Correcting for base rates in multidimensional "Who said what?" experiments

Alexander Bor

Department of Political Science, Aarhus University, Bartholins Allé 7, Aarhus C, 8000, Denmark, alexander.bor@ps.au.dk

Abstract

The 'Who said what?' protocol is a popular experimental paradigm and has been used for 40 years to study spontaneous mental categorization. This paper offers a crucial methodological improvement to calculate unbiased estimates in multidimensional 'Who said what?' studies. Previous studies predominantly corrected for base rates by first correcting the base rates and consequently aggregating errors for the two dimensions separately. The paper demonstrates that this procedure's estimates are biased. A large simulation of over 175,000 experiments and the re-analysis of a pivotal study show that this may increase both false-positive and false-negative error rates in treatment effects and might therefore, respectively, strengthen or weaken evidence for past hypotheses. The paper offers a simple remedy: researchers should first aggregate errors for each dimension and then correct for base rates relying on the method known from single-dimensional studies.

Keywords: Who said what; memory confusion protocol; categorization; experimental methodology; simulation

1 Introduction

Over the past four decades, research has indicated that social categorization is an important cognitive tool contributing to impression formation and stereotyping (Fiske & Neuberg, 1990; Taylor, Fiske, Etcoff, & Ruderman, 1978). People spontaneously sort others into (often implicit) social categories upon which they rely when forming impressions and determining appropriate behavior. The nature of these social categories is therefore of primary interest and has been the focus of social science for quite some time. The most popular and effective experimental method designed to reveal spontaneous social categorization has been introduced by Taylor et al. (1978). The "Who said what?" (WSW) paradigm has been particularly popular among evolutionary psychologists, who have utilized it to demonstrate categorization by kinship (Lieberman, Oum, & Kurzban, 2008), free-riding (Delton, Cosmides, Guemo, Robertson, & Tooby, 2012), deservingness (Petersen, 2012), morality (van Leeuwen, Park, & Penton-Voak, 2012), competence (Bor, 2017), tenure (Cimino & Delton, 2010) and accent (Pietraszewski & Schwartz, 2014), among others. An important "benefit of the categorization measure is that it allows us to

see how subjects spontaneously view this social world" (Delton & Robertson, 2012, p. 718). Using WSW, Kurzban et al. (2001) famously demonstrated that categorization by race is an artefact of our strong, innate propensity to categorize by coalition and can therefore be diminished when race becomes a poor predictor of coalition (see also Pietraszewski, 2016; Pietraszewski, Cosmides, & Tooby, 2014).

This paper provides a crucial methodological improvement to this important literature. It demonstrates that the overwhelming majority of multidimensional WSW studies relied on a faulty method when estimating categorization strength. Statistically, this canonical method yields biased categorization scores and inflates categorization effect size estimates. This might contribute to both false-positive and false-negative treatment effects, thereby, respectively, strengthening or weakening the evidence for past hypotheses. Fortunately, the problem may be ameliorated rather easily within the usual framework and without exploiting useful yet less accessible mathematical models (Klauer & Wegener, 1998).

The paper is structured as follows. First, the logic and standard procedure of WSW experiments are introduced using a simple one-dimensional example. Second, the two-dimensional case is demonstrated along with an intuitive explanation of the canonical and proposed estimation methods. Third, data from a large simulation of WSW experiments is analyzed, providing insights into the effects of the two estimation methods. This also allows interested readers to explore bias for specific parameter combinations. Fourth, a reanalysis of Voorspoels et al.'s (2014) replication of Kurzban et al.'s (2001) seminal study corroborates these findings and demonstrates the benefits of the proposed method.

2 The fundamental logic of WSW experiments

The basic procedure of the WSW experimental protocol is introduced below. For the sake of simplicity, the paper starts by describing experiments where only a single trait is manipulated (1D version) and then proceeds to the multi-trait case (2D version). In a conventional WSW study, participants are asked to watch and form an impression of a number of target individuals (usually eight), who are depicted one-by-one on the screen with a photograph, making one or more statements, or described by one or more sentences in random order. Importantly, the targets are carefully manipulated to differ along one or two dimensions. One such dimension might be any characteristic of the target, either encoded in their photograph or their statements/descriptions. The two categories within a dimension are usually balanced. If we are interested in categorization by gender, four of the eight targets will therefore be men while the other four are women. In the second part of the experiment, there is a short distractor task to clear short-term memory. Finally, there is a surprise recall phase in which the statements/descriptions appear one at a time and participants must pick which target individual (all depicted simultaneously) uttered the given statement.

The errors made by the participants are informative, as they can be sorted into within-category and between-category errors. For example, a sentence originally uttered by a woman and misattributed to another woman is a within-category (or same gender, sG) error. Misattributing the sentence to a man is a between-category (or different gender, dG) error. A within-category error might signal that the respondent relies on the given dimension to form mental groups (or categories) of the targets – correctly identifying the given sentence as belonging to someone

from that group – but fails to remember to whom exactly. Conversely, a between-category error provides no evidence of the given dimension being utilized to group the targets. Consequently, the larger the number of within-category (sG) errors relative to the number of between-category (dG) errors, the stronger support the study provides that the mind is using the given category to categorize targets. Importantly, correct responses are ignored, as it is impossible to know if a correct answer is a product of good memory, categorization, chance, or a combination hereof.

The two types of errors cannot be directly compared, however, as their base rates are different. This becomes clear if we assume that answers are given completely randomly. Following the example with four men and four women with one sentence each, a sentence uttered by a woman can be expected to produce one correct answer ($1corr$: the sentence is by chance attributed to the same woman), three within-category errors ($3sG$: the sentence is misattributed to one of the other three women) and four between-category errors ($4dG$: the sentence is misattributed to one of the four men). To correct for the fact that between-category errors are more likely to occur by chance alone, it is customary to multiply their aggregate number by the ratio of between-category to within-category errors, $(n-1)/n$, where n is the number of targets in a category.[1] This translates to $(4-1)/4 = 0.75$ in studies with eight targets. A categorization score (C) is usually calculated afterwards by subtracting the number of the corrected between-category errors from the number of within-category errors ($C_{gender} = sG - dG \times 0.75$).

3 Multidimensional WSW experiments

The basic protocol can be extended to two dimensions. This is beneficial in situations whenever competing or distracting features may add new insights; for example, a second dimension proved crucial to demonstrate that race encoding is less whenever a better (competing) cue of coalitional affiliations is present (Kurzban et al., 2001), and it helped demonstrate that categorization by morality is strong contrasted with competence (van Leeuwen et al., 2012) or that competence categorization is significant even if variation in likability competes for attention (Bor, 2017).

For an intuitive example, let the two dimensions be gender (male, female) and race (black, white). The two dimensions are usually orthogonal and we end up with four types of targets: two black males, two black females, two white males and two white females. The order of the target types is typically balanced. The experiment is executed as usual but sorting the errors becomes more complicated, as each response now conceals information on two dimensions and, thus, may belong to any of five categories. By chance alone, a sentence uttered by a black woman can be attributed to that same woman ($1corr$), to the other black woman ($1sGsR$: same gender, same race), to any of the two black men ($2dGsR$ different gender, same race), any of the two white women ($2sGdR$ same gender, different race), or to the two white men ($2dGdR$ different gender, different race).

The question then becomes: How do we correct for the base rates in this case? The standard practice is to multiply the number of errors in the last three groups ($dGsR, sGdR, dGdR$) by 0.5,

[1] Mathematically equivalent alternatives include dividing the number of all error types by their base-rate frequencies, multiplying the number of within-category errors by $n/(n-1)$ and so forth.

as they are twice as likely to occur by chance as the first type ($sGsR$). The errors are then aggregated for the two dimensions, for $C_{gender} = (sGsR + sGdR \times 0.5) - (dGsR \times 0.5 + dGdR \times 0.5)$, whereas for $C_{race} = (sGsR + dGsR \times 0.5) - (sGdR \times 0.5 + dGdR \times 0.5)$. This method extends the principle of correcting for the different base rates correctly, however it undermines estimating categorization scores for the two dimensions independently. In other words, if a researcher wants to make a statement about race and/or gender as two independent factors along which categorization may or may not occur (as opposed to categorization by one conditional on the other based on the four joint error-types), their estimates will be biased.

FIGURE 1 around here

This is easy to see with the following intuitive scenario relying on the same two-dimensional race and gender experiment. Let us assume that Participant 1 attributes all of the sentences to the same white woman (i.e., one correct attribution and seven errors) ($Participant\ 1: 1sGsR, 2sGdR, 2dGsR, 2dGdR$). This is illustrated in Figure 1, displaying the original sequence of the targets in the first row (with the subscripts distinguishing between the two targets in the same category) and the respective targets recalled by Participant 1 in the second row. Using the formulas above, her categorization scores will be 0 for both dimensions ($C_{gender} = C_{race} = (1 + 2 \times 0.5) - (2 \times 0.5 + 2 \times 0.5) = 2 - 2 = 0$). This is the result we would intuitively expect, as such a stubborn respondent provides no evidence for categorization.

Now let us assume Participant 2 selects the same white women seven times, but the eighth time she instead picks a black woman, thus (incidentally) committing a same-race, same-gender error ($Participant\ 2: 2sGsR, 1sGdR, 2dGsR, 2dGdR$). Importantly, her responses are identical to Participant 1's gender-wise but slightly more accurate regarding race. This is obvious comparing the answers of the two participants in Figure 1. Participant 2's categorization score for race thus becomes positive ($C_{race} = (2 + 2 \times 0.5) - (1 \times 0.5 + 2 \times 0.5) = 3 - 1.5 = 1.5$). Disturbingly, however, her gender categorization score has also increased ($C_{gender} = (2 + 1 \times 0.5) - (2 \times 0.5 + 2 \times 0.5) = 2.5 - 2 = 0.5$). Even though the two participants' gender responses (ignoring race) are identical, their categorization scores are different. More specifically, the positive change in categorization along race biased the categorization score upwards along the other dimension, gender. This hints at an important substantive implication: The canonical correction method increases false-positive (Type 1) error rates for dimensions crossed with another dimension, where categorization is stronger. Applying the canonical correction method might therefore yield statistically significant estimates even if the data provides no evidence of categorization.

Importantly, a literature review revealed how the vast majority of multidimensional WSW studies have fallen prey to this methodological pitfall. First, 68 published WSW studies were identified using Google Scholar, searching for "'Who said what?' paradigm", "category confusion paradigm", "memory confusion protocol", "memory confusion paradigm" and "statement recognition task". Studies with a single dimension or utilizing the multinomial model (Klauer & Wegener, 1998) were excluded from the analysis, because neither faces the problem of correcting for base rates in a multidimensional setting. This leaves 24 publications (listed in

Table S1 in the online supplementary materials). At least 18 published studies (75% of all relevant studies) report biased estimates.[2]

3.1 An unbiased base-rate correction method

The solution to the problem is straightforward: If we are interested in the two dimensions separately (which is usually the case), we should aggregate error types for the dimension of interest before employing the correction. Consequently, the two formulas for calculating categorization scores become ($C_{gender} = (sGsR + sGdR) - (dGsR + dGdR) \times 0.75$ and $C_{race} = (sGsR + dGsR) - (sGdR + dGdR) \times 0.75$).[3] It is easy to see how this approach circumvents the problem demonstrated by the previous method. The canonical method provides within-category errors by one dimension larger weight if they are paired with a within-category error by the other dimension (i.e., within-within) rather than within-category errors paired with between-category errors (i.e., within-between or between-within errors). The proposed method, however, grants that errors along one dimension receive the same weight notwithstanding the type of error committed along the other dimension with the same response.

Importantly, using the same weight as in one-dimensional studies has the added benefit that it allows for the comparison of unbiased categorization scores between one and two-dimensional experiments. Pietraszewski et al. (2014) correctly identified how it is meaningless to compare categorization scores for one-dimensional and two-dimensional WSW studies after applying different correction methods. However, their proposed solution, namely crossing responses from a 1D study with a random binary variable, relies on transforming 1D studies to 2D studies and applies the erroneous correction method, therefore yielding biased estimates.[4]

Although the base-rate correction method proposed here provides a remedy for most WSW experiments, some studies are interested in the joint error rates rather than comparing two dimensions. In other words, instead of comparing whether categorization is stronger by race or gender, some may be interested in e.g. whether participants are more likely to categorize along both dimensions (committing predominantly $sGsR$ errors; e.g., Stangor, Lynch, Duan, & Glass, 1992) or they might be particularly likely to recall gender but not race (prevalence of $sGdR$ errors). If our interest is in the joint error rates, we might use the canonical correction method; however, we should refrain from calculating "main effects" for the two dimensions separately.

---

[2] Estimates in two further studies may be affected, but one fails to describe the correction method in sufficient detail (van Knippenberg, van Twuyver, & Pepels, 1994), whereas the other excludes relevant trials that might affect the results (Cimino & Delton, 2010).

[3] By equaling this formula with the canonical methods from above and simplifying the equation, the two methods are seen to provide algebraically identical estimates (for gender) whenever $0.5sGdR = 0.25dGsR + 0.25dGdR$. The chances of this are trivial (see Section 4 for a precise estimate).

[4] Moreover, as demonstrated below, the size of the bias is influenced by the size difference between the categorization scores of the two dimensions. It is therefore unlikely that categorization scores from the "pseudo 2D" and real 2D studies are biased equally, which makes comparison across studies even more problematic.

Instead, we may test if any of the four types of error types occurs more frequently than the others with a simple one-way ANOVA or an OLS regression.

4 Exploring the effects of biased correction with a large simulation

To demonstrate the problem with the canonical correction method and prove that the proposed alternative provides a solution, numerous two-dimensional WSW experiments were simulated. The flexibility of the WSW protocol allows researchers to test diverse predictions concerning the relative or absolute strength of categorization by one or two dimensions or their changes across two or more experimental groups. While this contributed to the popularity of the method and led to many important insights, it also hinders efforts to summarize the effects of the two base-rate correction methods on substantive findings. Hence, this analysis primarily focuses on *statistical* bias in categorization scores and effect size estimates of categorization.

The simulated data is generated using five descriptive parameters. The ratio of uncorrected within-category and between-category errors along dimensions A (1) and B (2) are arguably the two most important parameters, as they determine the strength of categorization by the two dimensions. The proportion of correct recalls (3), sample size (4) and number of sentences to recall (5) may also influence the significance tests and are therefore included in the simulation. The parameters were selected to represent the whole range of actual findings based on a review of multidimensional WSW studies. Table S2 in the online supplementary materials shows the specific values of each parameter.

The simulation was performed as follows:[5] First, *m* (parameter 5) recall responses were simulated for 1,000 individuals. Parameters 1–3 set the probability of each response type's occurrence. Various error types were then tallied and (respondent) categorization scores were calculated with both correction methods for each individual. Next, 1,000 experimental groups of size N (parameter 4) were sampled with replacements from this pool of individuals. The biased and corrected (group) categorization scores and corresponding p-values and effect sizes (Pearson's *r*) were calculated for each group. Finally, the average (i.e., expected values of) biased and corrected categorization scores, effect size estimates and p-values were calculated, along with the average bias in categorization scores and effect sizes. The bootstrapped confidence intervals (95% range of values) were also exported for the key statistics (categorization scores, effect sizes and bias estimations).

It should be apparent to those familiar with WSW studies that this process mimics the analytical steps of actual experiments (tallying errors, correcting for base-rates, calculating individual categorization scores, performing statistical tests). The only two differences are that here the initial responses are generated from the parameters and that 1,000 groups are sampled to reveal the distribution of statistics of interest.

---

[5] All of the code and data necessary for the reproduction and replication of the analyses in this paper are available on the Open Science Foundation website: https://osf.io/9f7aw/?view_only=9f44632ed0a745d7bc7b8f82fe9e76b7. An easy-to-use function allows the retrieval of statistics of interest from the data.

This simulation process was repeated for each parameter combination. Because each parameter may take several values and all of the possible combinations of the five parameters were simulated, this yields a large database (over 175,000 observations). This allows for approximating a very wide range of studies.

Do the simulated data provide a good approximation of actual WSW studies? The simulated results were validated against all two-dimensional WSW experiments with published raw data (Bor, 2017; Pietraszewski & Schwartz, 2014; Voorspoels et al., 2014). The validation took place as follows: first, both biased and corrected categorization scores and effect sizes were estimated from the raw data (observed values). Next, the same raw data was used to calculate the specific values taken by the five parameters in a given experiment. Finally, the simulations corresponding to these specific parameter values were retrieved from the simulation data, revealing both biased and corrected categorization scores and effect size estimates (simulated values). The three studies – each with two treatment groups, two dimensions and two correction methods – yield ($3 \times 2 \times 2 \times 2$) 24 pairs of observed-simulated categorization scores and 24 pairs of effect size estimates. Reassuringly, the simulated and observed values are very highly correlated, both for categorization scores ($r = 0.99$) and effect sizes ($r = 0.98$). Table S3 in the online supplementary materials provides all of the observed and simulated values for the 24 categorization exercises.

What is the certainty of getting biased estimates? In specific cases the two correction methods may yield identical categorization estimates (see Footnote 3). To get a more precise estimate of the likelihood of statistical bias, the simulation registered how many of the 1,000 samples for a given parameter combination yield identical estimates. The modal value is 0, the average is 3 and the maximum is 50. In other words, the probability of getting biased estimates is 99.7% on average and never less than 95%.

FIGURE 2 around here

What is the direction of bias? Figure 2 demonstrates the importance of the accurate base-rate correction for a range of scenarios, limiting our attention to the two crucial parameters: the ratio of uncorrected within-category and between-category errors along the two dimensions. Estimates are averaged over the other three parameters. The left panel focuses on categorization scores. Categorization scores for Dimension A (the dimension of interest) calculated either with the canonical method (dashed lines) or with the method proposed in this paper (solid line) are displayed on the y-axis against the full range of uncorrected error ratios on the x-axis. The three shades denote three levels of categorization strength on Dimension B, the darkest shade denoting a very large error ratio (i.e., strong categorization) and the lightest denoting a very low error ratio (i.e., no categorization).

The plot offers a few important insights regarding estimates from both methods and their differences. Categorization score estimates from the canonical method are positively associated with larger error ratios in Dimension A (the lines sloping upwards with higher x-values) and Dimension B (darker lines are higher on the y-axis). Meanwhile, estimates from the proposed method are more strongly associated with changes in Dimension A (the solid lines are steeper) and unaffected by errors in Dimension B (the three solid lines perfectly overlap, appearing as one black line).

The distance between the dashed and solid lines informs us regarding the average bias. Small categorization scores (low x-values) tend to be biased upwards, whereas large categorization scores (high x-values) are biased downwards. The larger the difference between the categorization strengths of the two dimensions, the larger these biases are – the lightest dashed line is furthest from the solid line for large x-values and *vice versa*. Finally, estimates corresponding to higher error ratios are biased more by the canonical method than small values. A more formal analysis is offered in Section S4 in the online supplementary materials.

In turn, the right panel displays simulated effect sizes (Pearson's *r*) in a similar manner for the entire range of error ratios in Dimension A and at three levels of error ratios for Dimension B. Importantly, it reveals that effect size estimates are consistently biased upwards. Effect sizes are conventionally calculated from *t*-values and degrees of freedom using the formula $r = \sqrt{t^2/(t^2 + df)}$.[6] As the biased correction method appears to reduce the variance of categorization strength artificially, the *t*-value and, consequently, the *r*-estimate are both inflated. This has a particularly large effect in small (true) effect sizes pitted against strong categorization in the other dimension (black lines for low x-values).

Substantially, these statistical issues can lead to a number of concerns. First, the upwards bias for small categorization scores increases Type 1 (false-positive) error rates. Second, because large categorization scores are biased downwards, estimates of differences between categorization scores of the two dimensions are consistently diminished. The larger the difference between the two dimensions is, the larger the downward bias, which may increase Type 2 (false-negative) error rates. Finally, experimental treatments affecting exclusively one of the dimensions may "spill over" to the other dimension yielding a statistically significant change without having a real effect on it. This may yield either to false-positive or false-negative errors depending on the predictions of the researcher.

5 Reanalysis of a pivotal experiment

Although the simulation above forcefully demonstrates the importance of proper base-rate correction, it has several limitations. First, it necessarily focuses on categorization along a single dimension (in the presence of another dimension). However, most WSW studies test hypotheses comparing multiple dimensions or the same dimension in multiple experimental groups. Second, by definition, the simulation is artificially generated data. Despite its high correlation with observed data, it remains important to ensure – and interesting to see – how the main conclusions translate into specific studies. The final part of this paper therefore provides a reanalysis of a pivotal WSW experiment.[7]

Perhaps the most influential (certainly the most cited) two-dimensional WSW study is Kurzban, Tooby and Cosmides' (2001, p. 15387) "Can race be erased?" paper. The authors convincingly argue that "encoding by race is … a reversible byproduct of cognitive machinery that evolved to

[6] Section S5 in the online supplementary materials provides further insights regarding the properties of this effect size estimate.

[7] Two other studies are reanalyzed in the online supplementary materials (Sections S6 & S7).

detect coalitional alliances." Unfortunately, the original data is not publicly available. However, Voorspoels and colleagues' (2014) pre-registered replication study offers open access to all materials and data. We rely on this data to demonstrate the validity of the concerns raised above in a highly influential study.

In this experiment, the two orthogonal dimensions are race (Euro-American, African-American) and coalition (one of two rival basketball teams). The eight target individuals are engaged in a heated discussion between the two teams, 24 sentences being uttered in total. Whereas the order of the sentences is fixed (to maintain the logic of the argument), a target is randomly displayed for each sentence. After a short distractor task, respondents attribute each of the 24 sentences to a target. The experiment has two conditions. In the "no visual cue" condition, the coalition status of an individual can be inferred from their position in the debate but not from their appearance, as the targets all wear identical T-shirts. In the visual cue condition, targets wear their team t-shirts (yellow or gray). The authors demonstrate how categorization along race is diminished in the visual cue condition (Hypothesis 1) and that categorization along coalition is increased in the visual cue condition (Hypothesis 2).

TABLE 1 around here

My re-analysis follows the steps outlined by the original authors at every point except that I calculate categorization scores and effect sizes with both the canonical and correct base-rate correction method. Table 1 details categorization scores, standard errors, p-values and effect sizes for both the race and coalition dimensions in both conditions for both methods, along with tests of Hypotheses 1 and 2. It demonstrates all of the concerns raised in previous sections. First, all of the categorization scores are biased towards the categorization score of the other dimension. The true mean categorization by race in the no-cue condition (M = 2.52) is greater than the biased estimate (M = 1.77), as the coalition dimension reveals much weaker evidence of categorization (M = 0.89) and, thus, biases race downwards. The same happens to categorization by coalition in the cue condition. Vice versa, mean categorization by race in the cue condition (M_{biased} = 1.24, $M_{corrected}$ = 0.89) and by coalition in the no-cue condition (M_{biased} = 0.5, $M_{corrected}$ = 0.3) are biased upwards. In fact, the latter case provides a good example of a false-positive error, as the biased estimates yield a statistically significant effect (p = 0.01) even though the corrected estimate does not reach significance (p = 0.26). Second, all of the effect size estimates (Pearson's *r*s) in a single dimension are biased upwards. The bias is particularly large with small effect sizes (e.g., no-cue condition, coalition dimension r_{biased} = 0.18, $r_{corrected}$ = 0.07).

These biases affect the two hypothesis tests and thereby the substantive message of the study. The corrected treatment effect of the cue condition on race (M = –1.63, *r* = –0.20) is considerably larger than the biased estimate (M = –0.53, r = –0.09). In other words, the experiment provided much stronger support for the theory that categorization by race is reduced by better, cross-cutting signals of coalitions. The experiment also provides stronger evidence for H2 – that categorization by coalition is increased by the visual cues – relying on the proposed, unbiased correction method (M = 4.55, *r* = 0.41) than on the biased method (M = 3.00, *r* = 0.39). Importantly, this demonstrates that because the faulty correction method diminishes the true difference between the two dimensions in any study, an experimental treatment that has the

opposite effect on the two dimensions (diminishing one while enhancing the other) will be biased downwards.

6 Conclusions

This paper proposes a crucial methodological improvement for the popular "Who said what?" experimental paradigm (or memory confusion protocol). With the help of an intuitive example, a large simulation and a reanalysis of a prominent study, it demonstrated how the canonical method of correcting base rates simultaneously for the two dimensions and consecutively calculating main effects for the two dimensions is problematic. This method yields categorization scores that are statistically biased towards the crossed dimension and categorization effect sizes that are biased upwards. Future studies interested in the main effects of the two dimensions should instead aggregate first and then correct for varying base rates.

Importantly, this paper seeks to improve and not to attack the social categorization literature. Previous multidimensional WSW studies have contributed to many important insights in evolutionary, social and cognitive psychology and the analysis above suggests that the proposed correction method may even strengthen evidence for a hypothesis. Researchers are, nonetheless, encouraged to revisit their previous (published or unpublished) multidimensional WSW studies and seminal works in their field to investigate if data reanalyzed with the proposed correction method offers any new insights. This paper also provides a poignant reminder that it is important to question the assumptions of even popular tools and procedures.

7 Acknowledgements

I am grateful to Michael Bang Petersen, Florian van Leeuwen, Julianna Bor, and Andrew W. Delton for their help and advice at various stages of this project.

8 Appendix A. Supplementary materials and data

Supplementary materials to this article can be found online at https://doi.org/10.1016/j.evolhumbehav.2018.04.003. All of the code and data necessary for the reproduction and replication of the analyses in this paper are available in an OSF repository at https://osf.io/9f7aw/.

9 References

Bor, A. (2017). Spontaneous categorization along competence in partner and leader evaluations. *Evolution and Human Behavior*, *38*(4), 468–473. https://doi.org/10.1016/j.evolhumbehav.2017.03.006

Cimino, A., & Delton, A. W. (2010). On the perception of newcomers: Toward an evolved psychology of intergenerational coalitions. *Human Nature*, *21*(2), 186–202. https://doi.org/10.1007/s12110-010-9088-y

Delton, A. W., Cosmides, L., Guemo, M., Robertson, T. E., & Tooby, J. (2012). The psychosemantics of free riding: Dissecting the architecture of a moral concept. *Journal of Personality and Social Psychology*, *102*(6), 1252–1270. https://doi.org/10.1037/a0027026

Delton, A. W., & Robertson, T. E. (2012). The social cognition of social foraging: partner selection by underlying valuation. *Evolution and Human Behavior*, *33*(6), 715–725. https://doi.org/10.1016/j.evolhumbehav.2012.05.007

Fiske, S. T., & Neuberg, S. L. (1990). A Continuum of Impression Formation, from Category-Based to Individuating Processes: Influences of Information and Motivation on Attention and Interpretation. *Advances in Experimental Social Psychology*, *23*(C), 1–74. https://doi.org/10.1016/S0065-2601(08)60317-2

Klauer, K. C., & Wegener, I. (1998). Unraveling social categorization in the "who said what?" paradigm. *Journal of Personality and Social Psychology*, *75*(5), 1155–1178. https://doi.org/10.1037/0022-3514.75.5.1155

Kurzban, R., Tooby, J., & Cosmides, L. (2001). Can race be erased? Coalitional computation and social categorization. *Proceedings of the National Academy of Sciences*, *98*(26), 15387–92. https://doi.org/10.1073/pnas.251541498

Lieberman, D., Oum, R., & Kurzban, R. (2008). The family of fundamental social categories includes kinship: evidence from the memory confusion paradigm. *European Journal of Social Psychology*, *38*(6), 998–1012. https://doi.org/10.1002/ejsp.528

Petersen, M. B. (2012). Social Welfare as Small-Scale Help: Evolutionary Psychology and the Deservingness Heuristic. *American Journal of Political Science*, *56*(1), 1–16. https://doi.org/10.1111/j.1540-5907.2011.00545.x

Pietraszewski, D. (2016). Priming Race: Does the Mind Inhibit Categorization by Race at Encoding or Recall? *Social Psychological and Personality Science*, *7*(1), 85–91. https://doi.org/10.1177/1948550615602934

Pietraszewski, D., Cosmides, L., & Tooby, J. (2014). The content of our cooperation, not the color of our skin: An alliance detection system regulates categorization by coalition and race, but not sex. *PLoS ONE*, *9*(2). https://doi.org/10.1371/journal.pone.0088534

Pietraszewski, D., & Schwartz, A. (2014). Evidence that accent is a dedicated dimension of social categorization, not a byproduct of coalitional categorization. *Evolution and Human Behavior*, *35*(1), 51–57. https://doi.org/10.1016/j.evolhumbehav.2013.09.005

Stangor, C., Lynch, L., Duan, C., & Glass, B. (1992). Categorization of individuals on the basis of multiple social features. *Journal of Personality and Social Psychology*, *62*(2), 207–218. https://doi.org/10.1037/0022-3514.62.2.207

Taylor, S. E., Fiske, S. T., Etcoff, N. L., & Ruderman, A. J. (1978). Categorical and contextual bases of person memory and stereotyping. *Journal of Personality and Social Psychology*, *36*(7), 778–793. https://doi.org/10.1037/0022-3514.36.7.778

van Knippenberg, A., van Twuyver, M., & Pepels, J. (1994). Factors affecting social

categorization processes in memory. *British Journal of Social Psychology*, *33*(4), 419–431. https://doi.org/10.1111/j.2044-8309.1994.tb01038.x

van Leeuwen, F., Park, J. H., & Penton-Voak, I. S. (2012). Another fundamental social category? Spontaneous categorization of people who uphold or violate moral norms. *Journal of Experimental Social Psychology*, *48*(6), 1385–1388. https://doi.org/10.1016/j.jesp.2012.06.004

Voorspoels, W., Bartlema, A., & Vanpaemel, W. (2014). Can race really be erased? A pre-registered replication study. *Frontiers in Psychology*, *5*, 1035. https://doi.org/10.3389/fpsyg.2014.01035