Reply to the Comments of Referee 1 (Dr. Andrew Chang, Federal Reserve Board) on "Quis custodiet ipsos custodes?: Despite evidence to the contrary, the American Economic Review concluded that all was well with its archive"

I thank Dr. Chang for his considered remarks. In his discussion his makes three points, two minor and one major.

(1: minor) "I believe that the main point missing from McCullough's procedure to replicate Glandon (2011) (McCullough Section 3) is how you would acquire the necessary information from Glandon[.]"

I would request it from Glandon, as I discuss on page 11: "If I wanted to replicate Table 1, I should ask Glandon for the following: 1. A list of the nine papers he analyzed, and the procedure by which the nine were selected…. 2. …. 3. … 4…."

(2: minor) "Setting aside Glandon's summary statistics in Table 1, I think that attempting to replicate the nine attempted (unnamed) replications from Glandon's article would not lead to any new insights about data availability policies in general, or of the AER in particular, because the sample design of Glandon is ill equipped to do so."

I shall have to rewrite the introduction to clarify the purpose of the replication, which is not to yield insights about the data availability policy, but to highlight the way in which the Moffit misled his readers on the extent to which the *AER* was publishing reproducible research.

Dr. Chang suggests that I need to mention the fact that Glandon's sample is non-random. He writes: "As Glandon (2011) notes (emphasis mine): "A total of 39 articles (29 percent of the relevant empirical study population of 135) were selected primarily based on the students' interest in the topic." (pg 696)." and therefore the results of the study cannot be generalized. I specifically mention this on page 9: "(Even if a belief in an article's replicability, rather than an article's actual replicability, is the relevant criterion, any conclusions drawn about the sample cannot be extrapolated to the archive in general: the 39 articles were not randomly sampled but instead constitute a convenience sample.)"

(3: major) "As such, deviations of your replication estimates from the authors' estimates could come from anywhere, so most replication attempts that are at least somewhat constrained by replicator effort have to allow for some degree of deviation of replication estimates from published estimates and still qualify as a "success".

I disagree that deviations can come "from anywhere": when data and code are given, deviations can only from from computational sources. That said, I do not mean to imply that results must agree to machine precision, and the above point is absolutely correct. Where do you draw the line between reproducible and unreproducible? I do address this issue tangentially (p. 3, "We can quibble over how many significant digits constitute reproducibility, but in the end the decision is binary...") but incompletely. The important idea, as Dr. Chang notes, is to allow for some degree of deviation and still have a binary result. I shall have to elaborate on this important idea when I rewrite the paper.