ReproducibiliTea Frankfurt

Session 2 – 22.11.2023

Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting.

Siepe*, Bartoš*, Morris, Boulesteix, Heck & Pawel* (2023)

*equal contributions

Simulation studies are experiments

„Simulation studies are experiments and should be treated as such by authors and editors" - Hauck & Anderson (1984)

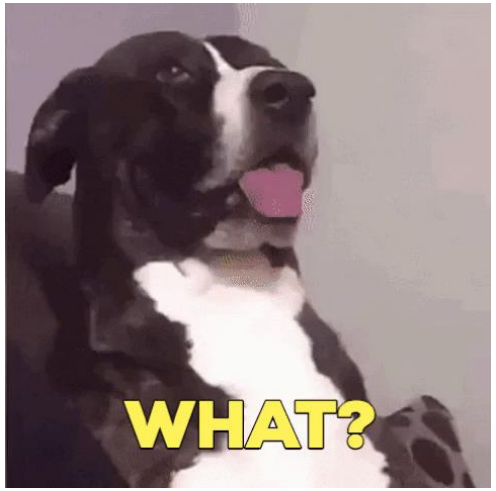
- Very important tool for methodological research
- Needed for evaluation most modern methods (proofs often not feasible)
- Can be very influential

Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives

L **Hu**, PM **Bentler** - Structural equation modeling: a …, 1999 - Taylor & Francis

This article examines the adequacy of the "rules of thumb" conventional cutoff criteria and several new alternatives for various fit indexes used to evaluate model fit in practice. Using a 2-…

☆ Save 99 Cite Cited by 109874 Related articles All 9 versions »



GIF source: Giphy

Preaching water, drinking wine

"Statisticians … often pay too little attention to their own principles of design, and they compound the error by rarely analyzing the results of experiments in statistical theory" - Hoaglin & Andrews (1975)

Problems in the literature:

- Huge researchers' degrees of freedom (Pawel et al., 2023)
- Over-optimism in your own methods (Boulesteix, 2015)
- Suboptimal reporting standards in various statistical fields (see Siepe et al., 2023)
- Barely any assessment of computational reproducibility so far (Luijken et al., 2023)
- Uncertainty? Sample size? Who cares about that? (Koehler et al., 2009 [only *slightly* paraphrased])

GIF source: Giphy

Summary of the ADEMP Planning and Reporting Structure for Simulation Studies.

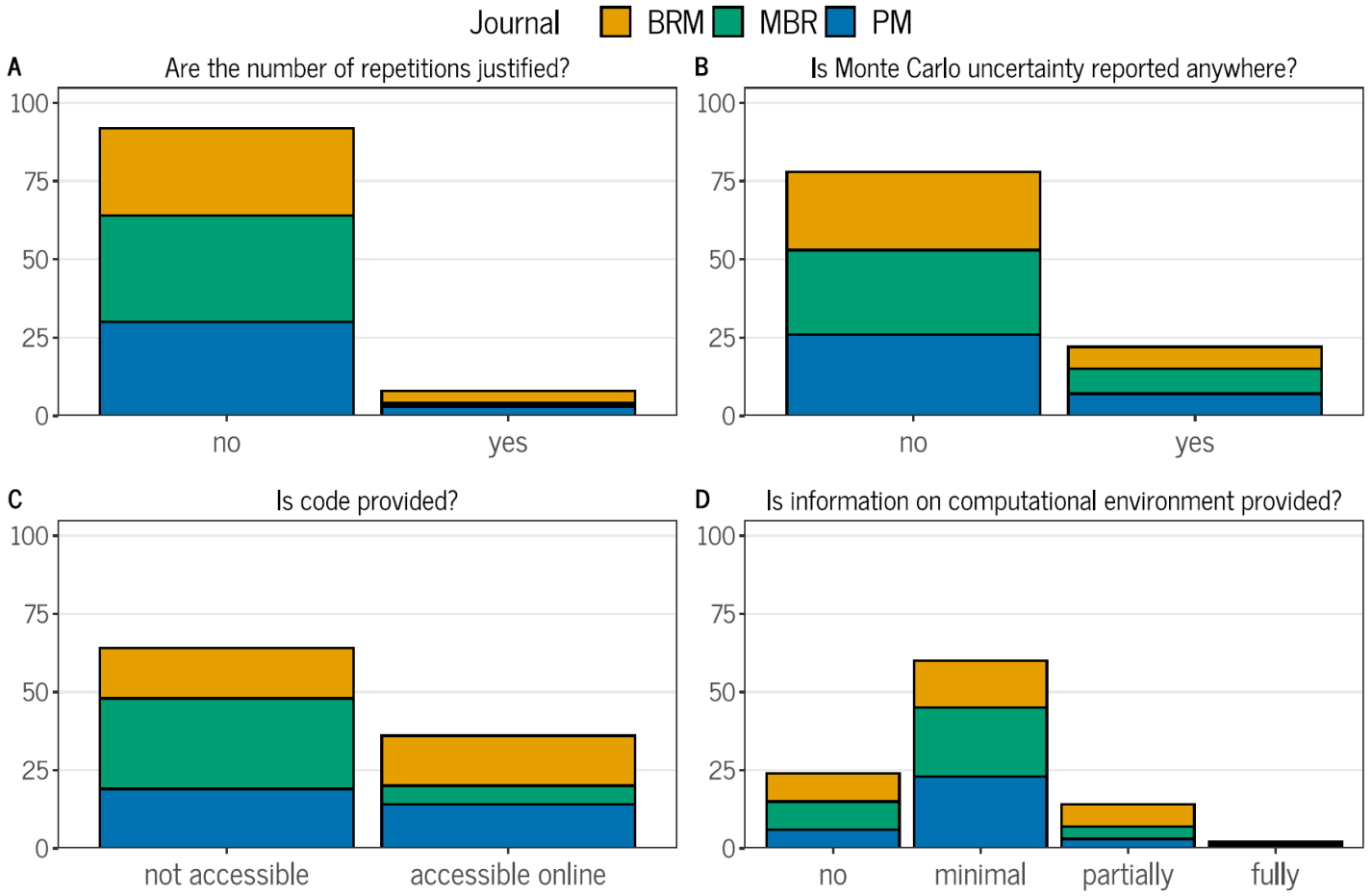
Step	Explanation	Example
Aims	What is the aim of the study?	To evaluate the hypothesis testing and estimation characteristics of different methods for analyzing pre–post measurements in terms of efficiency and robustness
Data-generating mechanism	How are data sets generated?	Pre–post measurements are simulated from a bivariate normal distribution for two groups, with varying treatment effects and pre–post correlations
Estimands and other targets	What are the estimands and/or other targets of the study?	The null hypothesis of no effect between groups is the primary target, the treatment effect is the secondary estimand of interest
Methods	Which methods are evaluated?	ANCOVA, change-score analysis, and post-score analysis
Performance measures	Which performance measures are used?	Type I error rate, power, and bias

Siepe et al. (2023), adapted from Morris et al. (2019)

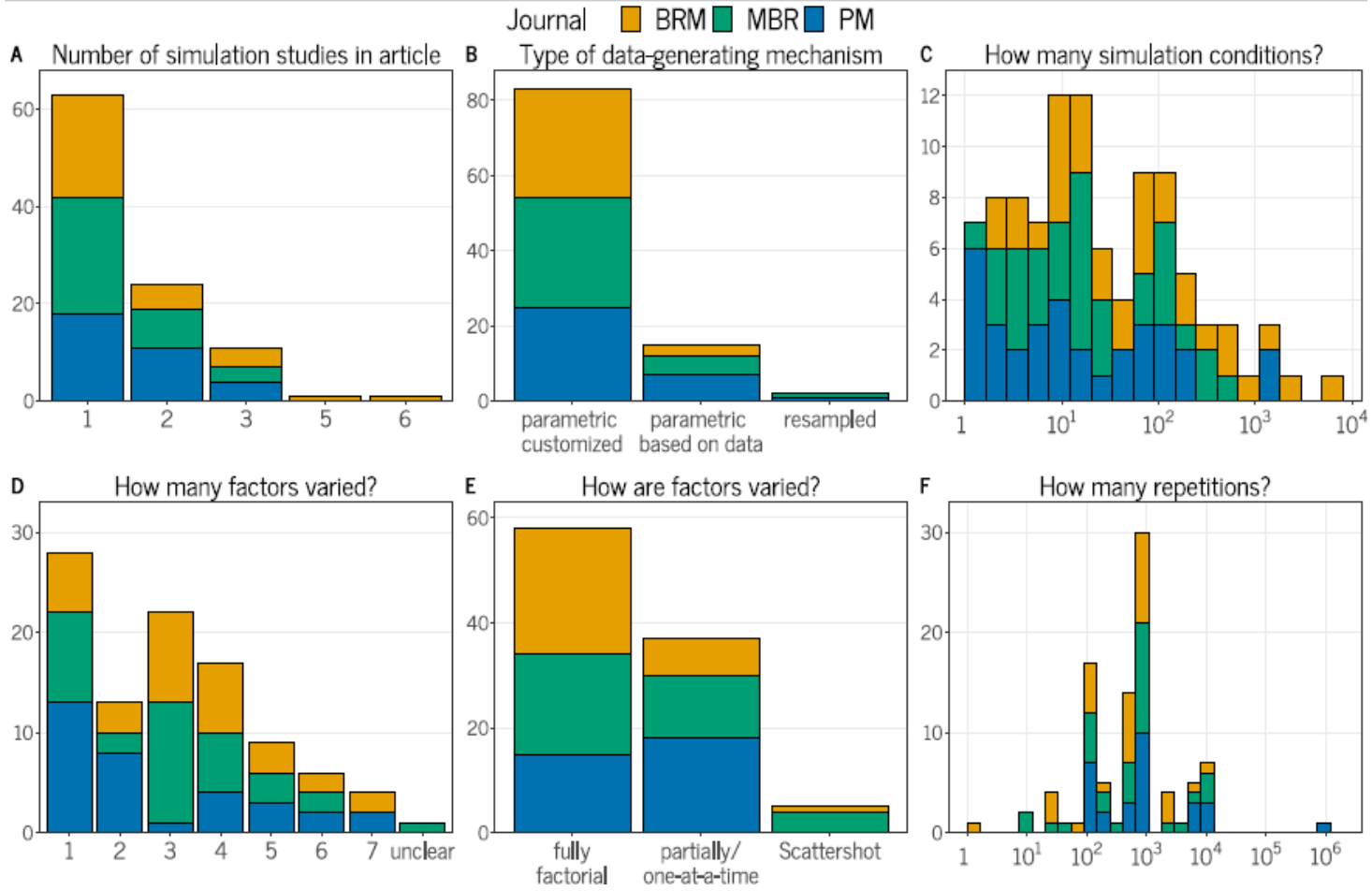
Surveying the psychological literature

- Survey of 321 papers in *Psychological Methods (PM)*, *Behavior Research Methods (BRM)*, *Multivariate Behavioral Research (MBR)*
- 100 contained a simulation study → coded different questions about reporting

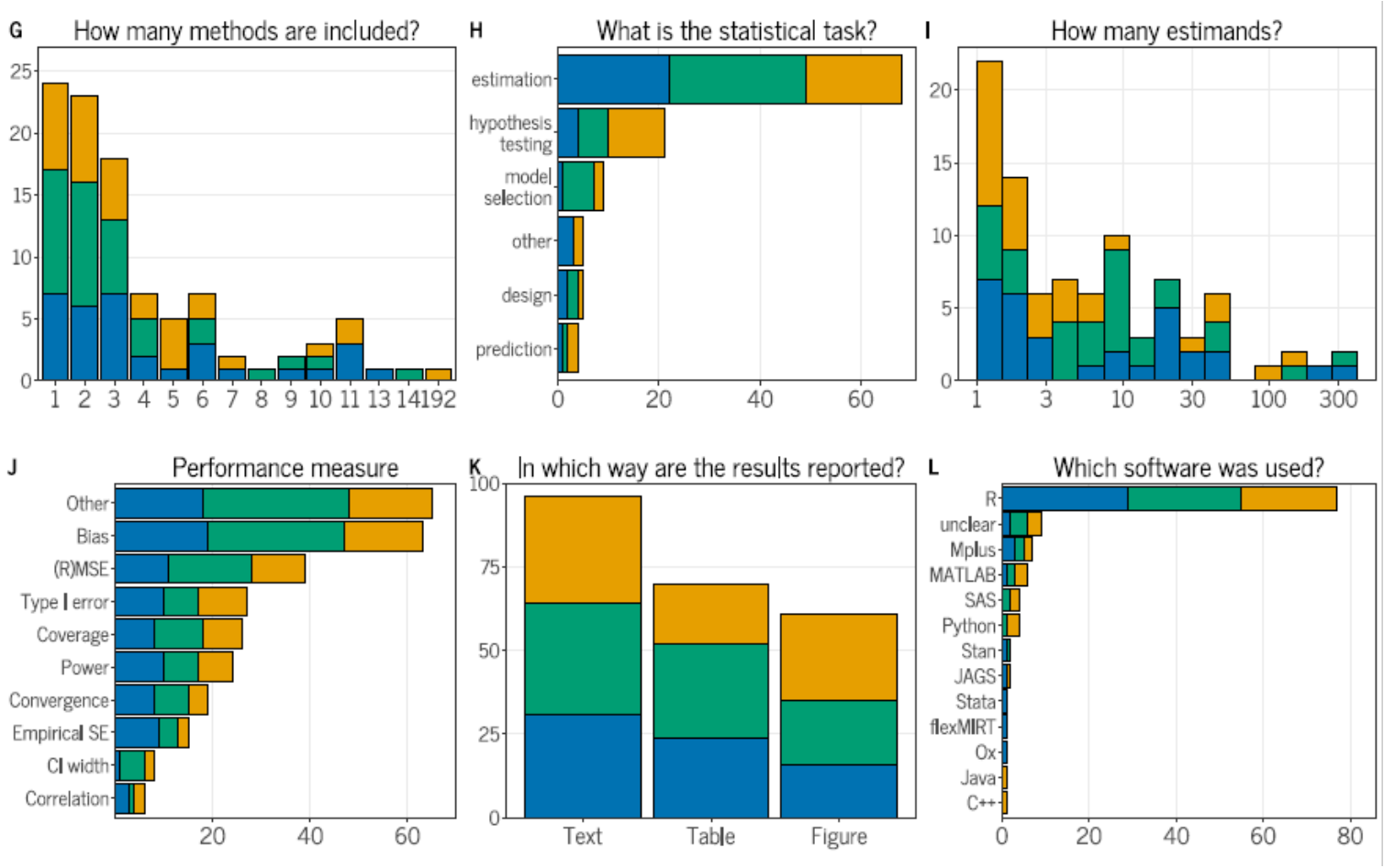
What are people doing in psychology?



What are people doing in psychology?



What are people doing in psychology?



Our proposal: ADEMP-PreReg

Example

Our primary performance measures are the type I error rate (in conditions where the true effect is zero) and the power (in conditions where the true effect is non-zero) to reject the null hypothesis of no difference between the control and treatment condition. The null hypothesis is rejected if the *p*-value for the null hypothesis of no effect is less than or equal to the conventional threshold of 0.05. The rejection rate (the type I error rate or the power, depending on the data generating mechanism) is estimated by

$$\widehat{\text{RRate}} = \frac{\sum_{i=1}^{n_{\text{sim}}} 1(p_i \leq 0.05)}{n_{\text{sim}}}$$

https://github.com/bsiepe/ADEMP-PreReg

Our proposal: ADEMP-PreReg

Performance measure	Definition	Estimate	MCSE	n_{sim}
Bias	$\mathrm{E}(\hat{\theta}) - \theta$	$(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta$	$\sqrt{S^2_{\hat{\theta}} / n_{\text{sim}}}$	$S^2_{\hat{\theta}} / \mathrm{MCSE}^2_*$
Relative bias	$\{\mathrm{E}(\hat{\theta}) - \theta\}/\theta$	$\{(\sum_{i=1}^{n_{\text{sim}}} \hat{\theta}_i / n_{\text{sim}}) - \theta\}/\theta$	$\sqrt{S^2_{\hat{\theta}} / (\theta^2 n_{\text{sim}})}$	$S^2_{\hat{\theta}} / (\mathrm{MCSE}^2_* \, \theta^2)$
Mean square error (MSE)	$\mathrm{E}\{(\hat{\theta} - \theta)^2\}$	$\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}$	$\sqrt{S^2_{(\hat{\theta}-\theta)^2} / n_{\text{sim}}}$	$S^2_{(\hat{\theta}-\theta)^2} / \mathrm{MCSE}^2_*$
Root mean square error (RMSE)	$\sqrt{\mathrm{E}\{(\hat{\theta} - \theta)^2\}}$	$\sqrt{\sum_{i=1}^{n_{\text{sim}}} (\hat{\theta}_i - \theta)^2 / n_{\text{sim}}}$	$\sqrt{S^2_{(\hat{\theta}-\theta)^2} / (4 n_{\text{sim}} \widehat{\mathrm{MSE}})}$	$S^2_{(\hat{\theta}-\theta)^2} / (4 \widehat{\mathrm{MSE}} \, \mathrm{MCSE}^2_*)$
Empirical variance	$\mathrm{Var}(\hat{\theta})$	$S^2_{\hat{\theta}}$	$S^2_{\hat{\theta}} \sqrt{2/(n_{\text{sim}} - 1)}$	$1 + 2(S^2_{\hat{\theta}})^2 / \mathrm{MCSE}^2_*$
Empirical standard error	$\sqrt{\mathrm{Var}(\hat{\theta})}$	$\sqrt{S^2_{\hat{\theta}}}$	$\sqrt{S^2_{\hat{\theta}} / \{2(n_{\text{sim}} - 1)\}}$	$1 + S^2_{\hat{\theta}} / (2\mathrm{MCSE}^2_*)$
Coverage	$\Pr(\text{CI includes } \theta)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{CI}_i \text{ includes } \theta) / n_{\text{sim}}$	$\sqrt{\widehat{\mathrm{Cov}}(1 - \widehat{\mathrm{Cov}}) / n_{\text{sim}}}$	$\widehat{\mathrm{Cov}}(1 - \widehat{\mathrm{Cov}}) / \mathrm{MCSE}^2_*$
Power (or type I error rate)	$\Pr(\text{Test rejects } H_0)$	$\sum_{i=1}^{n_{\text{sim}}} \mathbb{1}(\text{Test}_i \text{ rejects } H_0) / n_{\text{sim}}$	$\sqrt{\widehat{\mathrm{Pow}}(1 - \widehat{\mathrm{Pow}}) / n_{\text{sim}}}$	$\widehat{\mathrm{Pow}}(1 - \widehat{\mathrm{Pow}}) / \mathrm{MCSE}^2_*$
Mean CI width	$\mathrm{E}(\mathrm{CI}_{\text{upper}} - \mathrm{CI}_{\text{lower}})$	$\sum_{i=1}^{n_{\text{sim}}} (\mathrm{CI}_{i,\text{upper}} - \mathrm{CI}_{i,\text{lower}}) / n_{\text{sim}}$	$\sqrt{S^2_W / n_{\text{sim}}}$	$S^2_W / \mathrm{MCSE}^2_*$
Mean of generic statistic G	$\mathrm{E}(G)$	$\sum_{i=1}^{n_{\text{sim}}} G_i / n_{\text{sim}}$	$\sqrt{S^2_G / n_{\text{sim}}}$	$S^2_G / \mathrm{MCSE}^2_*$

Note. Table adapted from Table 6 in Morris et al. (2019)

https://github.com/bsiepe/ADEMP-PreReg

Recommendation

1. Provide a rationale for all relevant choices in design and analysis (e.g., justifications for data-generating mechanism conditions and analysis methods)
2. Use a standardized structure for planning and reporting of simulation studies (e.g., ADEMP)
3. Report Monte Carlo uncertainty (e.g., Monte Carlo standard errors)
4. Choose the number of simulation repetitions to achieve desired precision
5. Write (and possibly preregister) study protocol to guide simulation design and to disclose the state of knowledge, prior expectations, and evaluation criteria before seeing the results (e.g., using the ADEMP-PreReg template)
6. Avoid selective reporting of results that lead to desired outcomes
7. Acknowledge the limited generalizability of a single simulation study
8. Report software versions and environment (e.g., using `sessionInfo()` in R)
9. Upload code, data, results, and other supplements to a FAIR research data repository (e.g., OSF or Zenodo)
10. Journals/Editors/Reviewers: Promote higher reporting standards and open code/data

Question Time!

1. Did you ever conduct (some form of) a simulation study? How was your experience?
2. Is you daily work impacted by simulation studies?
3. Do you think preregistration of simulation studies can work? Not clear where data collection starts.
4. Should we spend more time replicating simulation studies?
5. Should journals have computational reproducibility checks? Is that too much too ask?

References

Boulesteix, A.-L. (2015). Ten simple rules for reducing overoptimistic reporting in methodological computational research. PLoS Computational Biology, 11 (4), e1004191. doi: 10.1371/journal.pcbi.1004191

Hauck, W. W., & Anderson, S. (1984). A survey regarding the reporting of simulation studies. *The American Statistician*, *38*(3), 214-216.

Hoaglin, D. C., & Andrews, D. F. (1975). The reporting of computation-based results in statistics. *The American Statistician*, *29*(3), 122-126.

Luijken, K., Lohmann, A., Alter, U., Gonzalez, J. C., Clouth, F. J., Fossum, J. L., ... & Groenwold, R. H. H. (2023). Replicability of Simulation Studies for the Investigation of Statistical Methods: The RepliSims Project. *arXiv preprint arXiv:2307.02052*.

Koehler, E., Brown, E., & Haneuse, S. J. P. (2009). On the assessment of Monte Carlo error in simulation-based statistical analyses. *The American Statistician*, *63*(2), 155-162.

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in medicine*, *38*(11), 2074-2102.

Pawel, S., Kook, L., & Reeve, K. (2023). Pitfalls and potentials in simulation studies: Questionable research practices in comparative simulation studies allow for spurious claims of superiority of any method. *Biometrical Journal*, 00, e2200091. https://doi.org/10.1002/bimj.202200091

Siepe, B. S., Bartoš, F., Morris, T. P., Boulesteix, A., Heck, D. W., & Pawel, S. (2023, October 31). Simulation Studies for Methodological Research in Psychology: A Standardized Template for Planning, Preregistration, and Reporting. *PsyArXiV*. https://doi.org/10.31234/osf.io/ufgy6

Resources

People reading image: https://c8.alamy.com/comp/2E2T78N/three-generation-of-surprised-women-reading-newspapers-isolated-on-white-2E2T78N.jpg
What GIF: https://i.giphy.com/media/3o7527pa7qs9kCG78A/giphy.gif
Sassy GIF: https://giphy.com/gifs/reactionseditor-sassy-sass-l0IymiszgmwwfB5K0