followed by the control group, and then by the Tibet group. This was true for both Blocks, whereas the differences between the groups were more pronounced in block 1. The overall number of reversals decreased from block 1 to 2. Red triangles indicate medians. In the Zen group, the median number of reversals was lower than in the other two groups. This was true in both Blocks

hypotheses (H). Mindfulness experts will report fewer reversals in perceptions of the Necker cube compared with the laypersons in block 1 (H1), and in block 2 (H2). Reversals refer here to a change in the person's perception of the Necker cube (from Position B to C or vice versa in Fig. 2). H3 expects a drop in the number of reversals from block 1 to block 2. H4 states that the drop of reversals is greater for the meditators than for the lay persons. Specifically, H4 expects that the drop is more pronounced for the members of the Zen than for the members of the Tibet group. We expected this order because individuals in the Zen groups focused their meditative practice on sustained attention, whereas individuals in the Tibet group exercised a variety of meditative practices. Put shortly, our hypotheses on the number of reversals (NR) were as follows; note that these hypotheses are observation based:

H1: $NR_{control} > NR_{Tibet} > NR_{Zen}$ in block 1.

H2: $NR_{control} > NR_{Tibet} > NR_{Zen}$ in block 2.

H3: $NR_{block1} > NR_{block2}$.

H4: $NR_{control} < NR_{Tibet} < NR_{Zen}$ for the drop of reversals between block 1 and block 2.

One observation had to be excluded (Zen group) due to a technical problem resulting in missing values. For

permutation tests, 10,000 trials were computed to achieve numeric stability for reproducibility. Outliers were left untreated in the sample.

In terms of OOM statistics, we report two variants of the PCC statistic: first, with ambiguous observations assigned to the condition with the highest frequency (PCC), and second, with ambiguous observations assigned to a class called “ambiguous”, thus lowering the PCC_{amb} . In addition, we report the accuracy gain, the c-value, and the CSI, as described above. Results are given for our revised approach as well as for Grice (2011) original approach.

We conducted all analyses in R (3.2.3; R Core Team, 2015) including several packages (Auguie, 2015; Aust and Barth, 2016; Dahl, 2014; Fox & Weisberg, 2015; R Core Team, 2015; Wickham, 2015; Wickham & Chang, 2015; Wickham & Francois, 2015). More details on computer environment, including the names and versions of the R packages used, are provided in S1 (CompEnv). For ensuring the reproducibility of the computations, we conducted all analyses based on a CRAN image of 2016-01-18, using the package “checkpoint” (Analytics, 2015). The R Code for the analyses can be freely accessed here: osf.io/u3nq9 (Sauer & Luebke, 2017b). The data set can be freely accessed here: osf.io/sgzzyd (Sauer & Luebke, 2017a). We confirm that we did not alter the data in any other way than as described here. Supplemental files can be freely accessed here: <https://osf.io/6vhja/>. This paper was not preregistered.

Results

Descriptive results

In sum, the median number of reversals was 44.0 in block 1, and 25.0 in block 2 (a difference of 19.0). Given a block duration of 180s, the average reversal time was 4.1s in block 1, and 7.2s in block 2. Descriptive statistics (Md, Min, Max) for the whole sample as well as for the three subgroups are given in Table 1.

As can be seen in Table 1, the median number of reversals differed between groups. The observations in the Tibet group and the control group tended to show more reversals than the Zen group; this was true in both blocks. Thus, on an aggregate level (Md), the Zen group appeared to demonstrate a clear advantage in comparison over the control group and the Tibet group, consistent with what we expected. Contrary to our expectations, the Tibet group showed more reversals, on a median level than the control group (cf. Fig. 3). In sum, this results provided partial support for H1 and H2.

Apart from comparing the three groups, we were interested in comparing the differences in change from block 1 to block 2 between the three groups. As can be seen in Table 1, the median reduction in reversals was most pronounced in the Tibet group (24.5, 45%). The relative reduction was stronger in the Zen group than in the control group (36% vs. 30%). However, the absolute reduction was stronger in the control group (17.5) than in the Zen group (16.0). Thus, the absolute reduction rates showed a mixed picture, and provide partial support for H4.

In each group, most of the individuals reduced their number of reversals. As expected, this effect was more pronounced in the Tibet group ($15/16 \approx .94$) than in the control group ($34/38 \approx .90$). However, contrary to expectation, the Zen group ($18/21 \approx .86$) showed a smaller proportion of participants who lowered their reversal number; the difference between the groups was small. Figure 4 depicts this pattern. On a side note, the reduction was apparent for extreme cases as well thus supporting the argument for not excluding extreme cases from this dataset.

OOM results

For the OOM analyses, we binned the 43 different values of the conforming variable (numbers of reversals) in order to reduce sparse bins and for easier comprehension (see supplementary information files S3 for problems with unbinned data and file S4 for OOM results without binning). Binning is a standard procedure, for example in exploratory data analyses for histograms. However, there is no single rule for the correct numbers of bins. The most widely used choices

Table 1 The median, minimum and maximum are shown for each subgroup as well as the whole sample, as well as by block (1 vs. 2) and change between blocks (raw and relative change)

block	Group	Md	Min	Max
1	Zen	44.00	0.00	101.00
2	Zen	25.00	2.00	86.00
Raw change	Zen	−16.00	−74.00	33.00
Relative change	Zen	0.37	−Inf	0.94
1	control	36.00	6.00	98.00
2	control	20.00	3.00	41.00
raw change	control	−14.00	−74.00	33.00
relative change	control	0.36	−4.12	0.94
1	Tibet	52.00	0.00	101.00
2	Tibet	27.50	5.00	86.00
raw change	Tibet	−18.50	−74.00	1.00
relative change	Tibet	0.45	−0.07	0.79
1	All	43.50	0.00	78.00
2	All	26.00	2.00	67.00
raw change	All	−14.50	−45.00	30.00
relative change	All	0.30	−Inf	0.93

Note. One individual reported no changes in block 1, and 30 in block 2; this increase is indicated as *Inf* in the table

include Sturges' rule (Sturges, 1926), the Freedman-Doane rule (Freedman & Diaconis, 1981), or the square root rule (Venables & Ripley, 2013), among others; see supplementary information S2 for details. For the present data, these rules resulted in 8, 14, and 9 bins in block 1, respectively. As Sturges' rule the Freedman-Doane rule are acknowledged as more adequate, we decided to conduct the OOM analyses with both 8 and 14 bins and to compare results. It should be noted that we do not exclude the possibility that different binning procedures are defensible. As a reference binning, we included a median-split binning.

In the results for both block 1 and 2, as well as for the change between the two Blocks, we found that the results of the different binning strategies differed very little. Thus, here, we present only the results of the Freedman binning (PCCs: block 1: .59; block 2: .52; change: .55). A more detailed analyses of the results can be found in the supplementary information (S2). As can be seen in Fig. 5, the PCC was moderate at best. This was true for block 1 (H1), block 2 (H2), and for the change in the number of reversals from block 1 to 2 (H4).

Thus, the OOM statistics provided only a little support for H1, H2, and H4.

A summary of the results of the OOM analysis is given in Table 2; results are based either on the Moore-Penrose pseudo inverse or on Grice' method). Comparing the tables, it is obvious that the two normalization procedures can differ to a moderate extent. H3 was supported by the OOM

Fig. 4 Change in number of reversals from block 1 to block 2 per group. In block 1, participants did not receive any instructions to hold their perception of the Necker cube. In block 2, participants were instructed to hold their perception constant for as long as possible. Black dots represent observations; red lines represent the median

number of reversals per group. In each group, most of the individuals reduced their number of reversals. (Note one extreme case in the Zen group with a contrasting pattern which called for a post-hoc investigation)

Fig. 5 Multigram for the three experimental groups, using Freedman's binning rule. Upper panel: OOM results for the number of reversals for block 1. Middle panel: OOM results for the number of reversals for block 2. Lower panel: OOM results for the number of reversals for the change in the number of reversals from block 1 to block 2. In either case, PCC was on par with the chance level. Accordingly, the chance values (*c*) corroborated that the empirical PCC occurred frequently in the permuted trials. Note that the classification results (correct/false/ambiguous) are labeled *c*, *f*, or *a*, respectively, in the facet headers of the plot

Table 2 Summary of the OOM analyses

hypothesis	method	PCC	PCCamb	CSImean	gain	c_value
H1	revised	0.59	0.56	0.59	1.16	0.32
H2	revised	0.52	0.48	0.52	1.03	0.95
H3	revised	0.67	0.67	0.67	1.35	0.01
H4	revised	0.55	0.51	0.55	1.08	0.65
H1	Grice	0.59	0.59	0.81	1.16	0.09
H2	Grice	0.40	0.40	0.69	0.79	0.89
H3	Grice	0.67	0.67	0.86	1.33	0.01
H4	Grice	0.49	0.49	0.74	0.97	0.52

Notes. Method *revised* indicates that the revised version of OOM based on the Moore Penrose Inverse was used; method *Grice* indicated that the method developed by Grice (2011) was employed. PCC: Percent correct classification. *PCCamb*: PCC with ambiguous observations counted as error. *CSImean*: Mean CSI across all observations

results, as can be seen in Fig. 6: The number of reversal times conformed well to block 1; data from the three groups were combined in this analyses. Although the PCC of .67 appears moderate, the c-value of .001 indicates that the classification was nevertheless well above the rate of chance. Note that in block 2, the number of correct classifications (green color) was much higher than in block 1. The reason is that in block 1 the variation of the reversals was much more spread out than in block 2.

Discussion

Discussion of OOM example

The initial analysis by Sauer et al. (2012) presented results that favored the main hypothesis, (i.e., that mindfulness can prolong the subjective now). However, our reanalysis reveals a more nuanced and less favorable picture. In sum, H1, H2, and H4 were moderately supported by the descriptive statistics (cf. Figs. 3 and 4), but they were not well supported by the OOM statistics (cf. Fig. 5, Table 2). H3 was well supported by the OOM statistics.

Why were the OOM results less favorable than the descriptive results? One explanation is that the differences were not pronounced enough for the OOM algorithm to separate the groups. This explanation is backed up by the marked overlap of the three distributions. Thus, although the descriptive analyses uncovered some difference between the groups, this difference should be regarded as small in size – especially on the level of the observations. We did not exploit the numeric (count) level of the data, consistent with the theoretical underpinnings of OOM.

In sum, the OOM analysis provided only weak evidence that some forms of attentive training can prolong the reversal times in bistable images. If it exists, this effect would be of interest as cognitive comprehension hinges on the size of working memory. If working memory capacity could be

Fig. 6 Multigram for comparing Blocks 1 and 2. Data from all three groups were merged for this analysis. Note that the classification results (correct/false/ambiguous) are labeled by *c*, *f*, or *a*, respectively, in the facet headers of the plot. The number of reversals could be conformed well to the Blocks

enhanced by mindfulness training, practical and theoretical benefits can be envisioned.

In the original paper, the results were based on a test of statistically significant differences in reversals between meditators and non-meditators (statistical significance was found in block 2 but not in block 1 in the original paper). The present reanalyses shed much more light on the particularities of the data. For example, the effects of mindfulness appeared to be moderated by the type of training, (i.e., Zen and Tibet data exhibited quite different results).

Future studies should provide more data and should try to find more reliable measures of when perception changes. Of course, limitations of the present study include its quasi-experimental setup. A true experimental setup is necessary to shed light on the effects and mechanisms of mindfulness with regard to attentive resources and processes. Of interest, previous research noted that the subjective now typically has a time window of around 3 seconds (Cohen, 1959). In our data, mean reversal times were often longer. However, it has been noted that whereas the distribution of reversal time may peak at 3 seconds, considerable variance prevails (Atmanspacher, Filk, & Römer, 2004). Thus, investigating the distribution of the raw reversal time would be of interest.

Overall discussion

The main objective of this paper was to present a refined version of OOM as an exploratory data modeling technique that may be suited to some of the data and research questions in the behavioral sciences, as its data analytic part is non-parametric and based on well-known methods such as cross-tabulation. Its most prominent features are the focus on counting individual observations and modeling based on sample data, but still offering possibilities for model checking by reporting the number of observations for which the model holds by presenting classification accuracy. We believe that OOM can offer an extension of the more common statistical armory in behavioral science and may help provide insights with respect to some scientific questions. One important aspect of OOM, which was not the focus of the present paper, involves the use of iconic or “integrated models” (Grice, 2011) (cf. Fig. 3 in S2). Such visual modeling forces the analyst to make assumptions that are more explicit than what is presented in standard path diagrams.

The initial OOM method by Grice (2011) is based on matrix algebra. This is an advantage for applied scientists with limited time and interest in statistical subtleties. However, such subtleties should be addressed when the assumptions of (standard) statistical methods are not met. Here, we adopted an approach that was based on the Moore-Penrose pseudo inverse. The advantage of this method is that it is known to optimize error as detailed above. The results of the two methods - Grice vs. Moore-Penrose - can differ.

We therefore advise analysts to report their normalization method.

In OOM, logical or arithmetic combinations of variables can be modeled – a procedure dubbed “logical hypothesis testing” by Grice (2011). This method allows researchers to potentially disentangle the effects of different causes. For monotonic relationships, for example, cross-tabulating, as is performed in OOM, may be too flexible. Also it may require data pre-processing (i.e., binning). Similarly, in situations with known stochastic models and distributions, Frequentist or Bayesian approaches may be more appropriate than OOM.

For the time being, we conclude that OOM is a promising modeling approach for some research settings as it circumvents some of the problems encountered in the application and interpretation of more common methods. There is a need for more research experience with this method and certainly room for more statistical revisions or extensions.

Acknowledgments Karsten Luebke developed the mathematical framework of this paper. He also wrote substantial parts of this paper. Without his efforts, this work would not have appeared. Unfortunately, limited time resources did not allow him to finalize the work on this paper; so he decided that he would not appear as author. I am deeply thankful for his contribution. I thankfully acknowledge Jana Lemke who collected the data. Christian Roeber and Jane Zagorski provided helpful comments on an earlier draft of the manuscript. I also thank the editor and the reviewers for their constructive comments.

References

- Agresti, A. (2003). *Categorical data analysis*, 2nd Edn. Hoboken, NJ: John Wiley & Sons, Inc.
- Analytics, R. (2015). Checkpoint: Install packages from snapshots on the checkpoint server for reproducibility. Retrieved from <https://CRAN.R-project.org/package=checkpoint>.
- Aronson, E., Wilson, T. D., & Sommers, S. R. (2015). *Social psychology*. New York City: Pearson.
- Atmanspacher, H., Filk, T., & Römer, H. (2004). Quantum Zeno features of bistable perception. *Biological Cybernetics*, 90(1), 33–40. <https://doi.org/10.1007/s00422-003-0436-4>
- Auguie, B. (2015). GridExtra: Miscellaneous functions for grid graphics. Retrieved from <https://CRAN.R-project.org/package=gridExtra>.
- Aust, F., & Barth, M. (2016). Papaja: Create apa manuscripts with rmarkdown. Retrieved from <https://github.com/crsh/papaja>.
- Ben-Israel, A., & Greville, T. N. (2003). *Generalized inverses: Theory and applications (Second.)* New York: Springer.
- Carter, O. L., Presti, D. E., Callistemon, C., Ungerer, Y., Liu, G. B., & Pettigrew, J. D. (2005). Meditation alters perceptual rivalry in Tibetan Buddhist monks. *Current Biology*, 15(11), R412–R413.
- Chiesa, A., & Serretti, A. (2014). Are mindfulness-based interventions effective for substance use disorders? a systematic review of the evidence. *Substance Use & Misuse*, 49(5), 492–512. <https://doi.org/10.3109/10826084.2013.770027>
- Cohen, J. (1990). Things I have learned (so far). *American Psychologist*, 45(12), 1304–1312. Retrieved from <http://psycnet.apa.org/psycinfo/1991-11596-001>.

- Cohen, L. (1959). Rate of apparent change of a necker cube as a function of prior stimulation. *The American Journal of Psychology*, 72(3), 327–344. <https://doi.org/10.2307/1420037>
- Craig, D. P. A., Varnon, C. A., Sokolowski, M. B. C., Wells, H., & Abramson, C. I. (2014). An assessment of fixed interval timing in free-flying honey bees: An analysis of individual performance. *PLoS ONE*, 9, e101262. Retrieved from <https://doi.org/10.1371/journal.pone.0101262>.
- Dahl, D. B. (2014). Xtable: Export tables to latex or html. Retrieved from <https://CRAN.R-project.org/package=xtable>.
- Dawid, A. P. (1976). Properties of diagnostic data distributions. *Biometrics*, 32(3), 647–658.
- Dinges, C. W., Avalos, A., Abramson, C. I., Craig, D. P. A., Austin, Z. M., Varnon, C. A., & Wells, H. (2013). Aversive conditioning in honey bees (*apis mellifera anatolica*): A comparison of drones and workers. *Journal of Experimental Biology*, 216(21), 4124–4134. <https://doi.org/10.1242/jeb.090100>
- Dougherty, J., Kohavi, R., & Sahami, M. (1995). Supervised and unsupervised discretization of continuous features. In *Machine learning: Proceedings of the twelfth international conference*, (Vol. 12 pp. 194–202).
- Edgington, E., & Onghena, P. (2007). *Randomization tests*. Boca Raton, FL: CRC Press.
- Edwards, W., Lindman, H., & Savage, L. J. (1963). Bayesian statistical inference for psychological research. *Psychological Review*, 70(3), 193–242. <https://doi.org/10.1037/h0044139>
- Efron, B., & Gong, G. (1983). A leisurely look at the bootstrap, the jackknife, and cross-validation. *The American Statistician*, 37(1), 36–48.
- Fisher, R. A. (1935). *The design of experiments*. Oliver & Boyd: Edinburgh.
- Fox, J., & Weisberg, S. (2015). Car: Companion to applied regression. Retrieved from <https://CRAN.R-project.org/package=car>.
- Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift Für Wahrscheinlichkeitstheorie Und Verwandte Gebiete*, 57(4), 453–476. <https://doi.org/10.1007/BF01025868>
- Gard, T., Hölzel, B., & Lazar, S. W. (2014). The potential effects of meditation on age-related cognitive decline: a systematic review. *Annals of the New York Academy of Sciences*, 1307, 89–103. <https://doi.org/10.1111/nyas.12348>
- Gigerenzer, G. (2004). Mindless statistics. *The Journal of Socio-Economics*, 33(5), 587–606. <https://doi.org/10.1016/j.socrec.2004.09.033>
- Godfrey, K. M., Gallo, L. C., & Afari, N. (2014). Mindfulness-based interventions for binge eating: a systematic review and meta-analysis. *Journal of Behavioral Medicine*, 38(2), 348–362. <https://doi.org/10.1007/s10865-014-9610-5>
- Grice, J. W. (2011). *Observation oriented modeling: Analysis of cause in the behavioral sciences*. Cambridge, MA: Academic Press.
- Grice, J. W. (2015). From means and variances to persons and patterns. *Frontiers in Psychology*, 6. <https://doi.org/10.3389/fpsyg.2015.01007>.
- Grice, J. W. (2017). *Observation Oriented Modeling (OOM) Software*. Amsterdam, NL: Elsevier. Retrieved from <http://www.idiogrid.com/OOM/>.
- Grice, J. W., Barrett, P. T., Schlimgen, L. A., & Abramson, C. I. (2012). Toward a brighter future for psychology as an observation oriented science. *Behavioral Sciences*, 2(1), 1–22.
- Grice, J. W., Cohn, A., Ramsey, R. R., & Chaney, J. M. (2015a). On muddled reasoning and mediation modeling. *Basic and Applied Social Psychology*, 37(4), 214–225. <https://doi.org/10.1080/01973533.2015.1049350>
- Grice, J. W., Craig, D. P. A., & Abramson, C. I. (2015b). A simple and transparent alternative to repeated measures anova. *SAGE Open*, 5(3). <https://doi.org/10.1177/2158244015604192>.
- Grice, J. W., Yepez, M., Wilson, N. L., & Shoda, Y. (2016). Observation-oriented modeling going beyond: is it all a matter of chance?. *Educational and Psychological Measurement*, 0013164416667985.
- Gu, J., Strauss, C., Bond, R., & Cavanagh, K. (2015). How do Mindfulness-Based Cognitive Therapy and Mindfulness-Based Stress Reduction Improve Mental Health and Wellbeing? A Systematic Review and Meta-Analysis of Mediation Studies. *Clinical Psychology Review*, 37, 1–12. <https://doi.org/10.1016/j.cpr.2015.01.006>
- Halsey, L. G., Curran-Everett, D., Vowler, S. L., & Drummond, G. B. (2015). The fickle P value generates irreproducible results. *Nature Methods*, 12(3), 179–185. <https://doi.org/10.1038/nmeth.3288>
- Hastie, T. J., Tibshirani, R. J., & Friedman, J. H. (2009). *The elements of statistical learning: Data mining, inference and prediction*, 2nd Edn. New York: Springer.
- Hofmann, S. G., Sawyer, A. T., Witt, A. A., & Oh, D. (2010). The effect of Mindfulness-Based therapy on anxiety and depression: a Meta-Analytic review. *Journal of Consulting and Clinical Psychology*, 78(2), 169–183.
- Hölzel, B., Carmody, J., Vangel, M., Congleton, C., Yerramsetti, S. M., Gard, T., & Lazar, S. W. (2011). Mindfulness practice leads to increases in regional brain gray matter density. *Psychiatry Research: Neuroimaging*, 191(1), 36–43.
- Hubbard, R., & Lindsay, R. M. (2008). Why p values are not a useful measure of evidence in statistical significance testing. *Theory & Psychology*, 18(1), 69–88. <https://doi.org/10.1177/0959354307086923>
- Kabat-Zinn, J. (2003). Mindfulness-Based Interventions in context: past, Present, and Future. *Clinical Psychology: Science & Practice*, 10(2), 144–156.
- Keng, S.-L. L., Smoski, M. J., & Robins, C. J. (2011). Effects of mindfulness on psychological health: a review of empirical studies. *Clinical Psychology Review*, 31(6), 1041–1056. <https://doi.org/10.1016/j.cpr.2011.04.006>
- Labatut, V., & Cherifi, H. (2012). Accuracy measures for the comparison of classifiers. CoRR, abs/1207.3790. Retrieved from arXiv:abs/1207.3790.
- Lambdin, C. (2012). Significance tests as sorcery: Science is empirical—significance tests are not. *Theory & Psychology*, 22(1), 67–90. <https://doi.org/10.1177/0959354311429854>
- Latané, B., & Nida, S. (1981). Ten years of research on group size and helping. *Psychological Bulletin*, 89(2), 308–324.
- Levine, M., Cassidy, C., Brazier, G., & Reicher, S. (2002). Self-categorization and Bystander Non-intervention: Two Experimental Studies1. *Journal of Applied Social Psychology*, 32(7), 1452–1463. <https://doi.org/10.1111/j.1559-1816.2002.tb01446.x>
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald, and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*, 46(4), 806.
- Michell, J. (1997). Quantitative science and the definition of measurement in psychology. *British Journal of Psychology*, 88(3), 355–383. <https://doi.org/10.1111/j.2044-8295.1997.tb02641.x>
- Michell, J. (2000). Normal science, pathological science and psychometrics. *Theory & Psychology*, 10(5), 639–667.
- Neyman, J., & Pearson, E. S. (1966). *Joint statistical papers*. Berkeley: Univ of California Press.
- Pearl, J. (2009). *Causality*. Cambridge: Cambridge university press.
- Pearson, E. S. (1939). Student as statistician. *Biometrika*, 30(3/4), 210–250. Retrieved from <http://www.jstor.org/stable/2332648>.

- Piet, J., & Hougaard, E. (2011). The effect of mindfulness-based cognitive therapy for prevention of relapse in recurrent major depressive disorder: a systematic review and meta-analysis. *Clinical Psychology Review*, 31(6), 1032–40. <https://doi.org/10.1016/j.cpr.2011.05.002>
- R Core Team (2015). R: A Language and Environment for Statistical Computing. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <http://www.r-project.org/>.
- Sauer, S., & Luebke, K. (2017a). Data for paper? Observation oriented modeling revised from a statistical point of view? Open Science Framework. <https://doi.org/10.17605/OSF.IO/SGZYD>.
- Sauer, S., & Luebke, K. (2017b). R-code for paper? Observation oriented modeling revised from a statistical point of view? Open Science Framework. <https://doi.org/10.17605/OSF.IO/U3NQ9>.
- Sauer, S., Lemke, J., Wittmann, M., Kohls, N., Mochty, U., & Walach, H. (2012). How long is now for mindfulness mediators? *Personality and Individual Differences*, 52(6), 750–754. <https://doi.org/10.1016/j.paid.2011.12.026>
- Sauer, S., Walach, H., Schmidt, S., Hinterberger, T., Lynch, S., Büssing, A., & Kohls, N. (2013). Assessment of mindfulness: Review on state of the art. *Mindfulness*, 4(1), 1–15. <https://doi.org/10.1007/s12671-012-0122-5>
- Simpson, E. H. (1951). The interpretation of interaction in contingency tables. *Journal of the Royal Statistical Society. Series B (Methodological)*, 13(2), 238–241.
- Slagter, H. A., Lutz, A., Greischar, L. L., Francis, A. D., Nieuwenhuis, S., Davis, J. M., & Davidson, R. J. (2007). Mental training affects distribution of limited brain resources. *PLoS Biology*, 5(6), 1–8. <https://doi.org/10.1371/journal.pbio.0050138>
- Stemmler, M. (2014). Person-centered methods. Springer International Publishing. <https://doi.org/10.1007/978-3-319-05536-7>.
- Stevens, S. S., Volkman, J., & Newman, E. (1937). A scale for the measurement of the psychological magnitude pitch. *Journal of the Acoustical Society of America (JASA)*, 8(3), 185–190. <https://doi.org/10.1121/1.1915893>
- Sturges, H. A. (1926). The Choice of a Class Interval. *Journal of the American Statistical Association*, 21(153), 65–66. Retrieved from <http://www.jstor.org/stable/2965501>.
- Szepannek, G., Luebke, K., & Weihs, C. (2005). Understanding patterns with different subspace classification. In *Machine learning and data mining in pattern recognition* (pp. 110–119). New York City: Springer.
- Thompson, B. (1999). If Statistical Significance Tests are Broken/Misused, What Practices Should Supplement or Replace Them? *Theory & Psychology*, 9(2), 165–181. <https://doi.org/10.1177/095935439992006>
- Thorngate, W., & Ma, C. (2016). Wiggles and curves: The analysis of ordinal patterns (analiza struktur porządkowych). *Problemy Zarządzania*, 14(60), 160–172. <https://doi.org/10.7172/1644-9584.60.10>
- Valentine, K. D., & Buchanan, E. M. (2013). JAM-Boree: An application of observation oriented modelling to judgements of associative memory. *Journal of Cognitive Psychology*, 25(4), 400–422. <https://doi.org/10.1080/20445911.2013.775120>
- Venables, W. N., & Ripley, B. D. (2013). Modern Applied Statistics with S-PLUS. New York City: Springer. Retrieved from <https://books.google.de/books?id=tovgBwAAQBAJ>.
- Wagenmakers, E.-J. (2007). A practical solution to the pervasive problems of p values. *Psychonomic Bulletin & Review*, 14(5), 779–804.
- Wassenhove, V., van Wittmann, M., Craig, D. P. A., & Paulus, M. P. (2011). Psychological and neural mechanisms of subjective time dilation. *Frontiers in Neuroscience*, 5.
- Wasserstein, R. L., & Lazar, N. A. (2016). The asa’s statement on p-values: context, process, and purpose. *The American Statistician*, 70(2), 129–133. <https://doi.org/10.1080/00031305.2016.1154108>
- Wickham, H. (2015). Tidy: Easily tidy data with ‘spread()’ and ‘gather()’ functions. Retrieved from <https://CRAN.R-project.org/package=tidy>.
- Wickham, H., & Chang, W. (2015). Ggplot2: An implementation of the grammar of graphics. Retrieved from <https://CRAN.R-project.org/package=ggplot2>.
- Wickham, H., & Francois, R. (2015). Dplyr: A grammar of data manipulation. Retrieved from <https://CRAN.R-project.org/package=dplyr>.
- Williams, J. M. G., Teasdale, J. D., Segal, Z. V., & Soulsby, J. (2000). Mindfulness-Based Cognitive therapy reduces over-general autobiographical memory in formerly depressed patients. *Journal of Abnormal Psychology*, 109(1), 150–155.
- Ziliak, S. T., & McCloskey, D. N. (2008). *The cult of statistical significance: How the standard error costs us jobs, justice and lives*. Ann Arbor, MI: University of Michigan Press.