Fractal Scaling and Implicit Bias: A Conceptual Replication of Correll (2008)

Mary Jean Amon (amonmj@mail.uc.edu)
University of Cincinnati, Department of Psychology
PO Box 210376, Cincinnati, OH 45221-0376

John G. Holden (john.holden@uc.edu)
University of Cincinnati, Department of Psychology
PO Box 210376, Cincinnati, OH 45221-0376

Abstract

A racial priming article claimed that, relative to a control condition, an exotic variety of variability, called $1/f$ noise, is altered when stereotypes impact participants' judgments in an implicit prejudice task (Correll, 2008). However, Madurski and LeBel (2014) recently described two powerful, faithfully cloned, and apparently decisive studies that each failed to return a successful *literal* replication of Correll's report. Madurski and LeBel outlined and subsequently eliminated several potential extraneous reasons for their replication failures, such as different participant demographics, participant non-compliance, poor psychometrics, and hardware discrepancies. By contrast, this article reports a successful *conceptual* replication of the pattern reported by Correll (cf. Schmidt, 2009). Notably, this conceptual replication required adjustments to Correll's original method and statistical analyses. All the changes were dictated by a systems theory of $1/f$ noise that was largely in place prior to Correll's report (Kello, Beltz, Holden, & Van Orden, 2007; Van Orden, Holden, & Turvey, 2003; 2005). Implications for the replication debate are discussed, with emphasis on contextualizing implicit cues.

Keywords: $1/f$ noise; prejudice; response time; replication; complexity science

A Conceptual Replication of Correll (2008)

Correll's (2008) implicit prejudice paradigm was modeled after a previously published weapon-identification task (Payne, 2001). It used racial priming to contrast automatic and controlled cognitive processes associated with stereotype activation. The task first presented participants with a racial prime (a photograph of a Black or White face) and then replaced the prime with either a stereotype-relevant or -irrelevant target (a gun or a tool). The task recorded response accuracy and response time as participants attempted to quickly identify targets by pressing either a "gun" or a "tool" button on successive trials.

Presumably, common racial biases might account for a participant's tendency to mistakenly identify a tool as a gun, subsequent to the presentation of a Black face as a racial prime. If so, participants' responses are thought to reflect either automaticity—perhaps driven by racial stereotypes—or the deliberate avoidance or invocation of stereotypes (Payne, 2001). That being said, issues persist regarding the generality and reliability of stereotype automaticity effects (e.g., Müller & Rothermund, 2014; Cesario & Jonas, 2014).

Correll (2008) extended Payne's statistical analysis by including a test for $1/f$ noise on each individual participant's successive series of response times—a trial-series—by analogy to a time-series. $1/f$ noise, also known as "pink noise," is a distinctive pattern of long-range correlation in successive measurements, taken more or less contiguously in time. The surprising aspect of pink noise is that it is a statistical fractal. It is comprised of proportionally nested, statistically self-similar patterns of fluctuation. This phenomenon is in stark conflict with the common Gaussian statistical intuition that uncontrolled variability lacks systematic structure.

Uncertain or variable task demands can perturb scaling, as participants attempt to accommodate competing performance goals. As such, task uncertainty forces unanticipated adjustments in the coordinative activity supporting responses in individual trials. Since they are unsystematic—a source of random variability, or *white* noise—their impact is to *whiten*, or weaken an observed scaling relation, relative to a baseline condition that presents more predictable events (Holden, Choi, Amazeen, & Van Orden, 2011; Kello, et al., 2007; Van Orden et al, 2003; Van Orden, Kello, & Holden, 2010; Van Orden, 2009).

Task uncertainty also *obscures* scaling. Fourier decompositions assume completely regular temporal sampling intervals. Unsystematic shifts in the trial-by-trial pace of the experiment introduce apparent but spurious shifts in the frequencies, amplitudes, and phases of the set of ideal sinusoidal functions that are used to decompose the complex empirical waveform. Other things being equal, a response time spectral scaling analysis that relies on unsystematic sampling intervals will be whitened approximately in proportion to the amplitude of the variability in the sampling rate (Holden et al., 2011).

Scaling in computer controlled response time studies arises at the interface of *extrinsic* cognitive and physiological variability and *intrinsic* sources of contextual variability introduced by the event cycle and demands of the task. One implication of this fundamentally adaptive and multi-scale reciprocal coordination is that the effects of laboratory manipulations are causally intertwined with this ongoing coordinative activity (Holden et al., 2011; Van Orden et al., 2010).

Correll's key manipulation gave differential instructions to participants to introduce competing task requirements, relative to a baseline condition. Participants were instructed to complete a weapon identification task while either deliberately using or explicitly avoiding use of racial stereotypes relevant to the priming photographs. Target identities and the depicted race primes were manipulated orthogonally. Thus, attending to race provides a potential to decrement performance. Attempting to explicitly ignore race also invokes a racial frame of reference, which also may decrement performance in a "*don't think of an elephant*" manner (e.g., Lakoff, 2005; Wegner, 1989). Instructing participants to attend to a functionally irrelevant task dimension makes the task just a bit more difficult than it would be otherwise. As such, by the uncertainty hypothesis, invoking and avoiding race is predicted to yield weaker scaling, more similar to white noise than a baseline condition.

Our review of the methodology used in both the original and replication studies suggested several potential methodological artifacts might obscure the impact of the racial priming manipulation on $1/f$ scaling. These artifacts could render the task less sensitive to perturbations induced by the manipulation itself. From the perspective of a widely agreed upon definition of a literal replication—do the same thing twice, and get the same result—the two replication failures of Madurski & LeBel (2014) were clear and decisive. Yet, as we explained, the impact of task uncertainty suggests the pattern of change predicted and reported by Correll is nevertheless plausible. We considered several potentially confounding methodological issues:

1. Both the original and replication studies presented only 200 trials. This is not enough observations to establish compelling scaling relations in either study. Generally, one seeks to establish $1/f$ scaling across at least two orders of magnitude of frequency. Technically, 200 trials appear to meet this criterion (i.e., 2, 20, 200 trials). However, to both comply with this rule of thumb, and to accommodate several major statistical pitfalls of spectral and scaling analyses, collecting at least 1,024 valid trials is strongly recommended (Eke, Herman, Kocsis, & Kozak, 2002; Holden, 2005). *Thus, we presented 1100 trials to participants in every condition of both our literal and conceptual replication attempts*.

2. The response boxes used by Correll were accurate to the nearest millisecond (ms). The response keyboards used to collect the replication data sets were accurate to ±7.5 ms. This difference in precision is inconsequential in studies that pursue differences in mean response time. However, it represents a relatively large amplitude source of unsystematic variability across trials. This added variability is capable of obscuring scaling differences (Holden et al., 2011). By itself, this issue *cannot* explain the replication failure. It is notable, however, that none of the replication control conditions yielded spectral exponents as large as those depicted in Correll's baseline condition. *Both our literal and conceptual replications used ms accurate keyboards, and adopted a symmetric response layout, in which the left-hand 'z' and right-hand '/' keys were mapped to the tool and gun responses, respectively*.

3. The procedure used to compute the power spectra in the original and replication studies is inconsistent with those commonly used in the $1/f$ scaling literature. The response times were log transformed, presumably to obviate the need for outlier censoring. A log transform has little discernable impact on the power spectrum of compact and symmetric Gaussian variables. However, response time distributions entail a potent positive skew. Informal contrasts of untransformed versus log transforms of comparable previously published data sets indicated that a log transform shrank scaling exponents by about 12%. This is problematic for standard subtractive statistical contrasts that use scaling exponents as dependent variables. As one approaches the floor of zero, scaling differences diminish. *We used previously established spectral techniques for all our analyses*.

4. The weapon identification response times are typically very fast (≈300 to 400 ms). As such, the relative time-course of the 1-second inter-trial interval is too long to reveal $1/f$ noise that is closely tied to the trials. The long inter-trial interval effectively shifts the minimum period of sampled change to about 2 sec. Moreover, a perceptibly long downtime between trials hampers the emergence of the close coordination between participant and task that reveals scaling. This uncertainty, in turn, becomes a source of unsystematic variability that both perturbs and obscures scaling—it impairs a task's ability to detect differences in scaling across contrasted cells. *We used a constant 500 ms inter-trial interval in our conceptual replication condition*.

As implemented, the Correll task is unlikely to be particularly sensitive to scaling changes. In light of the exact literal replication failure, we pursued both a literal and an optimized conceptual replication of the weapon identification task. We used four separate experimental cells. We increased the number of trials in all four cells from 200 to 1,100, and we used both ms accurate keyboards, along with a symmetric response-button layout to collect response time data in all cells.

Except for the response-button layout and increase in the number of trials, two cells used a method and event-timing identical to that described by Correll. We sought to optimize two additional experimental cells to detect a scaling difference with changes to ancillary methodological and procedural aspects of the original study. The goal was to use the uncertainty principle to guide improvements in the task's sensitivity to changes in $1/f$ scaling. The optimized cells reduced the inter-trial intervals from 1 sec to 500 ms and increased the presentation duration of the racial primes from 200 to 300 ms. In addition, an error message and beep that followed incorrect responses was removed in the optimized conditions, as these intermittent alerts perturb trial pacing and risk interrupting the coordination between the participant and task. The same experimenter explained the task to each participant in the optimized cells to reinforce the written instructions. Finally, we adopted spectral techniques modeled after those used previously to identify scaling differences response time studies.

Method

Participants A total of 128 undergraduate psychology students participated in exchange for course credit. They were recruited into a study described as investigating vigilance during social tasks. Participants ranged in age from 18 to 41, with a median age of 19. Sixty-five percent of participants were female, 34% were male, and 1% did not specify their gender. Eighty-three percent of participants identified as White, 7% as Black, 5% as Multiracial, 4% as Asian, and 1% as Hispanic. A White female graduate student greeted and consented participants individually. Once informed consent was obtained, participants were seated in front of a computer monitor. All research was carried out in accordance with the protocol approved by the University of Cincinnati's Institutional Review Board.

Design And Procedure Participants completed a two-option, forced choice response time task. Each trial began with the appearance of a *face prime*, drawn from a bank of five Black and five White male pictures. Next, either a hand-tool or handgun was shown and quickly masked by a series of black and white rectangles. The visual mask remained on the screen until the participant selected either the 'z' or the '/' key to identify the object as a tool or a weapon, respectively. All the pictures were presented in a black-and-white format. The task randomly interleaved 1,100 trials that balanced equal numbers of four trial identities: *Black Prime-Tool, Black Prime-Gun, White Prime-Tool,* and *White Prime-Gun*.

Participants were randomly assigned to one of four replication conditions. The first two cells mirrored Correll's control and avoid bias conditions, and maintained the use of a 1 sec inter-stimulus interval, a 200 ms prime presentation, and an error beep-and-message. In line with Correll's procedure, participants in the literal control and avoid bias conditions read instructions for the weapon-identification task off a computer screen.

The two remaining conditions implemented conceptually optimized versions of the control and avoid bias conditions. The inter-trial interval was reduced to 500 ms, and the prime duration was increased to 300 ms. Participants in the optimized control and avoid bias cells received both verbal and screen-printed instructions, in an attempt to maximize their potency. Participants in both versions of the avoid bias cells were informed that some people tend to respond more quickly and accurately to guns after a Black face than after a White face. They were directed to try to avoid racial bias when identifying objects. By contrast, instructions in both versions of the control cells did not raise the topic of race. Immediately following instructions, a single multiple-choice item was administered that asked participants in the optimized cells to indicate their primary goal while completing the computer task. Following their response, participants received feedback reiterating the instructions. The participant's multiple-choice responses were not recorded because of a programming error in the data collection script.

All participants completed 25 practice trials. During the practice trials, if no response occurred by 1 sec, a message to respond faster was displayed. The same warning was displayed after 1.5 sec in the two conceptual replication practice trials to allow participants to learn the task at their own pace. Participants required between 30 and 40 minutes to complete the literal replication cells and 20 to 30 minutes to complete the conceptual replication cells. Two self-report questions and a demographic form were administered once participant's completed the weapon-identification task. The self-report questions asked participants to rate task difficulty and their effort to avoid racial bias on a seven-point scale. Participants were then debriefed and thanked for their time and effort.

Results

The data sets for all consented participants were included in our subsequent statistical analyses. We adapted our statistical analysis for one participant that yielded an idiosyncratic response pattern. The participant produced an apparent 94% error rate (*d*-prime = -3.14), likely a consequence of reversing the keyboard response-mapping. We included this data, but assumed it represented a *d*-prime of 3.14 and a 6% error rate.

We adopted 1-tailed *p* equal to or less than .05 significance levels in all our statistical contrasts. We did this to accommodate the following facts: First, our uncertainty hypothesis makes a clear *a priori* directional prediction that scaling will be reduced, relative to controls, in the avoid bias cells. Increased scaling in the avoid bias condition would contradict the previously established theoretical narrative, the original Correll result, and the Madurski and LeBel (2014) replication failure. Thus, while it was a potential outcome, it would amount to a replication failure because it cannot be interpreted from any established perspective on the manipulation. Second, the previous studies suggest that if an effect is present, it is likely very weak. We used 128 participants, the integer nearest to the 126 participants recruited by Madurski and LeBel that is evenly divisible by 4. Our prediction was that the replication failure would hold in the literal conditions and be overturned in the optimized cells. As such, our key planned statistical contrast, between the two optimized cells, entailed a sample that was about 50% of the size required to yield a power level of .80 according to Madurski and LeBel (2014).

Scaling Analyses The data censoring procedures used to prepare each individual's response time trial-series for spectral analysis were modeled after the steps described in Holden (2005). Each 1,100 trial data set was sorted into the sequential order in which the trials were presented in the experiment. A two-stage procedure was then used to censor extreme observations from the resulting trial-series. First, responses less than 10 ms or greater than 2,000 ms were removed. Second, the series mean and standard deviation were computed, and observations that fell beyond ± 4 standard deviations from an individual's mean response time were also censored. A common ± 3 standard deviation criterion was too restrictive, eliminating more than 75 observations from four trial-series, and at least 50 observations from five additional series. Censorship procedures insure that extreme values do not overwhelm the statistical procedures. However, one must use the most conservative and inclusive thresholds possible, since omitting too many observations dilutes the sequential patterns the fractal analyses seek to reveal (Holden, 2005). Errors were included in the analysis to preserve trial order, (e.g., Gilden, 1997). If more than 1,024 observations remained following the censoring procedures, enough initial trials were deleted to yield a trial-series that contained 1,024 observations, and the series was then transformed into normalized Z-scores. If less than 1,024 observations remained after censoring, the trial-series was normalized into Z-scores, and padded with zeros until it contained 1,024 observations. Zero padding does not impact the scaling exponent. A 127-frequency power spectrum was computed using the procedures described by Holden (2005). A spectral scaling exponent was computed for each participant's trials series. Spectral exponents were derived from the slope of a regression line, computed from the 50 lowest log-log frequency-power data-pairs (Wijnants, Cox, Hasselman, Bosman, & Van Orden, 2012). The analysis was implemented on MATLAB software.

The scaling exponents were subsequently used as dependent variables for two planned contrasts. A between subjects ANOVA contrasted the scaling exponents in the optimized control and optimized avoid bias conditions, testing the hypothesis that instructions to avoid racial bias resulted smaller average scaling exponents—spectral whitening. We observed a barely reliable difference in the predicted direction, r^2 = .06, F(1, 62) = 3.88, p = .05, opt. control M = .39 (SD = .15), opt. avoid M = .31 (SD = .14). As predicted, scaling exponents in the optimized avoid bias cell were on average slightly whitened relative to that of the optimized control cell.

The outcome is consistent with the uncertainty hypothesis, but the difference is hardly compelling when considered in isolation. However, the "effect" was sussed from the ashes of a convincing and powerful replication failure, and the core of the original manipulation was entirely preserved. All our methodological changes were designed to make both optimized cells more sensitive to $1/f$ scaling, resulting in a contrast that was more sensitive to scaling changes. In every case, our methodological and statistical adjustments were either derived from established practice in the scaling literature, or dictated by the emergent coordination and sampling theories that motivated the uncertainty hypothesis. As expected, an identical contrast of the literal replication control and avoid bias conditions failed to reveal a reliable differences in the spectral scaling exponents, $p > .05$.

Figure 1 depicts the power spectra of the optimized control and optimized avoid bias conditions. They illustrate the quadratic shape typical of a power spectrum representative of a mixture of pink and white noise (Holden et al., 2011; Gilden, 1997; Thornton & Gilden, 2005). The optimized avoid bias spectrum diverges from the other spectra as a function decreasing frequency. However, this raises a question: *Is the sole basis of the observed effect an idiosyncratic difference in the lowest-frequency bands*?

This question is best addressed by an alternative statistical analysis called standardized dispersion analysis (SDA analysis, Bassingthwaighte, Liebovitch, & West, 1994; Holden, 2005). SDA is derived from the central limit theorem and is more sensitive to scaling differences in the higher-frequency range, where artifactual whitening introduces a *quadratic* trend that biases spectrum-based scaling exponents. (Thornton & Gilden, 2005; Holden et al., 2011). However, SDA itself is easily biased by low-frequency trends in a trial-series (Van Orden et al., 2003; 2005). For this reason, low-frequency trends are simply removed from each trial-series in advance of the analysis with least-squares linear de-trending.

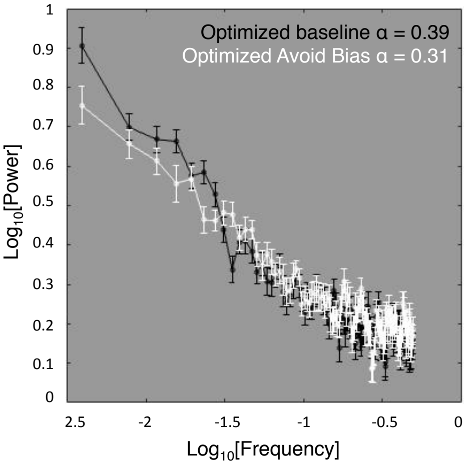

Figure 1: Power spectra of the optimized control and optimized avoid bias conditions averaged across participants. The black markers and lines correspond to the optimized baseline condition, the white markers and lines correspond to the optimized avoid bias condition. The whiskers indicate 1 SEM for each spectral frequency coefficient. The power spectrum of the optimized avoid bias is reliably whitened, relative to the baseline spectrum. The same contrast of the literal replication cells did not reveal a reliable difference.

The SDA analysis yields a fractal dimension statistic (FD) that is analogous to a spectral scaling exponent, but derived from a different statistical framework. An FD value that is statistically equivalent to 1.5 indicates white noise, values less than 1.5 but not greater than about 1.2 are symptomatic of pink noise. It is notable that, when viewed through the lens of SDA, the control and avoid bias conditions mirror each other in the literal and optimized versions of the replication (see Figure 2). SDA analysis does not rely on sinusoidal functions as a mathematical basis, and it is less susceptible to distortions resulting from irregular sampling in time. Neither of the two-cell literal and conceptual replication contrasts yielded a reliable difference. A 2 (task version) × 2 (instruction) ANOVA analyzing FD values based on task version and instruction set revealed a reliable instruction effect, $r^2 = .03$, $F(1, 124) = 3.99$ $p = .05$. The FD measurements distinguished the control and avoid bias instructions. The main effect of task version and the task × instruction interaction were non-significant, $p > .05$. Once again, the difference teeters on the margin of traditional statistical significance. We did not anticipate completing the SDA analysis, or this specific contrast, prior to conducting our replication attempt. We implemented it as an *ad-hoc* check for evidence of scaling differences that were not exclusively contingent on the low-frequency differences established with spectral methods. Our overall conclusion is the race manipulation is relatively weak, but the observed scaling differences are just sufficient to reasonably claim a successful conceptual replication.

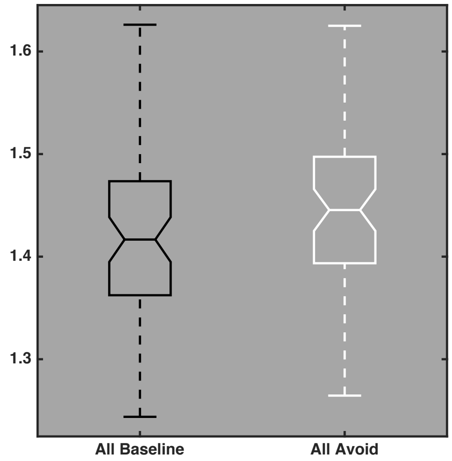

Figure 2: Boxplots of the FD statistics, used to detect differences in scaling. The baseline and avoid bias conditions were collapsed across replication type. SDA is not as sensitive to artifacts in the high frequency range as spectral analysis. A fractal dimension that is statistically equivalent to 1.5 indicates white noise. Fractal dimensions less than 1.5 but no less than about 1.2 indicate pink noise. The FD statistic was reliably larger in the avoid bias condition indicating whitened response time trial-series, relative the baseline condition.

The overall error rate for the different trial types was relatively low, 6.97% (*SD* = 6.32), a pattern consistent with Payne (2001), control *M* = 5.20% (*SD* = 3.48), opt. control *M* = 9.61% (*SD* = 6.30), avoid *M* = 3.69% (*SD* = 2.44), opt. avoid *M* = 9.37% (*SD* = 8.15)]. A *d*-prime statistic measures decision sensitivity. In the weapon identification task, this translates to the ability to both accurately identify weapons and tools, and to refrain from falsely identifying a tool for as a weapon and vise versa. A 2 (Type) × 2 (Instructions) between-subjects ANOVA compared *d*-prime as a function of replication type and bias instructions. There was a main effect of replication type, $r^2 = .19$, $F(1, 124) = 29.30$, $p < .05$, $M_{\text{Literal}} = 3.64$ ($SD = .74$), $M_{\text{Conceptual}} = 2.87$, ($SD = .87$), and no interaction or main effect of bias instructions. The cells in the literal replication yielded larger average *d*-prime values than the conceptual cells. This outcome is unsurprising since the literal replication cells included error feedback and the conceptual replication cells did not. Tests for a Race × Object interaction were non-significant for reaction times and error rates, $p > .05$.

Discussion

We managed to replicate the basic pattern of scaling changes predicted by Correll (2008). To do so, we changed a number of methodological details of the task protocol that were not closely linked to the issue of racial bias. Overall, the impact of the racial primes on participant's decision performance was somewhat weak. In fact, other than scaling changes, we found little compelling evidence for other effects that are typically associated with implicit racial priming in the weapon identification task, such those previously reported in mean response time and error rates. Of course, the task protocol was optimized to detect scaling changes, not differences in error rates or mean effects. In this case, the uncertainty hypothesis dictated methodological changes that overturned two compelling replication failures. It illustrates a strongly counterintuitive success of the emergent coordination framework.

Apparently, methodological details of the task events played a crucial role in the outcome of the study. These details were largely ancillary to the issue of racial priming. On one hand, the trial-series analyses tracked the principal instructional manipulation to avoid racial bias. On the other hand, conventional analyses failed to corroborate a role for implicit racial priming in errors or mean response times. As such, the basis for the difference is difficult to pin down. One possibility is that the scaling changes simply tracked the difference between verbal and written instructions, but that would not explain how the SDA analysis detected the same scaling difference when both the literal and optimized data were aggregated, nor does it explain the difference observed in the optimized cells. The only methodological detail that tracks the observed scaling patterns is the presence or absence of race in the instructions.

More generally, a counterintuitive implication of success of the optimized cells is the possibility that almost no factors are truly benign in experiments designed to isolate the influence of specific factors by holding all others constant. In fact, we take seriously the prospect that all psychological effects are context dependent. Similar paradoxes illustrate why some scientists take lessons learned from disciplines of quantum physics and nonlinear dynamics seriously (e.g., Atmanspacher, Römer, & Walach, 2002; Flach, Dekker, & Stappers, 2007; Gabora & Aerts, 2002; Wang, Solloway, Shiffrin, & Busemeyer, 2014). If context dependency is the norm, it has important implications for the replication crisis and the discipline at large.

Notably, a successful conceptual replication provides more and not fewer potential scientific stances on the effect of implicit bias. One might reasonably conclude the effect is so weak, and likely unreliable, that it is not worthy of inclusion in scientific discourse. On the other hand, while its impact is weak, it is theoretically aligned with other more powerful uncertainty-based scaling manipulations, and it lends credibility to a more general theoretical framework that serves the goals of a scientific enterprise.

References

Atmanspacher, H., Römer, H. & Walach, H. (2002). Weak quantum theory: Complementarity and entanglement in physics and beyond. *Foundations of Physics, 32,* 379-406. doi: 10.1063/1.2158709

Bassingthwaighte, J. B., Liebovitch, L. & West, B. J. (1994). *Fractal physiology*. New York: Oxford University Press.

Cesario, J., & Jonas, K. J. (2014). Replicability and models of priming: What a resource computation framework can tell us about expectations of replicability. *Social Cognition, 32,* 124-136.

Correll, J. (2008). $1/f$ noise and effort on implicit measures of bias. *Journal of Personality and Social Psychology, 94,* 48-59. doi: 10.1037/0022-3514.94.1.48

Eke, A., Herman, P., Kocsis, L., & Kozak, L. R. (2002). Fractal characterization of complexity in temporal physiological signals. *Physiological Measurement, 23*, R1. doi: 10.1088/0967-3334/23/1/201

Flach, J. M., Dekker, S., & Stappers, P. J. (2007). Playing twenty questions with nature (the surprise version): Reflections on the dynamics of experience. *Theoretical Issues in Ergonomic Science, 9*, 125–154.

Gabora, L., & Aerts, D. (2002). Contextualizing concepts using a mathematical generalization of the Quantum formalism. *Journal of Experimental and Theoretical Artificial Intelligence, 14*, 327-358. doi: 10.1080/09528130210162253

Gilden, D. L. (1997). Fluctuations in the time required for elementary decisions. *Psychological Science, 8,* 296-301. doi: 10.1111/j.1467-9280.1997.tb00441.x

Holden, J. G., (2005). Gauging the fractal dimension of cognitive performance. In M. A. Riley & G. C. Van Orden (Eds.), *Tutorials in Contemporary Nonlinear Methods for the Behavioral Sciences*. Retrieved from http://www.nsf.gov/sbe/bcs/pac/nmbs/nmbs.jsp

Holden, J. G., Choi, I., Amazeen, P. G., & Van Orden, G. (2011). Fractal $1/f$ dynamics suggest entanglement of measurement and human performance. *Journal of Experimental Psychology: Human Perception & Performance, 37*, 935-948. doi: 10.1037/a0020991

Kello, C. T., Beltz, B. C., Holden, J. G., & Van Orden, G. C. (2007). The emergent coordination of cognitive function. *Journal of Experimental Psychology: General, 136,* 551-568. doi: 10.1037/0096-3445.136.4.551

Lakoff, G. (2005). *Don't think of an elephant! Know your values and frame the debate.* White River Junction, VT: Chelsea Green Publishing.

Madurski, C., & LeBel, E. P. (2014). Making sense of the noise: Replication difficulties of Correll's (2008) modulation of $1/f$ noise in a racial bias task. *Psychonomic Bulletin & Review.* doi: 10.3758/s13423-014-0757-4

Müller, F., & Rothermund, K. (2014). What does it take to activate stereotypes? Simple primes don't seem enough: A replication of stereotype activation (Banaji & Hardin, 1996; Blair & Banaji, 1996). *Social Psychology, 45*, 187-193. doi:10.1027/1864-9335/a000183

Payne, B. K. (2001). Prejudice and perception: The role of automatic and controlled processes in misperceiving a weapon. *Journal of Personality and Social Psychology, 81,* 181-192. doi: 10.1037//0022-3514.81.2.181

Schmidt, S. (2009). Shall we really do it again? The powerful concept of replication is neglected in the social sciences. *Review of General Psychology, 13,* 90-100. doi: 10.1037/a0015108

Thornton, T.L. & Gilden, D.L. (2005). Provenance of correlations in psychological data. *Psychonomic Bulletin & Review, 12*, 409-441. doi: 10.3758/BF03193785

Van Orden, G. C. (2009). Voluntary performance. *Medicina, 46,* 581-594.

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2003). Self-organization of cognitive performance. *Journal of Experimental Psychology: General, 132*, 331-350. doi: 10.1037/0096-3445.132.3.331

Van Orden, G. C., Holden, J. G., & Turvey, M. T. (2005). Human cognition and $1/f$ scaling. *Journal of Experimental Psychology: General, 134*, 117-123. doi: 10.1037/0096-3445.134.1.117

Van Orden, G. C., Kello, C. T., & Holden, J. G. (2010). Situated behavior and the place of measurement in psychological theory, *Ecological Psychology, 22,* 24-43. doi: 10.1080/10407410903493145

Wang, Z., Solloway, T., Shiffrin, R. M., & Busemeyer, J. R. (2014). Context effects produced by question orders reveal quantum nature of human judgments. *Proceedings of the National Academy of Sciences of the United States of America, 111,* 9431-9436. doi: 10.1073/pnas.1407756111

Wegner, D. M. (1989). *White bears and other unwanted thoughts: Suppression, obsession, and the psychology of mental control.* New York: Viking/Penguin.

Wijnants, M. L., Cox, R. F. A., Hasselman, F., Bosman, A. M. T., & Van Orden, G. (2012). Does sample rate introduce an artifact in spectral analysis of continuous processes? *Frontiers in physiology*, *3.* doi: 10.3389/fphys.2012.00495