The Role Cross-Validation Can Play in Resolving the Replication Crisis in Psychology

Nalan Akın

Department of Psychology, Koç University

CSSM 502: Advanced Data Analysis in Python

David Carlson

June 3, 2023

Introduction

The replication crisis has become a widely discussed problem among researchers, especially after a large-scale study by Open Science Collaboration was published in 2015. The collaborators reported that only 36 out of 100 correlational and experimental studies published in psychology journals were replicated (i.e., had statistically significant results). The proposed reasons for this crisis are not limited to only one stage of research. From the operationalization of concepts to publishing a paper, related questionable practices such as HARKing, lack of transparency, and selective reporting are widespread according to surveys (Rubin, 2017; Baker, 2016). These are examples of using researcher degrees of freedom to hypothesize, analyze and report in ways that lead to p-hacking (Wicherts et al., 2016).

Yarkoni and Westfall (2017) bring a new perspective to this problem by pointing to the adoption of an almost purely explanatory approach that makes replication difficult. How they define replication crisis from their perspective is a failure of models, that were built to explain behavior, to predict the same behavior in a future sample. Their central thesis in this paper is that psychology can take advantage of the power of machine learning principles to become a more predictive science, which can also contribute to the performance and efficiency of explanatory models. The availability of big data and high-performance computing has created expectations of new research areas and applications in social sciences. In psychology, researchers have attempted to use machine learning algorithms to predict human behavior, although its relative performance to more traditional methods is sometimes found unimpressive in clinical settings (Christodoulou et al., 2019) and suicide research (Jacobucci et al., 2021). However, what is introduced in Yarkoni and Westfall (2017) is a more minimal but supposedly beneficial change through the cross-validation technique.

Yarkoni draws attention to model evaluation practices in psychology in both this paper with Westfall and Rocca and Yarkoni (2021), where they promote the adoption of

benchmarking practices in psychology. While machine learning has two phases of training and testing a model, psychology has only a training phase that corresponds to estimating the parameters of a model. This means the model's performance on an out-of-sample test is not evaluated. The R^2 metric is misused as an indicator of the predictive power of a model because it does not show the variance that a regression equation with specific intercept, beta, and error values can explain as assumed. It shows the variance that can be explained by different intercept, beta, and error values found with different samples. Therefore, it is overly optimistic for the regression equation made up of the specific values our model estimated (Yarkoni & Westfall, 2017). These specific values are the ones that will best fit the sample in which the model is being trained. Due to sampling error, the values that best fit each sample will be different. This refers to one of the key concepts explained in their article, *overfitting*. It occurs due to errors in separating noise from signal. The attention overfitting takes is at quite different levels within machine learning and psychology. While fit indices to evaluate the predictive capacity of machine learning models, which is more important for machine learning, are results of the testing phase, in psychology, what is reported is obtained from the training phase only. Thus, it is ignored that the test error is always higher than the training error. The difference between them does not always require action, but we should know which conditions increase the risk of overfitting. Yarkoni and Westfall (2017) clarify that "when predictors have strong effects and researchers fit relatively compact models in large samples, overfitting is negligible (p. 5)".

The problems related to overfitting or replication are not restricted to model evaluation practices. Whether decisions made until this stage involve a tendency to p-hacking is a critical indicator of the risks. P-hacking is done by exploiting degrees of freedom to distort significance, leading to false positives and inflation of effect sizes. Lakens (2015) contributed to the understanding of p-hacking by showing a better way to examine the distribution of p-

values by considering publication bias, power, and Type-1 errors, concluding that there are many studies with a significance level just below 0.5. Yarkoni and Westfall (2017) explain the connection between p-hacking practices such as optional stopping and the increased risk of overfitting. This is caused by making decisions compatible with the expected results or hypotheses in mind. Flexibility is used in psychology research to reduce bias in the model as much as possible. In this way, parameter estimates approach their true values. However, this increases the risk of overfitting because while the error is thought to consist only of bias here, a part of it is actually related to variance. Variance increases when bias decreases since the model becomes more complex, elevating the risk of overfitting. The high variance should be prevented for predictive models that aim to reduce overfitting because high variance makes the model unstable and sensitive to change (Yarkoni & Westfall, 2017). Therefore, their priority is to reduce the overall predictive error composed of bias and variance. A high variance may also reduce the generalizability and robustness of the results in explanatory studies.

How can we prevent overfitting by evaluating the performance of our model based on the prediction error instead of deciding only according to bias? Yarkoni and Westfall (2017) recommend using a machine-learning tool. They propose that cross-validation can be a close alternative to replication without any effort or resources to collect new data. K-fold cross-validation avoids the loss of statistical power by repeating cross-validation several times and using each data point in both training and testing data roles in different folds. In this way, researchers can test the generalizability of their model or previously published models and detect overfitting. The authors suggest that especially applied psychology research that aims to make accurate predictions of behaviors of employees, students, patients, etc., should adopt this practice of reporting cross-validation results. Moreover, they are not the only or first ones to offer cross-validation to improve reliability. Koul et al. (2018) introduced a new type

termed *simulated replication* with cross-validation for kinds of psychology studies that are inapplicable or hard to replicate, such as clinical-epidemiological studies. They provide a list of packages containing cross-validation, but one problem is that they are not the most widely used ones in psychology and generally require coding. PRoNTO (Schrouff et al., 2013) and PredPsych (Koul et al., 2017) are available software that does not require as much programming as others. Although holdout cross-validation was defined as "not worth the effort or the loss of degrees of freedom " by Murphy (1983) at the time, with newer techniques like Leave-One-Out Cross-Validation, it is easy to avoid loss of degrees of freedom and extra effort is minimal with new programs. Still, integrating it into SPSS and R statistics packages can make it easier to use. Murphy (1983) has other criticisms that can still be valid. He argues that violations of random sampling assumptions are not detectable through single-sample cross-validation done through partitioning a sample instead of collecting two independent samples from the same population. He could disapprove of new techniques due to inescapable sampling error regardless of the number of subsamples.

Present Study

Although Yarkoni and Westfall (2017) is an influential paper cited more than 1300 times, I have not encountered a research article that applies cross-validation in my readings yet. There have been papers published in psychology journals that promote cross-validation for several purposes, such as model selection (de Rooij & Weeda, 2020) and testing generalizability (Song et al., 2021), following a discussion of problems central to this paper. These efforts inspired me to test K-fold cross-validation with previously published research since this practice is at the intersection of my willingness to learn new tools beyond traditional methods and my interest in the issues regarding psychology research. Since both the tension and integration possibilities of explanatory and predictive approaches are ongoing debates (Hofman et al., 2021; Hullman et al., 2022), I decided to apply this technique to both research

goals. To that end, I searched the literature to find one paper that aims to predict an outcome and one that aims to identify underlying mechanism of behavior.

Method

Datasets

My search criteria were relevance to social psychology, open data availability, and data analysis with Regression or ANOVA models. For the first article, I looked for the mention of a goal to predict an outcome in the title or abstract. For the second one, an experimental study that features manipulation of a condition was needed. I restricted the source to the journal PLOS ONE due to their data availability principle. Trying to find data gave me a chance to observe the current situation of data sharing in psychology. It has not been a common practice despite the encouragement of the Center for Open Science and journals such as PLOS and The Royal Society (Houtkoop et al., 2018). This was also the reason for ending up with relatively new studies (published in 2022 and 2015) because the older data gets, the more challenging to find a published dataset.

I decided to try K-fold cross-validation for the findings of the study titled "The role of performance pressure, loneliness and sense of belonging in *predicting* burnout symptoms in students in higher education" (Dopmeijer et al., 2022). The data was taken from a large cross-sectional study that collected data from university students in the Netherlands. The sample size was 3141. 60% were composed of women, the mean age was 21.8, and 88.4 of the students were Dutch. The independent variables were performance pressure, loneliness, and sense of belonging. In addition, gender, age, study year, and living condition (alone or with parents) were covariates. They used scales for all variables except performance pressure, measured with 1 question. Burnout, the dependent variable, was measured with UBOS-A (Schaufeli et al., 2000), which had three subscales: Emotional exhaustion, depersonalization,

and a reduced sense of personal accomplishment. Analysis was conducted on SPSS and MPlus. The assumptions of linearity, multicollinearity, homoscedasticity, and normality were all confirmed.

The other chosen study was “Unskilled and Don't Want to Be Aware of It: The Effect of Self-Relevance on the Unskilled and Unaware Phenomenon” (Kim et al., 2015). Poor performers overestimate their success, which was previously linked to their low metacognitive ability to accurately evaluate their performance (Kruger & Dunning, 1999). They tested self-relevance as an alternative explanation to the Dunning-Kruger effect and hypothesized that stronger self-enhancement motivation leads to higher positive response bias when the task is relevant to the self. During the experiment, male and female university students (N= 143) were assigned to the same visual pun task, but it was introduced as a test of visual aptitude or language ability. In language ability condition, researchers emphasized the association of this ability with women to manipulate female students’ perception of self-relevance. In the end, they estimated how well they performed compared to others. Their actual performance in the task was measured before assignment to the conditions to rank them. Their factorial design was 2 (Relevance Condition) X 2 (Actual Performance) X 2 (Gender). The study was conducted in the United States, 86 participants were female, and the mean age was 20.09.

Plan of Analysis

Since the goal is replication, I needed to follow the same steps as the researchers. For the study predicting burnout (Dopmeijer et al., 2022), I determined which variables they use as categorical because the dataset included both scale and binary versions for most of them. However, only living condition, study year, and gender were added to the model after dummy coding. Researchers tested a different linear regression model with each burnout subscale as the dependent variable. I used OLS regression from the statsmodel package. OLS is a technique commonly used to estimate parameters in simple and multiple regression models.

After getting the result summary, I applied K-fold cross-validation and chose 10 folds with 200 repetitions to prevent randomness, this time following de Rooij & Weeda (2020), who detected default values to use unless the sample is too small (<40) considering bias-variance trade-off.

For the self-relevance effect study, the authors built a General Linear Model, but I reanalyzed it using the Linear Regression technique to obtain the R^2 value. Categorical variables gender and condition were coded as 1 and 2; therefore, 2s were replaced with 0 to transform it into dummy coding. Multicollinearity was prevented since the continuous predictor was centered by subtracting the mean from individual values in both the original study and my replication attempt. To test the 3-way interaction effect, the formula PERP ~ condition + gender + meancent + condition*gender*meancent' was implemented by preprocessing data with the ‘dmatrices’ function from ‘patsy’. The k-fold cross-validation procedure was the same.

Results

Predicting Burnout

The results from the original study can be found in Table 1. In the Jupyter Notebook, OLS summary parameter estimates are not standardized; therefore, they differed from the original study's reported values. I got standardized coefficients for my model by multiplying the ratio of the standard deviations of the independent variables and dependent variable with the unstandardized coefficient. As expected, there were no differences in their significance statuses.

A 10-fold cross-validation resulted in a drop in R^2 from .289 to .283 for unobserved data in emotional exhaustion model, from .365 to .357 in depersonalization model, and from .103 to .093 in personal accomplishment model. When repeated for emotional exhaustion

model 200 times, the mean R^2 was .280 with a standard deviation of .002. Overall, the results show that there is almost no overfitting.

Table 1

		Emotional Exhaustion			Depersonalization			Personal Accomplishment		
	Variable	Standardized β	Standard Error	p-value	Standardized β	Standard Error	p-value	Standardized β	Standard Error	p-value
Predictors	Performance pressure	.29	.02	.000	.10	.02	.000	.00	.02	.817
	Loneliness	.19	.02	.000	.12	.02	.000	-.15	.02	.000
	Sense of belonging	-.28	.02	.000	-.47	.02	.000	.23	.02	.000
Covariates	Gender	-.11	.02	.000	.04	.02	.006	.08	.02	.000
	Age	-.07	.02	.000	-.05	.02	.003	.02	.02	.338
	Living situation	.02	.02	.242	-.02	.02	.299	-.02	.02	.445
	Year of study	.07	.02	.000	.21	.02	.000	-.08	.02	.000
		$R^2 = .288$			$R^2 = .365$			$R^2 = .107$		

https://doi.org/10.1371/journal.pone.0267175.t003

When the sample size was reduced to 100 to experiment with the emotional exhaustion model, the new model R^2 was .381, and the coefficients of predictors were still significant. However, when the same 10-fold cross-validation procedure was repeated, this time, there was a dramatic drop in R^2 to -.05 with variable results for folds. 500 repetitions were added to avoid randomness, and the new R^2 was .20 with a standard deviation of .1.

Explaining Unskilled and Unaware Phenomenon

A 10-fold cross-validation resulted in a drop in R^2 from 0.18 to -0.07 for unobserved data. When repeated 200 times, the mean R^2 was .06 with a standard deviation.05.

Discussion

A better understanding of statistics was the most endorsed item when researchers evaluated a list of approaches to producing replicable studies (Baker, 2016). This paper attempted to try a machine-learning technique as a part of common data analysis procedure in psychology. Although their expectations from statistical models differ from psychology, Yarkoni and Westfall (2017) suggest using this ML tool as a model estimation and selection strategy in psychology after concisely explaining the overfitting issue it can detect.

In the first application, the fit of burnout models with a high sample size (N=3141) did not change after cross-validation. Big sample sizes are required to prevent overfitting,

especially when the number of predictors and model complexity are high. Therefore, big samples should be provided whenever possible despite the observed decrease in effect sizes when sample sizes are increased in psychology research (Ioannidis, 2008). When the sample was decreased to an extremely small size, R^2 was higher than the original version, but cross-validation failed and gave a negative value. Despite small sample size, regression coefficient values were still significant. This example shows the importance of the sample and statistical power against overfitting. It should also be noted that all students were from the same university, which might have caused a homogeneity that reminds Murphy's (1983) criticism of single-sample cross-validation. Although they mention this as one of the limitations of the study, they also note that "there's no reason to assume that the associations found in the present study will differ greatly between universities" (Dopmeijer et al., 2022, p. 8).

In the second application, we can say that the role of self-relevance in explaining the Dunnning-Kruger effect failed to replicate. It might be harder to grasp the value of evaluating an explanatory model based on its predictive power, but this gives us a chance to evaluate its generalizability and replicability. Rocca and Yarkoni (2021) show how "parameter estimates are always conditional on the model itself'. Therefore, although the interest is in the significance of parameter estimates and the magnitudes of their effects, model fit needs to be checked not only within the training phase but by predicting unobserved data. We need to ask ourselves whether all significant interaction effects in Kim et al. (2015) are meaningful after cross-validation uncovered significant overfitting.

Limitations

In this study, cross-validation is applied to only two datasets that belong to independent studies with different topics and goals. A more fruitful way of using cross-validation to reassess past research can be collecting different datasets on the topic of interest (e.g., from studies included in a meta-analysis) with diverse methods, sample sizes, and

characteristics to make conclusions about the reliability of an effect that was investigated in different ways.

Conclusion

In conclusion, despite its simplicity, cross-validation can be a valuable tool for evaluating the exact replicability of findings. Within-lab validation is expected to be a part of the solution to the lack of replicability, and cross-validation can be an important alternative rather than collecting and analyzing new data to assess that. Many researchers are not convinced that conducting “exact” replicability tests is possible since it assumes keeping the materials and conditions identical to the original study. Cross-validation can be the closest way to fulfill this assumption. It needs more recognition in the field to spread its use, explore its limitations for different contexts and goals, and discuss its optimal application.

References

Baker, M. (2016). 1,500 scientists lift the lid on reproducibility. *Nature*, *533*(7604), 452–454. https://doi.org/10.1038/533452a

Christodoulou, E., Ma, J., Collins, G. S., Steyerberg, E. W., Verbakel, J. Y., & Van Calster, B. (2019). A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *Journal of Clinical Epidemiology*, *110*, 12–22. https://doi.org/10.1016/j.jclinepi.2019.02.004

de Rooij, M., & Weeda, W. (2020). Cross-Validation: A Method Every Psychologist Should Know. *Advances in Methods and Practices in Psychological Science*, *3*(2), 248–263. https://doi.org/10.1177/2515245919898466

Dopmeijer, J. M., Schutgens, C. A. E., Kappe, F. R., Gubbels, N., Visscher, T. L. S., Jongen, E. M. M., Bovens, R. H. L. M., Jonge, J. M. de, Bos, A. E. R., & Wiers, R. W. (2022). The role of performance pressure, loneliness and sense of belonging in predicting burnout symptoms in students in higher education. *PLOS ONE*, *17*(12), e0267175. https://doi.org/10.1371/journal.pone.0267175

Hofman, J. M., Watts, D. J., Athey, S., Garip, F., Griffiths, T. L., Kleinberg, J., Margetts, H., Mullainathan, S., Salganik, M. J., Vazire, S., Vespignani, A., & Yarkoni, T. (2021). Integrating explanation and prediction in computational social science. *Nature*, *595*(7866), Article 7866. https://doi.org/10.1038/s41586-021-03659-0

Houtkoop, B. L., Chambers, C., Macleod, M., Bishop, D. V. M., Nichols, T. E., & Wagenmakers, E.-J. (2018). Data Sharing in Psychology: A Survey on Barriers and Preconditions. *Advances in Methods and Practices in Psychological Science*, *1*(1), 70–85. https://doi.org/10.1177/2515245917751886

Hullman, J., Kapoor, S., Nanayakkara, P., Gelman, A., & Narayanan, A. (2022). The Worst of Both Worlds: A Comparative Analysis of Errors in Learning from Data in Psychology and Machine Learning. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 335–348. https://doi.org/10.1145/3514094.3534196

Ioannidis, JPA. (2008). Why Most Discovered True Associations Are Inflated: *Epidemiology, 19*(5), 640–648. 10.1097/EDE.0b013e31818131e.

Jacobucci, R., Littlefield, A. K., Millner, A. J., Kleiman, E. M., & Steinley, D. (2021). Evidence of Inflated Prediction Performance: A Commentary on Machine Learning and Suicide Research. *Clinical Psychological Science*, *9*(1), 129–134. https://doi.org/10.1177/2167702620954216

Kim, Y.-H., Chiu, C.-Y., & Bregant, J. (2015). Unskilled and don't want to be aware of it: The effect of self-relevance on the unskilled and unaware phenomenon. *PLoS ONE*, *10*(6). Scopus. https://doi.org/10.1371/journal.pone.0130309

Koul, A., Becchio, C., & Cavallo, A. (2018). Cross-Validation Approaches for Replicability in Psychology. *Frontiers in Psychology*, *9*. https://www.frontiersin.org/articles/10.3389/fpsyg.2018.01117

Koul, A., Becchio, C., and Cavallo, A. (2017). PredPsych: A toolbox for predictive machine learning based approach in experimental psychology research. *Behavioral Research Methods.* doi: 10.3758/s13428-017-0987-2.

Lakens, D., 2015. Comment: What p-hacking really looks like: A comment on Masicampo and LaLande (2012). *Quarterly Journal of Experimental Psychology*, *68*(4), 829-832.

Murphy, K. R. (1983). Fooling Yourself with Cross-Validation: Single Sample Designs. *Personnel Psychology*, *36*(1), 111–118. https://doi.org/10.1111/j.1744-6570.1983.tb00507.x

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251. https://doi.org/10.1126/science.aac4716

Rocca, R., & Yarkoni, T. (2021). Putting Psychology to the Test: Rethinking Model Evaluation Through Benchmarking and Prediction. *Advances in Methods and Practices in Psychological Science*, *4*(3), 25152459211026864. https://doi.org/10.1177/25152459211026864

Rubin, M. (2017). When does HARKing hurt? Identifying when different types of undisclosed post hoc hypothesizing harm scientific progress. *Review of General Psychology*, *21*, 308–320.

Schaufeli, W., Dierendonck, D van. (2000). Maslach burnout inventory: Nederlandse versie [Maslach burnout inventory: Dutch version]. Swets y Zeitlinger.

Schrouff, J., Rosa, M. J., Rondina, J. M., Marquand, A. F., Chu, C., Ashburner, J., et al. (2013). PRoNTo: Pattern recognition for neuroimaging toolbox. *Neuroinformatics*, *11*, 319–337. doi 10.1007/s12021-0139178-1.

Wicherts, JM., Veldkamp Cl., Augusteijn HE. Et al. (2016) Degrees of freedom in planning, running, analyzing, and reporting psychological studies: A checklist to avoid p-Hacking. *Frontiers in Psychology*, 7, 1-12.

Yarkoni, T., & Westfall, J. (2017). Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. https://doi.org/10.1177/1745691617693393