JOURNAL OF THE EXPERIMENTAL ANALYSIS OF BEHAVIOR 2007, **87,** 405–407 NUMBER 3 (MAY)

WHEN IS A FAILURE TO REPLICATE NOT A TYPE II ERROR?

MARCO VASCONCELOS AND PETER J. URCUIOLI

PURDUE UNIVERSITY

AND

KAREN M. LIONELLO-DENOLF

UNIVERSITY OF MASSACHUSETTS MEDICAL SCHOOL/SHRIVER CENTER

Zentall and Singer (2007) challenge our conclusion that the work-ethic effect reported by Clement, Feltus, Kaiser, and Zentall (2000) may have been a Type I error by arguing that (a) the effect has been extensively replicated and (b) the amount of overtraining our pigeons received may not have been sufficient to produce it. We believe that our conclusion is warranted because (a) the original effect has not been replicated despite multiple attempts to do so and (b) the statement that more extended overtraining may be needed itself suggests that the original effect is not reliable.

Key words: work-ethic effect, Type I error, Type II error, within-trial contrast, overtraining, replication

Vasconcelos, Urcuioli, and Lionello-DeNolf (2007) report six experiments that attempted, unsuccessfully, to replicate the ''work ethic'' effect in pigeons reported by Clement, Feltus, Kaiser, and Zentall (2000). Experiments 1–5 closely modeled the Clement et al. methodology, whereas Experiment 6 followed a somewhat different design in order to obtain an independent index of the differential aversiveness of low- versus high-effort trials. All experiments returned the same pattern of null findings: On probe-trial preference tests, pigeons were, on average, indifferent between the S+ stimuli that followed high and low effort and, for the most part, between the two corresponding S– stimuli. Given these results, we concluded that ''…the *seminal* demonstration of the work-ethic effect in pigeons (Clement et al., 2000) is not a reliable finding'' (p. 396, italics added) or, as we put it in our introduction, ''…our null findings …underscore…the possibility that the *original* findings may have been a Type I error…'' (p. 383, italics added).

Zentall and Singer (2007) argue that our conclusion is not warranted because (a) the within-trial contrast effect has been extensively replicated and (b) the amount of overtraining used in our experiments was insufficient (i.e., had our pigeons been given sufficient overtraining, they would likely have exhibited preferences of the sort reported by Clement et al., 2000). We feel that the first rejoinder confuses the process of statistical decision-making with theoretical evaluation and that the second raises new issues that can only be settled by future research.

IS THE WORK-ETHIC EFFECT RELIABLE?

It is important not to confuse the work-ethic effect described by Clement et al. (2000) with the proposed explanation of it, within-trial contrast. Zentall and Singer (2007) appear to treat these two things interchangeably (see, for example, their Table 1). Our position is that for the purposes of statistical evaluation and decision-making, the work-ethic effect and within-trial contrast are not synonymous and should be kept separate because they pertain to two different domains of inquiry. The former term was coined to describe the observed preferences that the Clement et al. (2000) pigeons exhibited for a stimulus obtained following 20 pecks over a stimulus obtained following a single peck. Within-trial contrast, however, is a theoretical process hypothesized to underlie the work-ethic effect (cf. Zentall, Clement, Friedrich, & DiGian, 2006) as well as observable preferences following other experimental manipulations. In short, the former labels a particular empirical finding, whereas the latter refers to a proposed mechanism that could produce such a finding.

Address correspondence to: Marco Vasconcelos, Department of Psychological Sciences, Purdue University, 703 Third Street, West Lafayette, IN 47907. E-mail: marcov@psych.purdue.edu.

doi: 10.1901/jeab.2007.10-07

The concept of a Type I error does not have any meaning in the domain of theoretical proposals. Rather, the concept is meaningful only for empirical findings on which statistical decisions are based (see Bower & Mayer, 1989; Bulevich, Roediger, Balota, & Butler, 2006; Fernandez & Glenberg, 1985 for similar examples in the human memory literature). Hence, we were careful not to question whether or not within-trial contrast is a real and influential psychological process. Our concern was the reliability of the particular behavioral effect reported by Clement et al. (2000)—in other words, with direct (or systematic) replication rather than with conceptual replication.

In their Table 1, Zentall and Singer (2007) list many experiments in support of within-trial contrast. But most of these involve conceptual replications of the original (Clement et al., 2000) findings from which the idea of within-trial contrast emerged. We disagree, then, with the implication that attributing the original findings to chance (random) factors can be justified only with 250 null-effect experiments of the sort listed in that table. By this argument, attributing a significant effect following a particular contextual manipulation (e.g., Smith, 1979) to chance is justifiable only if there are many hundreds (thousands?) of other ineffective retrieval cue manipulations.

To our knowledge, only our series of experiments and Klein and Zentall (2002) attempted a direct or systematic replication of Clement et al. (2000). Our paper involved six separate experiments in which there were a total of 12 groups/conditions that were used to detect a possible work-ethic effect. None did. Counting Klein and Zentall's unpublished failure-to-replicate plus one additional assessment run in our lab that we did not report (but returned the same result), the count stands as one demonstration of the work-ethic phenomenon in pigeons and 14 null findings. Although some might view this as insufficient to attribute the original findings to chance, we feel that this conclusion is more parsimonious than attributing 14 failed replications to Type II errors.

WAS THE AMOUNT OF OVERTRAINING INSUFFICIENT?

Zentall and Singer (2007) argue that the amount of overtraining used in our experiments may not have been sufficient to observe the effect and suggest that when sufficient overtraining is provided a reliable effect can be found. With an operational definition of ''sufficient'' stated in advance, this is testable, and it is entirely possible that a preference would be apparent with more sessions of overtraining. Determining what parameters are necessary to reliably produce a phenomenon is important for empirical and theoretical development. At this stage, however, it remains an open question whether the work-ethic effect in pigeons is a real phenomenon. If it is, then knowing the conditions under which it appears and those under which it does not will benefit our eventual understanding of the underlying psychological processes. If it is not, then that information, too, is important in the context of preferences arising from ostensibly related manipulations for similar reasons.

If, as Zentall and Singer (2007) argue, the lower limit to observe the effect is about 20 overtraining sessions and ''…at that level the reliability of the effect is questionable'' (p. 402), then it appears we are in agreement. In other words, if the argument is that under the *particular* training conditions used by Clement et al. (2000) researchers will obtain a statistically significant work-ethic effect only once in every 20 or so attempts, that, to us, means that such a result, when obtained, is probably due to random factors.

CONCLUSIONS

Replication is one of the most important self-corrective mechanisms in science. Effects caused by random factors periodically will appear in the literature and, as scientists, we are susceptible to their influence just as we are for true effects. Careful replication can help us make that important distinction and clarify what variable or combination of variables produces reliable behavioral effects. Naturally, failures to replicate should be carefully scrutinized for procedural integrity, measurement sensitivity, statistical power, and the like. We believe our experiments meet these criteria, and Zentall and Singer (2007) seem to agree. We contend that the work-ethic effect in pigeons is not reliable under the conditions originally reported. What conditions, if any, will generate this effect in pigeons will be decided by future experiments.

REFERENCES

Bower, G. H., & Mayer, J. D. (1989). In search of mood-dependent retrieval. *Journal of Social Behavior and Personality, 4*, 121–156.

Bulevich, J. B., Roediger, H. L. III, Balota, D. A., & Butler, A. C. (2006). Failures to find suppression of episodic memories in the think/no-think paradigm. *Memory & Cognition, 34*, 1569–1577.

Clement, T. S., Feltus, J., Kaiser, D. H., & Zentall, T. R. (2000). ''Work ethic'' in pigeons: Reward value is directly related to the effort or time required to obtain the reward. *Psychonomic Bulletin & Review, 7*, 100–106.

Fernandez, A., & Glenberg, A. M. (1985). Changing environmental context does not reliably affect memory. *Memory & Cognition, 13*, 333–345.

Klein, E. D., & Zentall, T. R. (2002). [Failure to replicate the ''work ethic'' effect in pigeons]. Unpublished raw data.

Smith, S. M. (1979). Enhancement of recall using multiple environmental contexts during learning. *Memory & Cognition, 10*, 405–412.

Vasconcelos, M., Urcuioli, P. J., & Lionello-DeNolf, K. M. (2007). Failure to replicate the ''work ethic'' effect in pigeons. *Journal of the Experimental Analysis of Behavior, 87*, 383–399.

Zentall, T. R., & Singer, R. A. (2007). Within-trial contrast: When is a failure to replicate not a Type I error? *Journal of the Experimental Analysis of Behavior, 87*, 401–404.

Zentall, T. R., Clement, T. S., Friedrich, A. M., & DiGian, K. A. (2006). Stimuli signaling reward that follow a less-preferred event are themselves preferred: Implication for cognitive dissonance. In E. A. Wasserman, & T. R. Zentall (Eds.), *Comparative cognition: Experimental explorations of animal intelligence* (pp. 651–667). NY: Oxford University Press.

Received: January 26, 2007
Final acceptance: January 29, 2007