Scientific prognostication:

The challenges of pre-registering research studies

Thomas Lindsay
Alicia Hofelich Mohr

University of Minnesota
Driven to Discover℠

Pre-registration

- Register hypotheses before data collection begins
- Pre-registration is (usually) publicly available
- Address “HARKing”
- Distinguish confirmatory from exploratory analysis

Pre-registration

Pre-registration isn't primarily designed to:

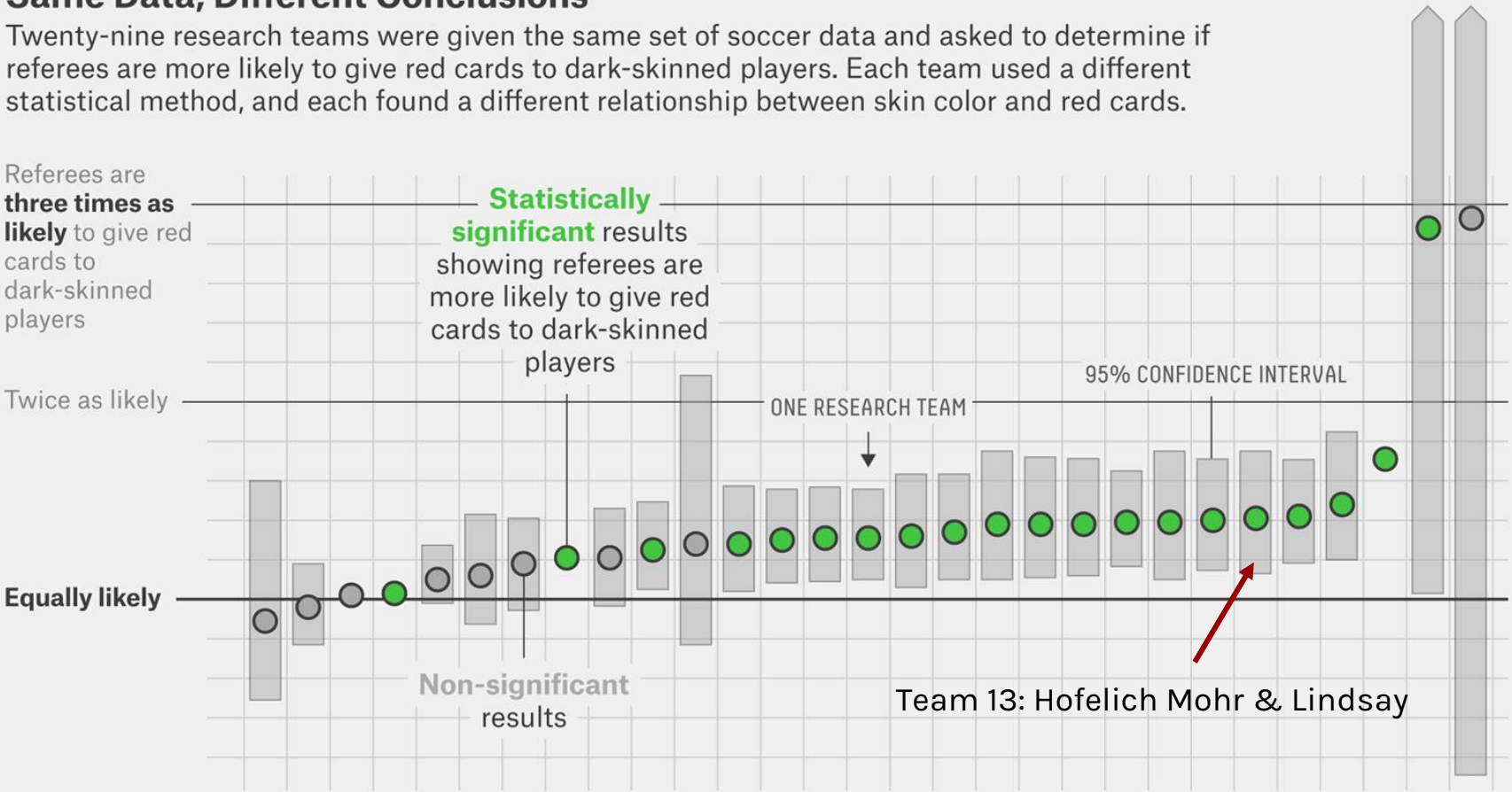
- Prevent blatant dishonesty
- Guarantee replication
- Control quality
- Limit analysis

Pre-Registered Analysis Plan

- Prevent P-hacking
- But how specific does it need to be?

Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.



FIVETHIRTYEIGHT

In press at AMPPS https://psyarxiv.com/qkwst/

SOURCE: BRIAN NOSEK ET AL.

Sources of Variance in analysis

- Covariates
- Outliers
- Correct model specification

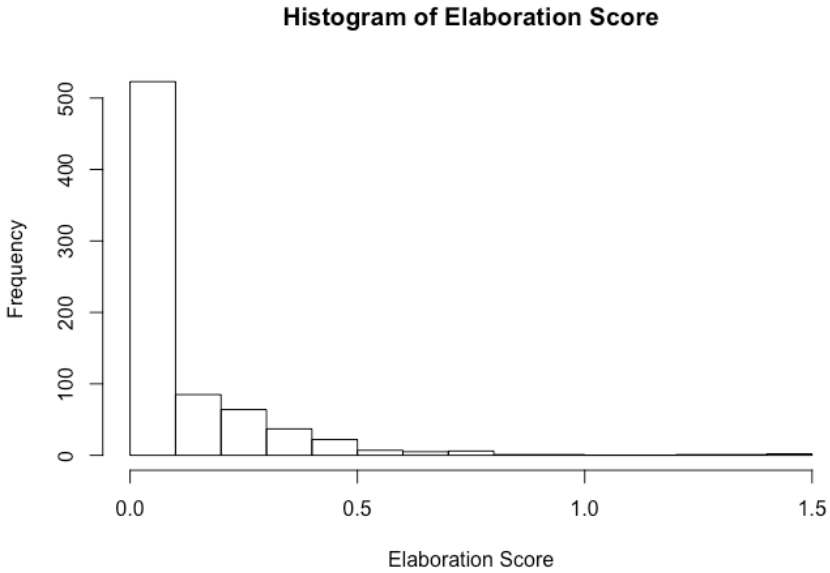
Team	Analytic Approach	N covariates	Treatment of Non-Independence	Distribution
10	Multilevel regression and logistic regression	3	Variance component	Linear
1	Ordinary least squares with robust standard errors, logistic regression	7	Clustered SE	Linear
4	Spearman correlation	3	None	Linear
14	Weighted least squares regression with referee fixed-effects and clustered SE	6	Clustered SE	Linear
11	Multiple linear regression	4	None	Linear
6	Linear Probability Model	6	Clustered SE	Linear
17	Bayesian logistic regression	2	Variance component	Logistic
15	Hierarchical log-linear modeling	1	None	Logistic
31	Logistic regression	6	Clustered SE	Logistic
30	Clustered robust binomial logistic regression	3	Clustered SE	Logistic
3	Multilevel Binomial Logistic Regression using Bayesian inference	2	Variance component	Logistic
23	Mixed model logistic regression	2	Variance component	Logistic
2	Linear probability model, logistic regression	6	Clustered SE	Logistic
5	Generalized linear mixed models	0	Variance component	Logistic
24	Multilevel logistic regression	3	Variance component	Logistic
28	Mixed effects logistic regression	2	Variance component	Logistic
32	Generalized linear models for binary data	1	Clustered SE	Logistic
8	Negative binomial regression with a log link analysis	0	None	Logistic
25	Multilevel logistic binomial regression	4	Variance component	Logistic
9	Generalized linear mixed effects models with a logit link function	2	Variance component	Logistic
7	Dirichlet process Bayesian clustering	0	None	Miscellaneous
21	Tobit regression	4	Clustered SE	Miscellaneous
12	Zero-inflated Poisson regression	2	Fixed effect	Poisson
26	Three-level hierarchical generalized linear modeling with Poisson sampling	6	Variance component	Poisson
16	Hierarchical Poisson Regression	2	Variance component	Poisson
20	Cross-classified multilevel negative binomial model	1	Variance component	Poisson
13	Poisson Multi-level modeling	1	Variance component	Poisson
27	Poisson regression	1	None	Poisson
32	Generalized linear models for binary data	1	Clustered SE	Logistic

Our project

- Replication and extension of our published paper:
    - Hofelich Mohr, A., Sell, A., & Lindsay, T. (2016). Thinking inside the box: Visual design of the response box affects creative divergent thinking in an online survey. Social Science Computer Review, 34(3), 347-359. https://doi.org/10.1177/0894439315588736
- New data collection, similar but different hypotheses
- Pre-registered before data collection (https://osf.io/mbrs8/)
- Data collected, time for analysis...

Pre-registration dilemma

- No mention of what to do if we found skew… and we did.
- Correction changed results
- Not correcting was statistically irresponsible



Is this a common oversight?

Is it just us? Or are pre-registrations failing to capture important analytical choices?

- Examined analytical specificity in other pre-registrations
- Open Science Framework (OSF)'s Pre-Registration Challenge

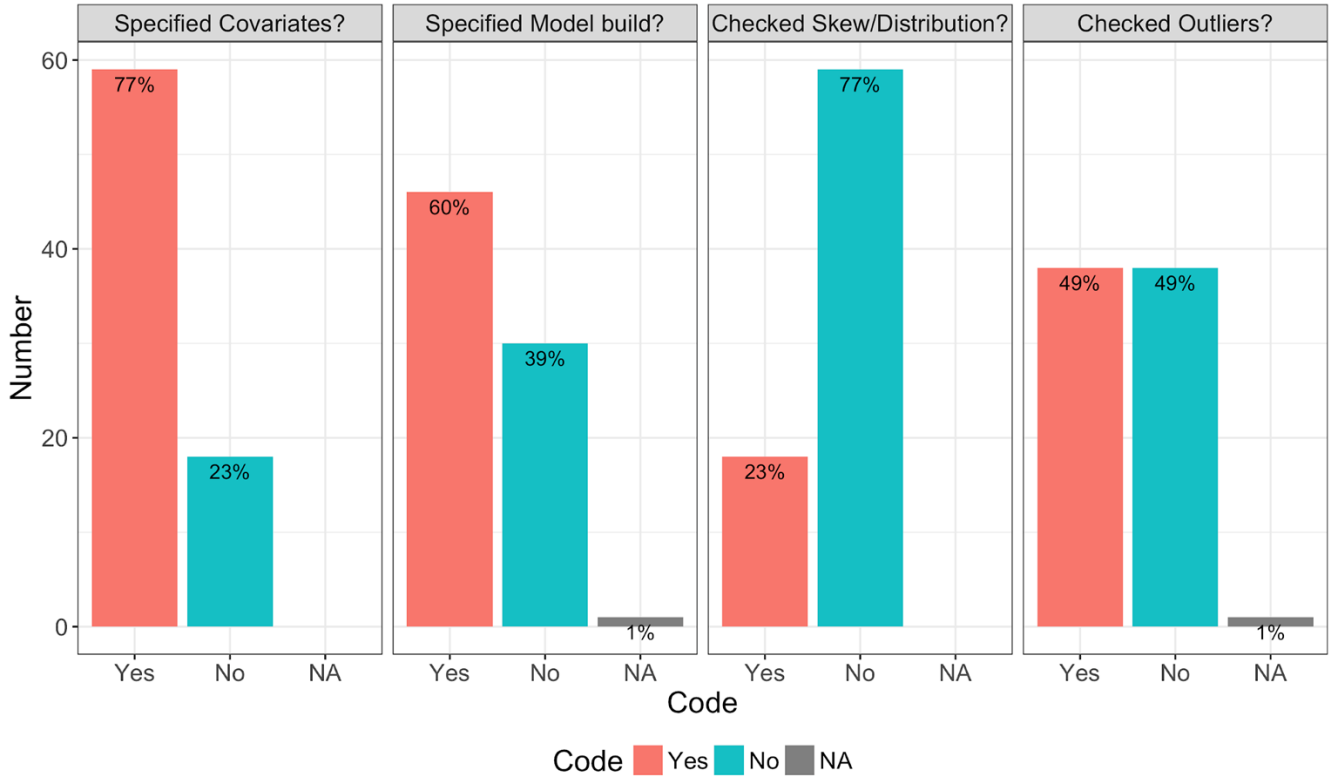
Pre-reg Methods

- OSF API pull of all Pre-Registration Challenge entries
  - Dec 18, 2015 - April 29, 2018
  - N = 2434
- Coded a subset of Pre-registration Analysis Plan sections:
  - Challenge Completers: All pre-registration winners as of Jan 2018 (N=37)
  - Others: Random selection from rest of population (N=40)
- Yes/No coding of inclusion of covariates, model specification, skew/distribution checks, and outliers

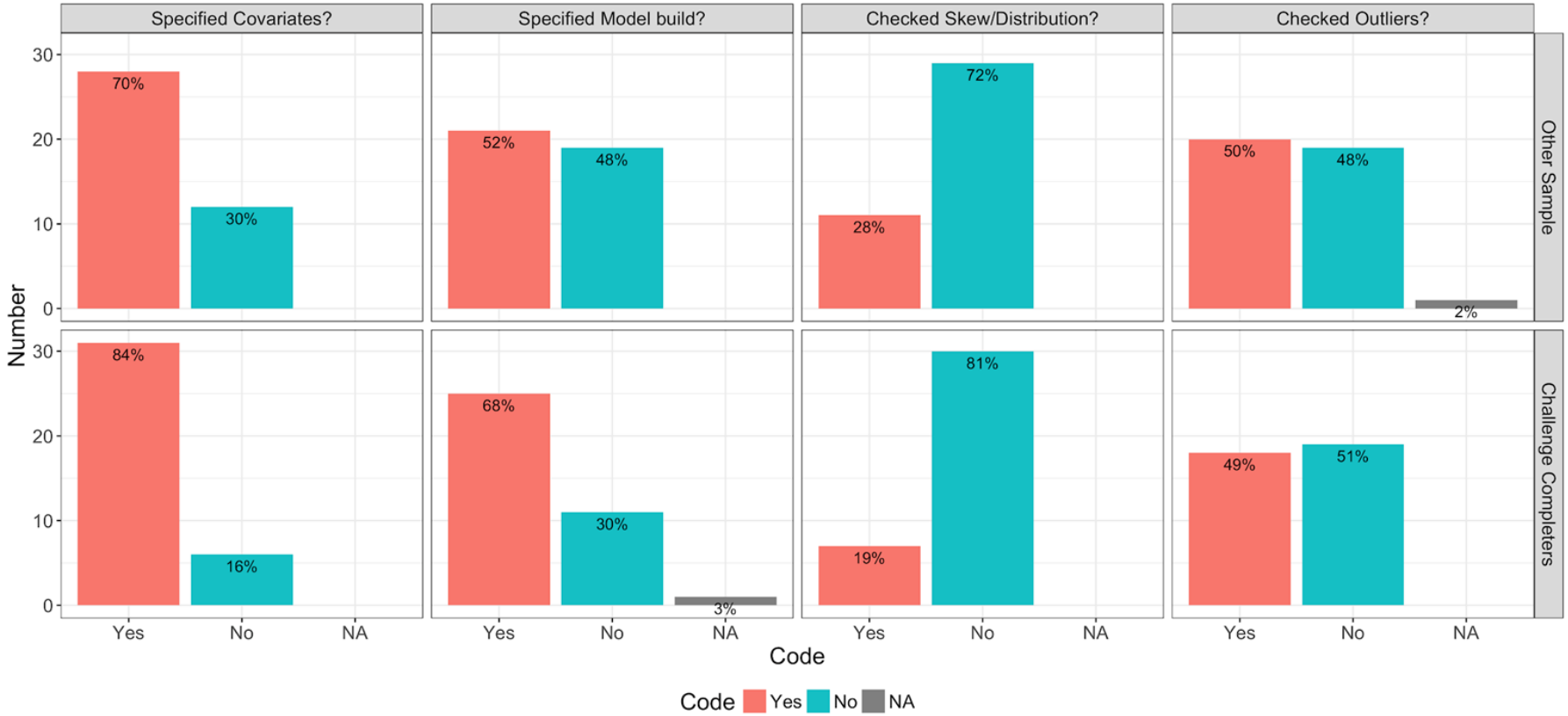
Pre-Reg results

In analysis plans:

- Majority specified covariates and model build
- Few checked variable distributions
- Outliers mentioned inconsistently



Similar for Completers and Others



What's the best strategy?

- Be as specific as possible
  - For challenge completers, absent covariates WERE mentioned in article (n=6)
- Analysis decisions can contribute "degrees of freedom"
  - Think through up front
  - Conditional analysis can leave flexibility
- Learning process
  - How to publish in a tidy manner?

Thank you!
Questions?

lindsayt@umn.edu
hofelich@umn.edu