arTs Università degli Studi di Trieste
Archivio della ricerca – postprint

Åhs et al.'s (2018) Systematic review on biological preparedness and resistance to extinction: A commentary and reanalysis

Marco Del Giudice *

University of New Mexico, USA

Åhs and colleagues (2018) systematically reviewed 23 studies testing the hypothesis that "biologically prepared" or "fear-relevant" conditioned stimuli (CS) are more resistant to extinction than neutral stimuli. Specifically, the authors focused on skin conductance responses to pictures of snakes and spiders; in most of the studies, the hypothesis was tested by the interaction between differential conditioning (CS + vs. CS−) and fear-relevance (snakes/spiders vs. neutral stimuli) during the extinction phase.[1] An experiment was coded as supporting the hypothesis if the critical test was statistically significant with $p < .05$. Out of the 27 tests reported in the authors' Table 1, only 10 yielded statistically significant results. The authors interpreted these findings as evidence that preparedness effects have poor replicability, and speculated that publication bias may explain the initial string of positive results obtained by Öhman and collaborators between 1975 and 1981.

Unfortunately, the vote-counting method employed in the review is severely limited, as it fails to take statistical power into account. If the studies testing a true hypothesis are systematically underpowered, they will still yield a high proportion of "null" findings. In fact, a majority of null findings may reflect a relative *lack* of publication bias coupled with low power. Note that, in this case, the effects of interest are attenuated interactions, which typically account for small amounts of variance and require large samples for adequate power (Blake and Gangestad, 2020). Åhs et al. acknowledged the limitations of vote-counting and regretted not being able to perform a formal meta-analysis, because descriptive statistics were not reported in the original papers and the raw data were no longer available.

Even though detailed descriptive statistics are lacking, the papers in the systematic review report *F*-ratios and degrees of freedom for all the significant tests and some of the non-significant ones. I used this information to compute exact *p*-values and an index of effect size (ES), the adjusted partial eta squared ($Adj.\eta_p^2$; see Mordkoff, 2019).[2] The values are reported in Table 1. Inspecting the original studies, I found three inconsistencies in the systematic review. (1) The *N* for the interaction test in Dawson et al. (1986) was 79 instead of 144. (2) The study by Hugdahl and Öhman (1977) did not test the critical interaction because of assumption violations, and should be counted as uninformative rather than null. (3) Åhs et al. counted the study by Schell et al. (1991) as a negative finding because extinction was preceded by reconditioning, and there was no significant interaction before reconditioning. But the hypothesis addressed in the review specifically concerns *extinction*; moreover, the reconditioning consisted of only 4 trials with a single reinforcement of the CS+, which were followed by 92 extinction trials. Even if the procedure was somewhat non-standard, there is a reasonable case for testing the target hypothesis on the extinction trials, and counting this as a positive finding. Note that the reanalysis I report below gives the same results regardless of whether the Schell et al. study is included or excluded.

To assess the evidentiary value of the study set, I performed a *p*-curve analysis (Simonsohn et al., 2014) with the online application at https://p-curve.com. The *p*-curve is the distribution of statistically significant *p*-values in a set of studies, and can be used to draw inferences about the likelihood that the positive findings reflect a true effect versus selective reporting, including publication bias and "p-hacking". In addition, *p*-curve analysis estimates the average power of the studies if an effect exists. Crucially, this method does not suffer from the "file drawer problem" because it relies only on *significant* tests. The results are shown in Fig. 1. The analysis suggested that the set of studies has evidentiary value—i.e., the positive findings are not solely explained by selective reporting. The estimated average power was 71 % (with 90 % CI [39 %, 90 %]). Excluding the study by Schell et al. (1991) yielded virtually identical results ($p < .0001$ for right-skewness, 70 % power). Removing the first and second smallest *p*-value as a robustness check did

* Corresponding author at: Department of Psychology, University of New Mexico, Logan Hall, 2001 Redondo Dr. NE, Albuquerque, NM, 87131, USA.
E-mail address: marcodg@unm.edu.

[1] Åhs et al. did not specify which skin conductance measure they selected when multiple ones were available. I always selected the first interval anticipatory response (labeled FIR or FAR), which was the only measure reported in all the studies; the results of the tests agreed with those reported by Åhs et al. (2018).

[2] The formula is: $Adj.\eta_p^2 = \frac{(F-1) \times df_{effect}}{F \times df_{effect} + df_{error}}$.

Table 1
Revised statistics for the fear-relevance effects reviewed in Åhs et al. (2018).

Study	Test *N*	Test statistic	$Adj.\eta_p^2$	*p*-value	
Björkstrand et al. (1990)					
Experiment 1	64	Not reported		> .050	
Experiment 2	64	Not reported		> .050	
Booth et al. (1989)					
Experiment 1	48	Not reported		> .050	
Experiment 2	48	Not reported		> .050	
Cook et al. (1986)					
Experiment 1–6	292	$F(1, 288) = 1.24$	.001	.266	
Dawson et al. (1986)	79	$F(1, 75) = 3.11$	.027	.082	
Fredrickson et al. (1976)	48	$F(1, 44) = 4.07$	.064	.050	*
Fredrickson & Öhman (1979)	32	$F(1, 30) = 18.29$	.358	< .001	*
Ho & Lipp (2014)	40	Not reported		> .050	
Hugdahl et al. (1977)	64	$F(1, 60) = 47.18$	.431	< .001	*
Hugdahl & Kärker (1981)	45	$F(2, 42) = 3.44$	.100	.041	*
Hugdahl and Öhman (1977)	56	Effect not tested		N/A	
Kirsch & Boucsein (1994)	28	Not reported		> .050	
Kirsch & Boucsein (1997)	42	Not reported		> .050	
Lovibond et al. (1993)					
Experiment 2	96	$F(1, 92) = 1.1$	.001	.297	
Lipp & Edwards (2002)	64	$F(1, 60) = 19.63$	.234	< .001	*
McNally & Foa (1986)	38	Not reported		> .050	
McNally (1986)	24	Not reported		> .050	
Neumann & Longbottom (2008)					
Experiment 1	64	$F's < 1.30$		> .050	
Experiment 2	74	Not reported		> .050	
Öhman et al. (1975a)	120	$F(2, 108) = 4.31$	.057	.016	*
Öhman et al. (1975b)	64	$F(2, 120) = 6.77$	.086	.002	*
Öhman et al. (1976)					
Experiment 1	60	$F(2, 57) = 4.09$	.095	.022	*
Experiment 3	40	$F(1, 36) = 10.29$	.201	.003	*
Schell et al. (1991)	163	$F(1, 147) = 9.46$	.054	.003	*
Stussi et al. (2018)					
Experiment 3	40	$F(1.73, 67.62) = 4.68$	.084	.016	*
Thompson & Lipp (2017)	25	$F's < 1.72$		> .050	

Note. See Åhs et al. (2018) for complete references. Asterisks indicate $p < .050$. The *p*-value for the study by Fredrickson et al. (1976) is .04978, displayed as .050 in the table. $Adj.\eta_p^2$ = adjusted partial eta squared.

not change the evidentiary value of the set, but reduced the estimated power to 48 % and 31 %, respectively. These figures are more in line with the actual proportion of statistically significant effects in the study set (11 out of 26 or 42 %; see Table 1), also considering that additional null results remain in the file drawer (Åhs et al., 2018, p.436).

The eta squared values compiled in Table 1 should be interpreted with caution. First, they are likely to be substantially biased upward since they mostly come from statistically significant tests. Second, the size of the partial eta squared partly depends on the specifics of the analysis (e.g., which other factors were included; see Lakens, 2013), so that values from different studies may not be fully comparable. To obtain a rough estimate of the underlying ES, I converted the eta-squared values to Cohen's *f*, calculated their *N*-weighted mean, then converted back to eta-squared.[3] To reduce bias, I imputed the missing values for non-significant tests with the weighted mean for the available non-significant tests. The mean ES thus estimated was $\overline{Adj.\eta_p^2} = .035$. Excluding the studies conducted by Öhman's research group reduced the estimate to $\overline{Adj.\eta_p^2} = .012$, consistent with the idea that initial findings were inflated by selective reporting. With the typical design for this kind of study (2 within-subjects × 2 between-subjects), achieving sufficient power with $\eta_p^2 = .012$ may require hundreds of participants,

[3] The conversion formulas are: $f = \sqrt{\frac{\eta_p^2}{1-\eta_p^2}}$ and $\eta_p^2 = \frac{f^2}{1+f^2}$.

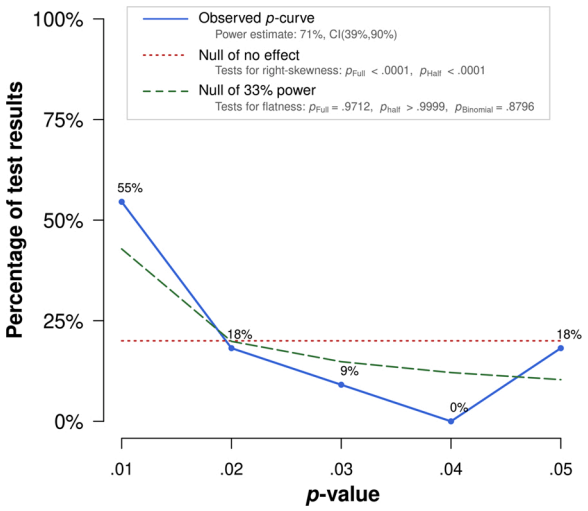

Fig. 1. *P*-curve analysis of the 11 significant *p*-values ($p < .05$) reported in Table 1. The *p*-curve (solid blue line) is the empirical distribution of statistically significant *p*-values in the study set. The dotted red line is the expected distribution of *p*-values under the null hypothesis of no effect. The dashed green line is the expected distribution of *p*-values for a set of studies with an average power of 33 %. The *p*-curve becomes right-skewed when the null hypothesis is false (i.e., the true effect is not zero); hence, the overall evidentiary value of the study set can be assessed by testing the curve for right-skewness (both tests were significant with $p < .05$). The additional flatness tests were not statistically significant, indicating that the *p*-curve is not significantly flatter than expected under 33 % power.

depending on the strength of the correlation between CS + and CS− measures. Since this information was not reported in the original studies, I calculated the required *N* to achieve 80 % power with $r = .30$ ($N = 230$), $r = .50$ ($N = 164$), and $r = .70$ ($N = 100$). Assuming the same ES, the "median-sized study" for this set ($N = 54$) would have a power of about 27 % with $r = .30$, about 36 % with $r = .50$, and about 54 % with $r = .70$ (calculated with G*Power 3.1; Faul et al., 2009).

In sum: despite some indications of selective reporting, the available data tentatively support the hypothesis that fear-relevant stimuli are more resistant to extinction. However, the magnitude of this effect seems much smaller than indicated by the initial findings, and most of the published studies lacked sufficient power to reliably detect it. While preparedness effects may be too weak to account for the etiology of phobias, they remain theoretically important; the present reanalysis tempers Åhs et al.'s original conclusions, and suggests that preparedness theory should not be prematurely dismissed. At the same time, the evidence available to date remains far from conclusive. The best way to resolve the debate would be to run one or more high-powered, preregistered studies, with sufficient sensitivity to detect the likely effect of preparedness.

References

Åhs, F., Rosén, J., Kastrati, G., Fredrikson, M., Agren, T., Lundström, J.N., 2018. Biological preparedness and resistance to extinction of skin conductance responses conditioned to fear relevant animal pictures: a systematic review. Neurosci. Biobehav. Rev. 95, 430–437.

Blake, K.R., Gangestad, S., 2020. On attenuated interactions, measurement error, and statistical power: guidelines for social and personality psychologists. Pers. Soc. Psychol. Bull. doi: 0146167220913363.

Dawson, M.E., Schell, A.M., Banis, H.T., 1986. Greater resistance to extinction of electrodermal responses conditioned to potentially phobic CSs: a noncognitive process? Psychophysiology 23, 552–561.

Faul, F., Erdfelder, E., Buchner, A., Lang, A.G., 2009. Statistical power analyses using G* Power 3.1: tests for correlation and regression analyses. Behav. Res. Methods 41, 1149–1160.

Hugdahl, K., Öhman, A., 1977. Effects of instruction on acquisition and extinction of electrodermal responses to fear-relevant stimuli. J. Exp. Psychol. Hum. Learn. 3, 608–618.

Lakens, D., 2013. Calculating and reporting effect sizes to facilitate cumulative science: a practical primer for t-tests and ANOVAs. Front. Psychol. 4, 863.
Mordkoff, J.T., 2019. A simple method for removing bias from a popular measure of standardized effect size: adjusted partial eta squared. Adv. Methods Pract. Psychol. Sci. 2, 228–232.
Schell, A.M., Dawson, M.E., Marinkovic, K., 1991. Effects of potentially phobic conditioned stimuli on retention, reconditioning, and extinction of the conditioned skin conductance response. Psychophysiology 28, 140–153.
Simonsohn, U., Nelson, L.D., Simmons, J.P., 2014. *P*-curve: a key to the file-drawer. J. Exp. Psychol. Gen. 143, 534–547.