Evolution and Human Behavior xxx (xxxx) xxx–xxx

Contents lists available at ScienceDirect

Evolution and Human Behavior

journal homepage: www.elsevier.com/locate/ens

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

ARTICLE INFO

Keywords:
Who said what
Memory confusion protocol
Categorization
Experimental methodology
Simulation

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.

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, Tooby, and Cosmides (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, Bartlema, and Vanpaemel'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

E-mail address: alexander.bor@ps.au.dk.

https://doi.org/10.1016/j.evolhumbehav.2018.04.003
Received 8 September 2017; Received in revised form 1 March 2018; Accepted 14 April 2018

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, 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.

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 Fig. 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 Fig. 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

[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.