An invalid test of the original experiment: Comment on the Reproducibility Project's attempt to replicate Amodio, Devine, & Harmon-Jones (2008)

David M. Amodio[1], Patricia G. Devine[2], & Eddie Harmon-Jones[3]

[1]New York University, [2]University of Wisconsin Madison, [3]University of New South Wales

Correspondence: david.amodio@gmail.com

Abstract: We describe several methodological deviations of this replication attempt from the original study which likely contributed to its failure. In cases such as these, the result should be described as inconclusive rather than a nonreplication.

We very much appreciate the time and effort of the replication team, and we certainly acknowledge the importance of replication as part of scientific progress. However, the attempt to replicate our study (Amodio, Devine, & Harmon-Jones, 2008, Study 2) was limited by critical methodological problems that rendered the test invalid—that is, unable to provide a psychometrically valid test of the original hypothesis. We describe the most critical problems below. Importantly, the conclusion reported by the replication team, on their web site and in *Science* magazine, is incorrect: the authors did not fail to replicate our finding. Rather, the replication test was invalid.

The goal of the original study was to test whether a previously-observed individual difference in the cognitive control of racial stereotypes reflected a domain specific or domain-general capacity for cognitive control. To test this hypothesis, we used two well-validated and highly-replicated cognitive control tasks—a version of the Eriksen Flanker Task, to assess domain-general control, and Payne's (2001) Weapons Identification Task, to assess stereotype-based control. We first reported support for our hypothesis in Study 1 of Amodio et al. (2008) and then replicated this effect in Study 2; hence, the effect tested by the replication team had already been successfully replicated.

In reviewing the data collected by the replication team, we noted a failure to produce the normally-observed stereotype-based conflict effect in the Weapons Tasks, i.e., a pattern of greater errors on Black-tool trials than Black-gun trials. It is this highly-replicated pattern of error rates that, by design, makes this task a valid index of stereotype-based cognitive conflict and control. Without this basic pattern, the task does not provide a valid index of stereotype-based cognitive control, rendering it unable to test our hypothesis about individual differences in control. This pattern—the lack of an error rate difference

between Black-gun and Black-tool trials, was noted in the replication team report, but its implication for task validity was not noted or discussed.

We are not sure why the stereotype control task was not valid in the replication study. But we noticed two critical anomalies in the replication that might have contributed.

First, the scripts that we provided for running the Weapons Task were used incorrectly. These scripts were designed for use with a CRT monitor running at 100 Hz, but in the replication study they were run on computers using LCD monitors running at 60 Hz. As a result, the prime and target stimuli were each presented 60% longer than in the original study, for 333 ms rather than 200 ms, and the pacing of the overall task was 60% slower. In the world of priming tasks, this is a large discrepancy that could have undermined task validity.

Second, our original study examined the activation and control of racial majority group member's bias toward Blacks, and our sample primarily comprised White subjects (71% White, 25% Asian, 4% Hispanic) and no Black subjects. [It is notable that Study 1 of Amodio et al. (2008), which reported support for the same test, included 98% White subjects.]

In the replication study, the sample subjects were approximately 39% White, 28% Asian, 9% Hispanic, 3% Black; 2% South Asian, 3% Other, and 15% not reported. The important point here is that, despite the goal of this study to examine White's racial biases toward Blacks, the majority of subjects in the replication appeared to be non-White. This was obviously a major deviation from the goals and methods of the original study. This difference in race of the subjects constitutes a second important discrepancy that undermined the validity of the replication attempt.

These two methodological discrepancies—the differences in stimulus timing and the racial/ethnic makeup of the sample—were not described in the Reproducibility Project's report (e.g., in the section *Differences from the Original Study*).

This commentary is not exhaustive, but rather focuses on three critical problems with the replication attempt that, in technically (i.e., psychometrically and methodologically), rendered the replication study invalid.

At the same time, we gratefully acknowledge the work and cooperation of the replication team. Although the study was ultimately invalid, their hard work is nonetheless appreciated.

As a final note, we believe that problems with the study would have been detected had the research been submitted to peer review by external experts—i.e., the same type of peer review that is the standard for empirical reports in our field. We doubt that this report would have passed peer-review, given the relatively obvious problems with the task timing, error rate pattern, and sample demographics. It's unclear why the empirical findings reported in *Science* (i.e., the 100 replication studies) were not subjected to peer review. [According to Reproducibility Project members, reports were submitted to internal review for compliance with formatting requirements, but not to external scientific review.]

It is notable that we, the authors, had the opportunity to review the data earlier, but were delayed due to our own time commitments and our initial difficulties trying to piece together the multiple raw data files provided by the Reproducibility Project. But even if we had discovered the validity problems sooner, it would have been too late—these problems occurred during data collection. We also acknowledge that the original tests of our effect (in Studies 1 and 2 of Amodio et al., 2008) were underpowered, due to small samples, and thus the effects would benefit from valid replication using larger samples.

In conclusion, we report that the replication attempt of our study (Amodio et al., 2008) was invalid and thus should not be included among either the successful or failed replication studies reported by the Open Science Collaboration in *Science* magazine.

References

Amodio, D. M., Devine, P. G., & Harmon-Jones, E. (2008). Individual differences in the regulation of intergroup bias: The role of conflict monitoring and neural signals for control. *Journal of Personality and Social Psychology, 94,* 60-74.

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.