Final Report: Direct Replication of Correll (2008, Study 2, *JPSP*)

Etienne P. LeBel
Department of Psychology, Western University, Ontario, Canada

Introduction

Reaction times in behavioral tasks are often used to infer psychological processes. Typically, latencies are averaged across different sets of trials, which ignores a great deal of information about trial-by-trial variation in those latencies. Correll (2008) investigated whether such trial-by-trial variability in latencies could shed new light on the psychological processes underlying implicit racial bias. Specifically, Correll aimed to "test the possibilities that (a) trial-to-trial fluctuations vary in a non-random fashion and (b) the pattern of variability depends on participants' task-related effort" (p. 49). Overall, Correll found that trial-by-trial variation in latencies in a weapon identification task revealed non-random patterns reflecting 1/f noise. Furthermore, it was found that effort to avoid racial bias modulated such non-random 1/f noise in latency variability of the implicit measure whereby increased effort led to less non-random 1/f noise as compared to a baseline control condition.

The target finding for replication is the main finding of Study 2 where participants were assigned to one of three experimental conditions ([1] *use race* during the weapon identification task, [2] *avoid race*, or [3] *control condition* where race went unmentioned). Correll found that participants instructed to *use race* and *avoid race* exhibited less 1/f noise than participants in the control condition (planned contrast: average of experimental conditions compared to control condition). 1/f noise was assessed by applying a fast Fourier transform (FFT) to each participant's wave of latency variability data, yielding power spectral density (PSD) slopes which involve plotting the power of the component waves against their frequency (less negative PSD slopes assumed to reflect more effort).

Methods

Power Analysis

The main result that is the target of replication is Correll's (2008) planned contrast showing a statistically significant difference in the magnitude of the PSD slopes between the control condition and the average of the two experimental conditions, $F(1, 68) = 5.52$, $p < .02$ (p. 56). Correll's F-value of 5.52 translates to a ω^2 of 0.06, which translates to an f of 0.252 (Cohen, 1988). Accordingly, power analyses indicated that sample sizes of 126, 144, 168, and 207 are required to achieve power levels of 80%, 85%, 90%, and 95%, respectively (power estimated using G-Power 3.1; Faul, Erdfelder, Buchner, & Lang, 2009).

Planned Sample

The planned sample size was chosen to be N = 144 (achieving a power of 85%), based on feasibility considerations. The sample of individuals will be drawn from an introductory psychology subject pool from a large Canadian university in London, Ontario, Canada (Western University) where students participate for course credit (same sampling frame as Correll). Demographically, Western University students are primarily White Canadians, middle to upper class, with a sizable international student population. No pre-selection rules used.

Materials

"*The weapon-identification task was based on the paradigm developed by Payne (2001, Study 1). The task consisted of a 25-trial practice phase followed by a test phase of 200 trials. Each*

trial began with the presentation of a prime face (200 ms). Primes consisted of black-and-white photographs (faces only) of five Black men and five White men. Immediately after the prime, a target object appeared (200 ms). Target stimuli consisted of black-and-white photographs of 5 guns and 5 power tools (drills, screwdrivers, etc). After 200 ms, the target image was replaced with a mask (a random pattern of black and white rectangles), which remained on screen until the participant responded. A 1,000-ms inter-trial interval preceded the onset of the next prime. …The test phase of Study 2 imposed no response deadline. To encourage participants to respond quickly, the practice phase imposed a deadline of 1,000 ms. Participants who responded too slowly during the initial 25 trials received a message to that effect."

This was followed precisely. All original stimuli were acquired from the original author. There are only two minor exceptions: (1) I used a standard keyboard to record responses (rather than a response box as used by Correll) because response boxes were not available in the multi-cubicle lab rooms used and (2) on trials where an incorrect response was made, a different beeping sound was heard (though the exact same text information was presented, i.e., "XXX" in red, font 60, presented for 500 ms) because the original author did not have access to the original sound file.

Other methodological details confirmed via email with original author (which were not mentioned in published article): (1) A different random order for the trials was used across participants. (2) The response key for "gun" was on the right whereas response key for "tool" was on the left. (3) Feedback for incorrect responses was presented for both practice and test trials.

Procedure

"…individuals were randomly assigned to one of three conditions, each imposing a different task-relevant goal. Participants in the control condition were given no specific instructions; participants in the avoid-race condition were instructed not to use racial information when making their decisions; participants in the use race condition were instructed to let racial cues guide their behavior during the task. This study followed a 3 (instructions: control vs. avoid race vs. use race) × 2 (prime race: Black vs. White) × 2 (object type: gun vs. tool) mixed-model design, with repeated measures on the last two factors."

This was followed precisely. Exact instruction manipulation wording:

Control condition	Avoid Race condition	Use Race condition
In this study, you will be asked to classify objects that appear on the screen. Immediately before you see each object, a face will appear on screen. You do not need to respond to the face. It simply indicates that an object is coming. When you see the object appear, classify it as follows: If you see a tool, press the "A" key. If you see a gun, press the "5" (NUMPAD) key. Make this classification as ACCURATELY as you can!	In this study, you will be asked to classify objects that appear on the screen. Immediately before you see each object, a face will appear on screen. You do not need to respond to the face. It simply indicates that an object is coming. When you see the object appear, classify it as follows: If you see a tool, press the "A" key. If you see a gun, press the "5" (NUMPAD) key. Make this classification as ACCURATELY as you can! Also, try to respond QUICKLY. You will	In this study, you will be asked to classify objects that appear on the screen. Immediately before you see each object, a face will appear on screen. You do not need to respond to the face. It simply indicates that an object is coming. When you see the object appear, classify it as follows: If you see a tool, press the "A" key. If you see a gun, press the "5" (NUMPAD) key. Make this classification as ACCURATELY as you can! Also, try to respond QUICKLY. You will

Also, try to respond QUICKLY. You will have less than one second! Please call the experimenter now.	have less than one second! ** NOTE ** The faces will be from either White (European American) or Black (African American) people. Research has shown that the race of the face sometimes impacts the ways that people classify the second object. People are sometimes faster and more accurate in responding to guns after a Black face than after a White face. You have been randomly assigned to take the perspective of a completely unbiased person. Regardless of your personal views, we would like you to base your responses only on whether the second object looks more like a gun or tool. *** Try not to let the race of the face influence your decisions. *** Please call the experimenter now.	have less than one second! ** NOTE ** The faces will be from either White (European American) or Black (African American) people. Research has shown that the race of the face sometimes impacts the ways that people classify the second object. People are sometimes faster and more accurate in responding to guns after a Black face than after a White face. You have been randomly assigned to the RACIAL PROFILING condition. Regardless of your personal views, we would like you to play the role of someone engaged in racial profiling. That is, try to make correct classifications, but... *** We would like you to use the race of the faces to help you identify the object in question. *** Please call the experimenter now.

"A White male experimenter introduced the study as an investigation of vigilance. He seated participants in individual rooms, each equipped with a computer. A randomly selected computer program delivered the instructions, which constituted the experimental manipulation. These instructions were taken directly from Payne et al. (2002)."
This was followed precisely except participants will be run in groups of 1 to 4 rather than individually given the much larger sample size (N = 144 vs. N = 71). Also, to parallel the roughly equal sample sizes across conditions in the original study, random assignment to conditions will be determined a priori based on a list of block randomized numbers (1, 2, or 3, using http://www.randomizer.org/). The experimenter will be blind to conditions.

Analysis Plan
The target analysis will follow precisely Correll's analytic strategy, using SAS syntax provided in the appendix of the original article. This will involve two steps. First, two within-subject regressions are performed yielding participant-specific PSD slopes. Second, a between-subjects ANOVA will be performed using two planned orthogonal contrasts, (1) comparing the PSD slopes in the control condition to the average of the PSD slopes in the two experimental conditions (codes: control = -1, avoid race = +1/2, use race = +1/2) and (2) comparing the PSD slopes across the avoid race and use race conditions (codes: control = 0, avoid race = -1/2, use race = +1/2).

Details of the within-participant regressions in the first step:
"A regression was performed for each participant to analyze 1/f noise, modeling the log-transformed latency on a given trial as a function of the target type (Black gun, White gun, Black tool, White gun), accuracy (correct response vs. incorrect response), and trial number. These regressions were designed to remove known sources of variability from the reaction times. Trial number was included to account for increases in speed over the course of the task—a nonstationarity that can disrupt the Fourier analysis. PSD was calculated for each participant

using the Statistical Analysis Software SPECTRA procedure with a Tukey–Hanning window (see Appendix for example syntax). With 200 trials, the FFT decomposed each trial series into 100 component waves. The analysis thus provides estimates of the power and frequency of 100 waves for each participant. A second within-subject regression was performed to quantify the power–frequency relationship for each participant. The goal was to estimate the relationship at low frequencies, but this effort is complicated by the general prevalence of white noise at high frequencies (Gilden, 2001). In essence, the power–frequency relationship is not a straight line; it has a kind of elbow. Clayton and Frey (1997) addressed this nonlinear relationship by excluding high-frequency data and estimating 1/f noise as the linear relationship for the component waves below the elbow. Accordingly, the linear relationship between power and frequency was estimated for log frequencies below zero."

Differences from Original Study

(1) Canadian undergraduates will be used rather than American undergraduates
(2) Participants will be run in groups of 1 to 5 rather than separately as in original study
(3) Keyboard will be used to record responses rather than a response box
(4) A different beeping sound will be heard for incorrect responses

None of these differences are anticipated – based on published and intuitive bases – to alter the likelihood of replicating the effect.

(Post Data Collection) Methods Addendum

Actual Sample

The replication sample consisted of 148 undergraduate students from the University of Western Ontario (108 females, 40 males; mean age = 18.6, *SD* = 3.01) who participated for course credit (ethnicity/race breakdown: 62% Caucasian, 32% Asians (incl. Indians), 2% Blacks, and 5% Other).

Differences from pre-data collection methods plan

None.

Results

Data preparation

Data were prepared exactly as in original paper, using the SAS syntax provided by the original author. For each participant, log-transformed RTs were regressed onto an accuracy code (correct vs. incorrect response; effect coded), trial-type race code (effect coded), trial-type object code (effect coded), race x object product term, and trial number. Residuals of these within-person regressions were saved. SAS' SPECTRA procedure was then used (again at the within-person level) to decompose each participants' 200 residualized RTs into 100 component waves that varied in terms of power and frequency. A second within-person regression was executed yielding person-specific PSD slopes by regressing log-transformed power values (of the 100 component waves) onto log-transformed frequency estimates for log frequencies below zero only.

Confirmatory analysis

The target analysis involved a between-subjects ANOVA using a planned orthogonal contrast comparing the PSD slopes in the control condition to the average of the PSD slopes in the two experimental conditions (codes: control = -1, avoid race = +1/2, use race = +1/2). Inconsistent

with Correll's original finding, this analysis yielded a non-significant contrast, $F(1, 145) = .79$, $p > .37$, $d = .15$ whereby the average of the mean PSD slopes in the experimental conditions ($M_{avoidRace} = -.28$, $SD = .26$ and $M_{useRace} = -.35$, $SD = .33$) were not less negative than in the control condition ($M_{control} = -.37$, $SD = .38$; contrast estimate = .05, with a 95% C.I. of [-.06, .17]), even though the sample was more than twice as large as the original. The current sample achieved a power level of .86 to detect an effect as large as the one reported by Correll (i.e., contrast estimate = .18, $d = .56$).

Mean PSD slopes for participants in the avoid-race condition did not differ from those in the use-race condition, $F(1, 145) = .89$, $p > .35$ (contrast code used: control = 0, avoid race = -1/2, use race = +1/2), which is consistent with that reported by Correll.

Exploratory analyses

Diffusion analyses executed on the RT and error data (Wagenmakers, van der Maas, & Grasman, 2007) performed on both the replication sample data and the original sample data (graciously provided by Correll) revealed that the drift rate (i.e., speed of information accumulation in the stimuli) was lower in my replication sample (M=.32, SD=.09) compared to Correll's sample (M=.37, SD=.11), $F(1,212)=11.24$, $p<.001$, d=.50 (95% C.I. of mean difference = [.02,.08] or .05 +/-.03). This was very surprising given that drift rate is typically manipulated by altering features of the stimuli (e.g., degrading the quality of the stimuli), which of course should have been constant across studies given all experimental stimuli was acquired from Correll. However, this made me consider whether the stimuli appeared physically smaller in my sample, potentially due to a higher screen resolution and/or larger computer monitor. I enquired about this with Correll (somehow neither of us thought to confirm these details prior to data collection) and indeed a smaller screen resolution and computer monitor was used in his original study, meaning that the stimuli physically appeared about 23% smaller in my study. Hence, it is possible (though unlikely; more on this in the Discussion section) that this methodological difference was (in part) responsible for the different pattern of results observed in my sample. I am currently running a follow-up direct replication (data collection under way) duplicating all procedural and methodological details from my first direct replication but increasing the size of the stimuli by 32% so that the stimuli appear to participants the same physical size as in Correll's original study.

Discussion

Summary of Replication Attempt

The current direct replication attempt failed to replicate the original result reported by Correll (2008, Study 2, JPSP). Though the effect was in the expected direction (PSD slopes were numerically less negative in the avoid- and use-race conditions compared to control, presumably reflecting higher effort), the effect – which was about a quarter of the size of the original effect -- was not statistically significant at the $p < .05$ level even though the replication sample size was more than doubled that of the original study and hence I had an 86% chance of detecting an effect as large as the one originally reported given it exists.

Commentary

Though there were some minor (mostly procedural) differences between the original and present study, it seems unlikely that these differences were responsible for the different pattern of results observed. First, the fact that a Canadian rather than American undergraduate sample was used is considered to be an unlikely critical difference given that the behavioral evidence of racial bias on the weapon identification task (in terms of increased # of stereotypically-congruent errors and RTs in the avoid- and use-race conditions compared to control) was actually stronger

in my sample than in Correll's original study (mean differences across conditions for both errors and RTs were statistically significant in my sample whereas only one effect was marginally significant in Correll's sample). This suggests that the Canadian participants had sufficient knowledge of the African-American stereotype. Second, the fact that participants were run in groups of 2 to 5 seems like an unlikely important difference given that participants were wearing headphones in separate cubicles and great care was taken to minimize distractions. Third, using a keyboard rather than a response box also does not seem likely to be an important difference given that standard keyboards are typically accurate to about +/- 7.5 ms (Segalowitz & Graves, 1990). Finally, the different beeping sound used for incorrect responses also seems like an unlikely critical difference.

Furthermore, the fact that the task stimuli physically appeared about 23% smaller to participants in my study could have played a role, but it seems unlikely that this would make that much of a difference based on a priori theoretical grounds (and the fact that neither myself nor Correll thought to confirm this detail prior to data collection). Nonetheless, a follow-up direct replication keeping this methodological detail constant is under way to rule out this concern. The additional N=148 sample will also add to the overall evidence base supporting or contradicting Correll's original finding.

No objections or challenges have been raised by the original author regarding my replication attempt.