Research Questions:

- Is psychological science getting better?
- Is the movement to improve psychological science working?
- Are we running better-powered studies?
- Are we using fewer questionable research practices?
- How have the new reporting standards implemented in 2014 at Psychological Science improved replicability and diminished questionable research practices?
- How do different measures of research replicability relate to each other?
- Does replicability vary by the prestige of the first-author's affiliation (i.e., are first-string researchers doing more replicable research than second-string researchers?)?
- Does more replicable research get cited more?

General Predictions:

The improving science discussion that became louder and more mainstream in 2012 (when Perspectives on Psych Science published a special issue dedicated to the subject) may have helped usher in more replicable and less questionable research practices. Alternatively, this discussion may not have any effect on actual research practices.

Specific Predictions:

1. Studies in 2013 and 2014 will have more statistical power than studies in 2003 and 2004.
2. Indices of research replicability (e.g., R-Index, Test of Insufficient Variance, N-Pact, and p-curve) will be greater in 2013 and 2014 than in 2003 and 2004.
3. The 2014 volume in Psychological Science will score significantly better on measures of research replicability than all other journals, including the 2003, 2004, and 2013 volumes of Psychological Science.
4. The different measures of research replicability will correlate positively with each other, and, they should be able to identify the papers that were retracted after outright fraud (e.g., Stapel, Smeesters, Sanna) or papers that remain in the published record despite likely excessive questionable research practices (e.g., Bem, 2011).
5. Replicability will not vary by the prestige of the university of the researchers.
6. There is too much variability in the reasons why papers get cited for any relationship to emerge between citations and replicability index scores.

Study Inclusion Rules:

- We will not code meta-analyses, simulation studies, or Bayesian analyses for this project and thus will exclude them from all analyses.
- We will code sample sizes and p-values reported for analyses using structural equation modeling, confirmatory factor analysis, exploratory factor analysis, cluster analysis, multilevel modeling, and other advanced statistics. We will NOT try to interpret their effect sizes.

Population of articles:
2266 articles were published in these journals during these years. We are beginning the process of coding a randomized subset of 30% of these articles, which will give us a final sample of approximately 750 articles.

Analysis Plan:
For all studies that meet the above criteria, we will compute:

- Statistical power
- Effect size
- Actual p-value for a given test statistic
- Percentage of significant p-values out of all p-values reported in each study

For all years and journals, we will compute:

- Average statistical power
- R-Index
- p-curves, and the corresponding Chi-Square statistic
- N-Pact (median power)
- Test of Insufficient Variance
- Percentage of p-values that are inappropriately rounded down (e.g., if p = .055, saying that p < .05)

After computing these, we will report means, standard deviations, 95% confidence intervals for the above indices by journal and year.

We will compute Cohen's d to represent the standardized differences between all journals and years.

We will report intercorrelations between all of the above-mentioned scientific integrity indices to look at their convergent validity. To supplement this, we will also look at these indices for retracted papers (e.g., Stapel, Smeesters, Sanna, & Forster) and compare them to the non-retracted, still-in-print corpus of articles that were coded. Retracted papers should score worse on these measures of scientific integrity.

We will then examine whether university prestige relates to scientific integrity indices. To do this, we will use previously-collected third-party perceptions of prestige collected by The Reproducibility Project (i.e., Nosek et al. / Open Science Collaboration, 2015). Prestige was collected on a continuous Likert-type item, so we will look at the simple correlations between the scientific integrity indices and the third-party judgments of university prestige.

We will also report the relationship between citations and the scientific integrity indices. There is likely too much noise to find a relationship here, but if there is any signal whatsoever, this could suggest that more replicable research that uses fewer questionable research practices is more impactful than less replicable research that uses more questionable research practices.