aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Perspectives on Psychological Science
2020, Vol. 15(3) 809–816

DOI: 10.1177/1745691620902467
www.psychologicalscience.org/PPS
SAGE

Machine Learning and Psychological Research: The Unexplored Effect of Measurement

Ross Jacobucci[1] and Kevin J. Grimm[2]
[1]Department of Psychology, University of Notre Dame, and [2]Department of Psychology, Arizona State University

Abstract
Machine learning (i.e., data mining, artificial intelligence, big data) has been increasingly applied in psychological science. Although some areas of research have benefited tremendously from a new set of statistical tools, most often in the use of biological or genetic variables, the hype has not been substantiated in more traditional areas of research. We argue that this phenomenon results from measurement errors that prevent machine-learning algorithms from accurately modeling nonlinear relationships, if indeed they exist. This shortcoming is showcased across a set of simulated examples, demonstrating that model selection between a machine-learning algorithm and regression depends on the measurement quality, regardless of sample size. We conclude with a set of recommendations and a discussion of ways to better integrate machine learning with statistics as traditionally practiced in psychological science.

Keywords
machine learning, data mining, measurement error, structural-equation modeling, psychometrics

The impact of machine learning (e.g. data mining, big data, artificial intelligence) has been felt across science, with more recent applications in the field of psychology (e.g., Adjerid & Kelley, 2018). Machine learning has led to new ways of collecting data (e.g., from social media) as well as to new ways in which the relationships between variables are modeled. Although machine learning has allowed for the statistical analysis of relationships that were previously not possible (e.g., $p > N$ in genetic and brain research), it has also allowed for the reanalysis of existing data, with hopes that these more flexible algorithms will improve our prediction or understanding of variables of interest.

Although machine learning has prompted improvements in the accuracy of modeling, particularly in areas assessing biological or genetic variables, in which most variables have an objective form of assessment (e.g., Just et al., 2017), the overall promise has not been matched by gains in performance in the behavioral sciences. In psychology specifically, the impact of machine learning has not been commensurate with what one would expect given the complexity of algorithms and their ability to capture nonlinear and interactive effects. For example, in the prediction of suicide (broadly speaking to include suicidal ideation and attempts), some research has noted the need for machine learning to account for the complexity of risk factors (Walsh, Ribeiro, & Franklin, 2017), indeed finding improved predictive performance with machine-learning algorithms (e.g., see Burke, Ammerman, & Jacobucci, 2019). However, many articles do not compare performance across machine-learning and traditional statistical models, making it difficult to assess the utility of algorithms that sacrifice interpretation to some degree and those that do find marginal, at best, improvements. This has been noted in a recent review comparing machine learning with more traditional statistical models such as linear and logistic regression (Jie, Collins, Steyerberg, Verbakel, & van Calster, 2019), which found no evidence of superior performance of the machine-learning algorithms.

The integration of machine-learning algorithms with psychological research requires a great deal of nuance, a point that has been made by a number of researchers

Corresponding Author:
Ross Jacobucci, Department of Psychology, University of Notre Dame, 390 Corbett Family Hall, Notre Dame, IN 46556
E-mail: rjacobuc@nd.edu

(e.g., Adjerid & Kelley, 2018; Chen & Wojcik, 2016; Dwyer, Falkai, & Koutsouleris, 2018). As one example, traditional concepts in psychometrics, such as validity, do not receive near the same extent of coverage in machine learning, requiring both the translation of ideas across disciplines and refinement in how the concept of validity can be applied with new types of data (Bleidorn & Hopwood, 2019). Machine learning also opens up a greater possibility of exploratory analyses, which could be seen to be at odds with an increase in a recent focus on performing strictly confirmatory analyses (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). This allows researchers to increase their focus on prediction and less on explanation, as many of the machine-learning algorithms provide less information on the relationships in the data (Yarkoni & Westfall, 2017). One area of integration that has received very little focus is that of reliability and its impact on the conclusions derived.

Psychological data often contain traditional types of variables such as demographics, health, personality, or other psychosocial factors. In most cases, these variables are not perfectly measured. Although the effect of measurement quality (error) in predictor variables is well understood in the case of linear models, pioneered by psychometric work in the middle of the 20th century (e.g., see Lord & Novick, 1968), less is understood in nonlinear methods, such as the algorithms in machine learning. A further contrast is that most articles in machine learning are concerned with prediction (e.g., examining R^2) as opposed to explanation (examining β coefficients; see Yarkoni & Westfall, 2017). Our focus is the effect of measurement error in predictor variables on prediction, particularly changes in the explained variance, which differs from most previous research that has examined the effect of measurement error on the regression (path) coefficients (e.g., Cole & Preacher, 2014; Rhemtulla, van Bork, & Borsboom, 2020).

Our goal is to highlight that the use of machine learning does not negate the influence of measurement in behavioral-science data. To put it more plainly, throwing the same set of poorly measured variables that have been analyzed previously into machine-learning algorithms is highly unlikely to produce new insights or findings. We specifically refer to the influence of measurement error, as this can be calculated to a relatively accurate degree in psychology studies, in contrast to the more general notion in machine learning of "garbage in, garbage out." To demonstrate why this is the case and how poor measurement quality affects the results of analyses, we present a simulated example, a small simulation study, and discuss the findings.

Simulated Example

For this example, we simulated two x variables and one y variable with the following nonlinear relationship:

$$y = \cos(x_1) + \sin(x_2) + \tan(0.1 \times x_1 \times x_2) + N(0,.1) \quad (1)$$

This model contains two nonlinear conditional effects for x_1 and x_2, a nonlinear interaction between the variables, and normally distributed errors. These marginal relationships are displayed in Figure 1. The predictor variables are perfectly reliable, and we directly modeled their relationship with the outcome. Given the nonlinearity inherent in these data, a linear model would be

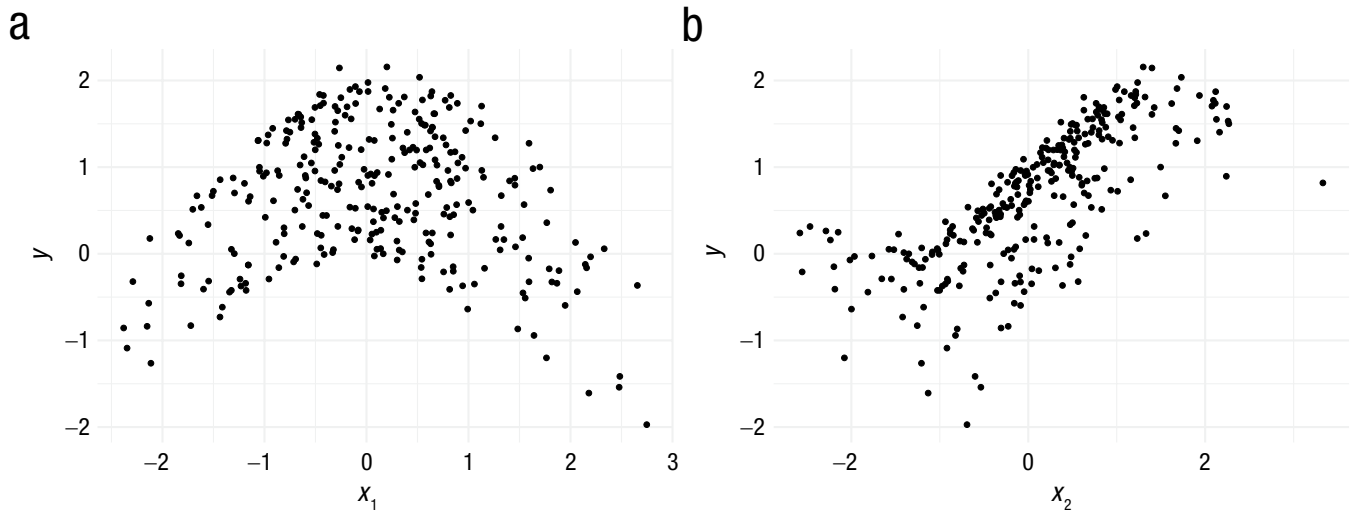

Fig. 1. Two simulated nonlinear relationships. In (a), the graph shows a simulated relationship according to a cosine-functional form, whereas in (b), the graph shows a sine-functional form.

expected to fit more poorly than a popular machine-learning method, such as *boosting* (Freund & Schapire, 1995; Friedman, 2001), that can account for nonlinear effects and complex interactions. Boosting is a method that combines the use of hundreds (or thousands) individual decision trees in an attempt to overcome the drawbacks of single trees. Boosting creates a sequence of trees on the basis of reweighted forms of the data set, namely the residuals derived from the prediction of the prior set of trees.

A linear-regression model that included an interaction term achieved an R^2 of .54 and .53 without the interaction, as assessed across 25 bootstrap samples. In contrast, a stochastic boosting model—using the *gbm* package (Ridgeway, 2017)[1] for the R software environment (R Core Team, 2017)—achieved an R^2 of .93 across 25 bootstrap samples, whereas a correctly specified nonlinear regression model explained 97% of the variance. Using this model comparison, there is clearly more evidence in favor of nonlinearity and possible interaction effects in the data.

One fundamental assumption of linear regression is the inclusion of perfectly measured predictors (e.g., Cohen, Cohen, West, & Aiken, 2014). To investigate the ability of models to recover these nonlinear relationships when this assumption does not hold, we added measurement error to the two predictors so that each variable has a reliability of 0.25. We can think of the perfectly measured variables as ideal forms of what we desire to measure. However, in practice, most constructs that we wish to assess cannot be perfectly measured. Instead, we generally specify a model with observed variables that we know are not perfect because they contain some degree of noise resulting from question wording, participant fatigue, or inattention; differences in the latent variable and observed variable measure; or additional factors. The degree of noise (i.e., error) variance in each observed indicator can be depicted as 1 – reliability, or more formally, x_2:

$$x_1^* = \sqrt{\text{reliability}_{x_1^*}} \times x_1 + N(0, 1 - \text{reliability}_{x_1^*})$$
$$x_2^* = \sqrt{\text{reliability}_{x_2^*}} \times x_2 + N(0, 1 - \text{reliability}_{x_2^*}) \quad (2)$$

where x_1 and x_2 are the perfectly reliable variables, x_1^* and x_2^* are observed versions of x_1 and x_2 that are measured with error, and $N()$ refers to the normal distribution. Simply speaking, the higher the reliability, the more similar a respondent's true and observed responses will be. Now, x_1^* and x_2^* are measured with some error, which will lead to repercussions for examining the associations with other variables. This error attenuates the association between x_1 and y by the degree of error, or 1 – reliability, in x_1^*. For example, if the association, β, between x_1 and y was 0.5, and the variance of x_1 was 1, the expected regression coefficient between x_1^* and y would be $0.5 \times \sqrt{\text{reliability}_{x_1}}$. The attenuation of regression coefficients due to measurement error in predictors in linear-regression modeling has been well understood by methodologists in the social and behavioral sciences since the beginning of the 20th century. What is less well understood is how the effects of measurement error generalize beyond the linear framework, specifically to nonlinear associations. To give an example of this generalization, we added measurement error to x_1 and x_2 following Equation 2 reliabilities equal to 0.5 for x_1 and x_2. Given the degree of error imparted into each of these variables, a drastically reduced model of fit would be expected. We also wished to see whether our model-comparison approach would lead us to conclude that a linear model fits better than boosting despite the true associations being nonlinear. To better grasp the effects of measurement error, scatterplots of the association between x_1^* and y and the association between x_2^* and y are displayed in Figure 2.

In Figure 2, it is hard to distinguish any association, let alone a nonlinear relationship between each of the observed variables and y. To determine how reliability influences our prediction performance, we used the same set of models as before. Now, with our two imperfectly measured predictors of y, the linear model without interaction terms explained 16% of the variance, whereas the boosting model explained 14% of the variance. Not only does our prediction worsen to a large degree, but our assessment of models changes, with a linear model performing comparably to (or better than) boosting.

Our point in this example is twofold. First, we want to highlight that using machine-learning algorithms does not negate the necessity of quality measurement. Second, in performing model comparison, our results are conditional not only on what variables are included in the model but also the quality of these variables. The fact that we use more powerful machine-learning methods does not negate the term garbage in, garbage out.

Simulation Study

To drive our point home, we conducted a small simulation that varied some of the conditions of the above example. We simulated reliability index values of 0.3, 0.6, and 0.9 while also varying the sample size of 200, 500, and 2,000 to determine whether collecting larger samples can mitigate some of the deleterious effects of measurement error. The results obtained from the same methods (R^2 evaluated using 25 bootstrap samples and hyperparameters for boosting) and number of indicators as above are displayed in Figure 3.

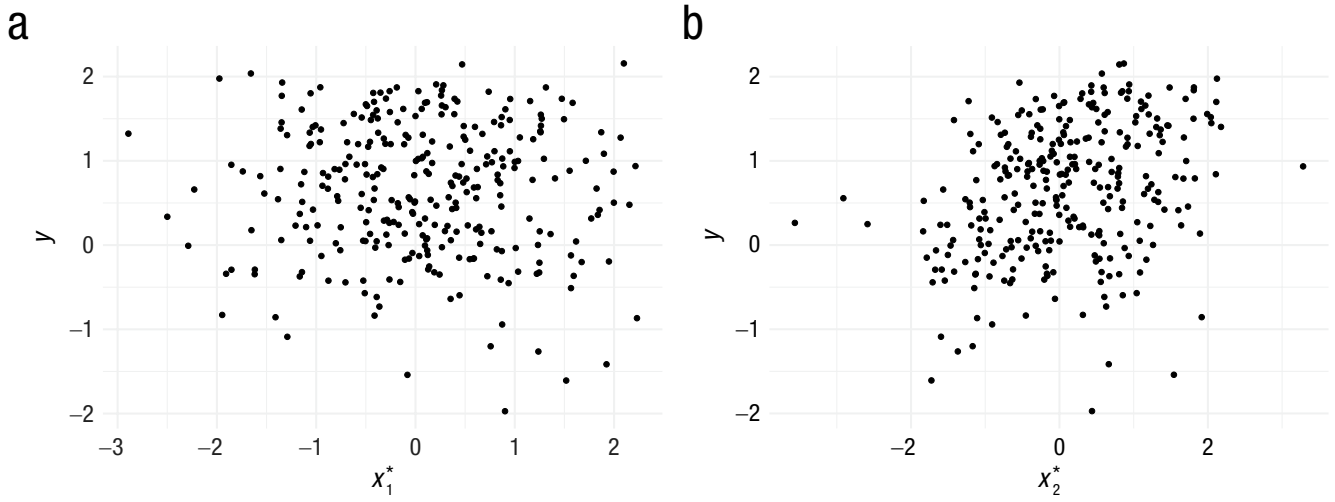

Fig. 2. Two simulated observed indicators of the first true nonlinear relationship. x_1^* and x_2^* are observed versions (that are measured with error) of perfectly reliable variables x_1 and x_2, respectively.

In Figure 3, we can clearly see the effects of measurement error. Given that the true relationship between predictors and the outcome is nonlinear, it is not surprising that the boosting model has the highest R^2 compared with the linear model. Although sample size had a small effect on the boosting results, there seems to be almost zero effect for the linear model, which is in line with prior research (van der Ploeg, Austin, & Steyerberg, 2014). An interesting shift in these results occurs at the reliability indexes of 0.6 and 0.3. The linear model demonstrates a less dramatic decline in R^2, with higher R^2 values than boosting at a reliability index of 0.3. Across sample sizes, 97% of the replications had higher R^2 values in the linear model than in boosting when the reliability was 0.3. If the indicator variables had reliabilities of 0.3, most often researchers would conclude that the linear model fit best despite the underlying nonlinear relationship. Simply put, poor measurement prevents the discovery of these interesting relationships.

To determine whether these results hold in the case of simulated linear relationships, we repeated the simulation but with linear relationships as in the following:

$$y = x_1 + x_2 + 0.1 \times x_1 \times x_2 + N(0,.1) \quad (3)$$

Here, we would expect the results from both the linear regression and boosting to be similar. Linear regression follows the underlying true model, whereas boosting should be able to model the linear relationship to a large degree. Results from this simulation are displayed in Figure 4.

The first thing to note is the unsurprising increased performance of linear regression. Although both models saw a decrement in performance as the reliability was decreased, boosting was also influenced by sample size, with smaller sample sizes resulting in slight decreases in R^2 values across the use of both observed and latent predictors. In contrast, linear-regression results were relatively impervious to sample size.

What to Do?

Our first recommendation is to assess the reliability of both the outcome and predictors. This is most likely more relevant to predictors rather than an outcome, as a variable needs to be paired with other similar variables to assess reliability. We have no specific recommendations with respect to selecting among the many reliability metrics and instead refer readers to McNeish (2018). Not assessing the measurement error inherent in the variables of interest precludes identifying a possible rationale for results such as poor predictive performance or bias in parameter estimates. After assessing reliability, researchers can then be more accurately informed as to the modeling options available and how to use these options to maximize their research aims. We detail four modeling scenarios and the corresponding model framework in Table 1.

Psychological researchers generally assume that the relationships between the covariates and outcome are truly linear, utilizing one of the many forms of generalized linear models. Our purpose was to highlight the disconnect between what exists in truth and what we can model in our data. If there is measurement error among predictors or outcomes, or both, these can be modeled in the structural-equation modeling (SEM) framework.[2] In contrast, if the relationship between covariates and

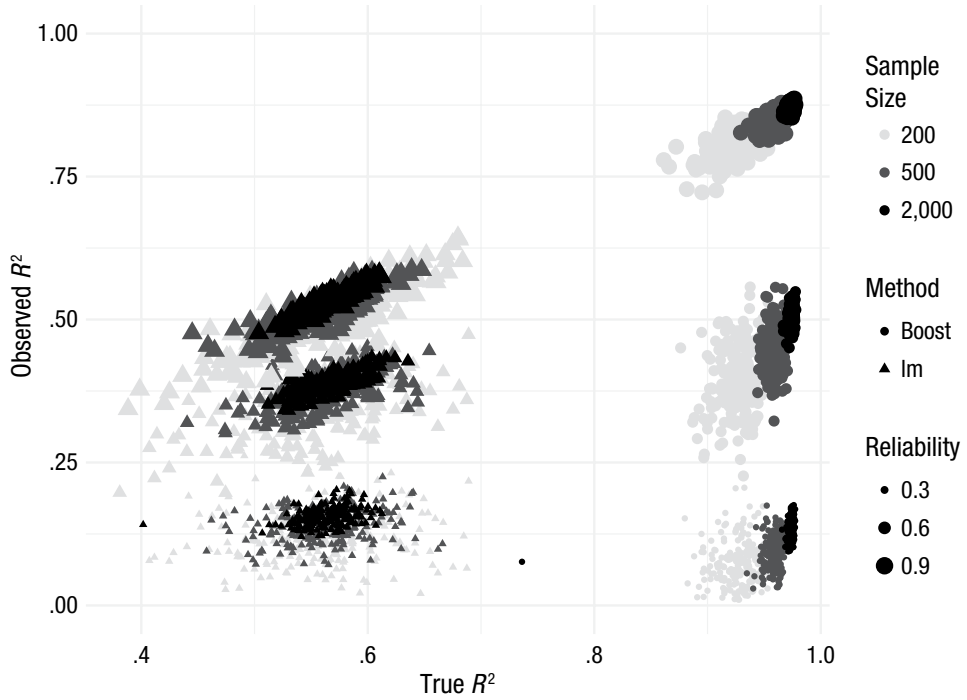

Fig. 3. Results from a nonlinear-relationship simulation with varying reliability of eight predictor variables. Boost = boosting; lm = linear regression.

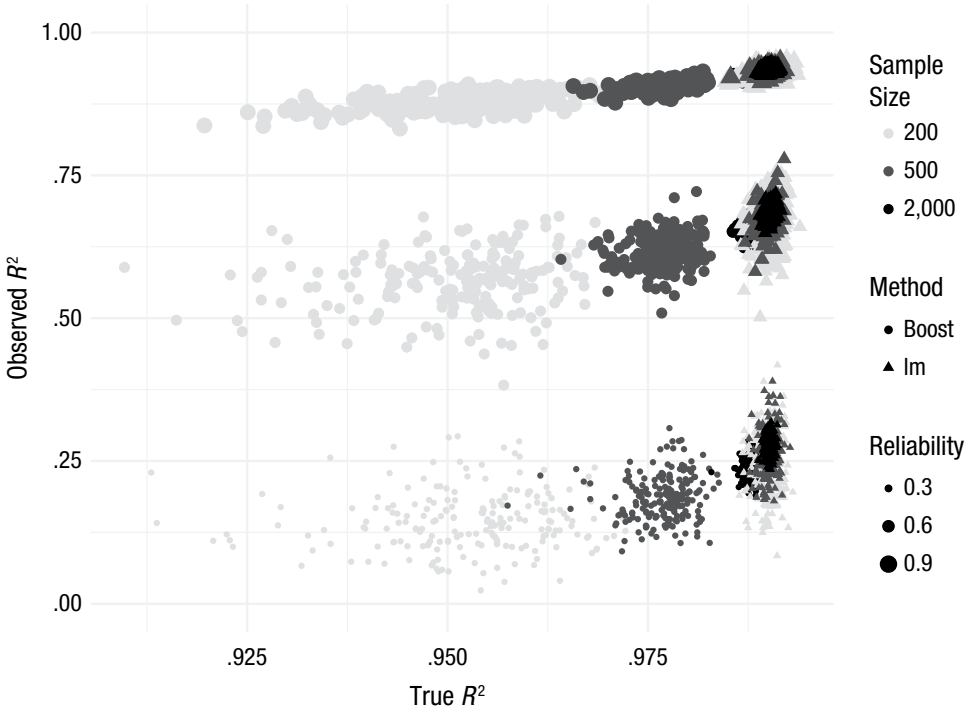

Fig. 4. Results from a linear-relationship simulation with varying reliability of eight predictor variables. Note that the results for boosting at a sample size of 2,000 are essentially identical (partially hidden) to those from linear regression. Boost = boosting; lm = linear regression.

Table 1. Various Options for Pairing Data and Statistical Models

Relationships in data	Model frameworks
Linear relationships and well-measured predictors	Generalized linear model
Linear relationships and poorly measured predictors	SEM framework—forming latent variables among similar variables and then using these as predictors; if single indicators, one can also fix the factor loading to 1.0 and the unique variance to a value based on the estimate of reliability (e.g., see Williams & O'Boyle, 2008)
Nonlinear relationships and well-measured predictors	Machine learning
Nonlinear relationships and poorly measured predictors	Bayesian SEM with quadratic effects from latent variables or fusion of machine learning and SEM (e.g., SEM trees)

Note: SEM = statistical-equation modeling.

outcome are nonlinear (either all or a subset of) and the variables have a low degree of measurement error, one of the many machine-learning algorithms (e.g., random forests or support vector machines) can be used.

If the covariates have low degrees of reliability, these variables can be modeled as indicators of latent variables, which in turn are used as predictors of the outcome of interest. If there is a nonlinear relationship with the outcome of interest, additional modeling steps are necessary for using SEM because it is a fundamentally a linear framework. There has been recent research done on extending SEM to the use of nonlinear relationships (e.g., Umbach, Naumann, Brandt, & Kelava, 2017), which has been aided by a recent surge in Bayesian SEM (Van De Schoot, Winter, Ryan, Zondervan-Zwijnenburg, & Depaoli, 2017). Software such as Stan (Carpenter et al., 2017) can aid in modeling nonlinear relationships with latent variables while also scaling to larger numbers of variables. In addition, some methods that fall under the umbrella of machine learning have recently been integrated into the SEM framework such as SEM trees (Brandmaier, von Oertzen, McArdle, & Lindenberger, 2013), allowing for more automatic forms of modeling nonlinear relationships with latent variables. Although SEM trees can model latent variables as outcomes, the estimation of factor scores is necessary for predictor variables given the partitioning algorithm. In addition, recent proposals have improved the scalability of SEM to big data (numbers of variables), most notably with the integration of regularization (e.g., Jacobucci, Grimm, & McArdle, 2016; van Kesteren & Oberski, 2019). We expect this area to be an active area of research in the near future.

The selection and assessment of predictors is an important facet of research design, which can have important downstream effects when it comes to modeling an outcome of interest. Although a large amount of recent research has focused on the impact of poor measurement quality of new types of data collection facilitated by machine learning (e.g., Qiu, Chan, & Chan, 2018), this does not mean that most work done in assessing traditional psychological scales is done sufficiently. Using psychological scales in ways in which they were not intended or selecting scales that have not been previously validated are just two of the many ways in which researchers threaten aspects of validity (e.g., see Flake & Fried, 2019). We advocate for researchers to go beyond just traditional survey validation (e.g., see Maul, 2017), taking into account other factors that can influence the relationship between X and Y. We would expect similar simulation results if we had imparted noise due to careless responding (Patton, Cheng, Hong, & Diao, 2019), random guessing (Hong & Cheng, 2019), or other factors that can influence the quality of responses. Although our focus in this article has been on how well-observed variables represent a latent construct, similar principles of measurement apply to beyond just this narrow context.

Note that we have assumed our latent variables are indicated by multiple observed variables. In this modeling scenario, we can calculate each variable's reliability as well as model the unreliabiliy in the SEM framework. In contrast, we may have an unreliable variable but no similar variables that purport to measure a similar construct (for more detail on this topic, see Loken and Gelman, 2017). In research scenarios such as this, new forms of collecting data can provide additional avenues for augmenting single variables. For example, if personality variables are of interest as predictors, data collected from the Big Five Inventory could be paired with language-based scores extracted from social-media posts (for a general overview, see Bleidorn & Hopwood, 2019). This multimethod assessment can shed light on aspects of construct validity, and the joint use of predictors extracted from different forms of assessment could add incrementally to the modeling of the outcome. In addition, detail with respect to sample size has been largely left out of our discussion despite its critical role in the discourse surrounding replication (Morey & Lakens, 2016). SEM is most notably a large-sample technique. Thus, assessing latent variables, let alone modeling nonlinear relationships, applies chiefly in research scenarios in which the number of participants is in the hundreds, if not thousands.

One of the most problematic aspects of the pairing of machine learning and unreliable variables is with respect to model comparison. In the context of machine learning, although some researchers apply only one statistical method, it is advantageous to compare algorithms across degrees of complexity, such as through the use of linear models (either regularized or nonregularized forms of linear or logistic regression) to nonlinear algorithms (random forests or boosting), as this can shed light on the forms of relationships in the data (Hong, Jacobucci, & Lubke, 2020). Applying only linear models can lead to overlooking nonlinear effects, whereas applying only nonlinear algorithms can lead researchers to conclude that more complex relationships exist than actually do. However, as discussed previously, choosing the "best" model for the data is conditional on the measurement of the variables. Understanding the reliability of the predictors (and outcome) can be informative as to whether new forms of measurement of the constructs of interest are necessary to overcome the poor performance and why a linear model fit best when nonlinear or interactive effects were hypothesized.

Conclusion

We highlighted one possible explanation for the lack of innovations in pairing traditional behavioral-science variables with machine-learning algorithms. The results of our simulation highlight contradictory concerns regarding the pervasive fear of overfitting the data or concluding that more complex relationships exist in our sample than actually do. Measurement error can lead to severely underfitting the true relationships. In addition, from a hypothesis-driven modeling perspective, the results of our simulation present a challenge to model comparison and selection, namely that the choice of models, as well as hypotheses, is conditional on the measurement of the variables regardless of whether linear or nonlinear relationships exist. We advocate for comparing results across methods (i.e., linear regression and boosting), as well as assessing the measurement error of both the outcome and predictors.

Transparency

Action Editor: Laura A. King
Editor: Laura A. King
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.

ORCID iD

Ross Jacobucci https://orcid.org/0000-0001-7818-7424

Notes

1. Parameters were set to a shrinkage of 0.1, interaction depth of 2, and 50 trees.
2. However, researchers should exert caution that the latent variables are properly specified (e.g., Rhemtulla, van Bork, & Borsboom, 2020).

References

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*, 899–917.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*, 190–203.

Brandmaier, A. M., von Oertzen, T., McArdle, J. J., & Lindenberger, U. (2013). Structural equation model trees. *Psychological Methods*, *18*, 71–86.

Burke, T. A., Ammerman, B. A., & Jacobucci, R. (2019). The use of machine learning in the study of suicidal and non-suicidal self-injurious thoughts and behaviors: A systematic review. *Journal of Affective Disorders*, *245*, 869–884.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., . . . Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, *76*, 1–37.

Chen, E. E., & Wojcik, S. P. (2016). A practical guide to big data research in psychology. *Psychological Methods*, *21*, 458–474.

Cohen, J., Cohen, P., West, S. G., & Aiken, L. S. (2014). *Applied multiple regression/correlation analysis for the behavioral sciences* (3rd ed.). Mahwah, NJ: Erlbaum.

Cole, D. A., & Preacher, K. J. (2014). Manifest variable path analysis: Potentially serious and misleading consequences due to uncorrected measurement error. *Psychological Methods*, *19*, 300–315.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*, 91–118.

Flake, J. K., & Fried, E. I. (2019, January 17). *Measurement schmeasurement: Questionable measurement practices and how to avoid them*. doi:10.31234/osf.io/hs7wm

Freund, Y., & Schapire, R. E. (1995). A decision-theoretic generalization of on-line learning and an application to boosting. In P. Vitanyi (Ed.), *Computational learning theory: Second European conference, EuroCOLT '95, Barcelona, Spain, March 13–15, 1995. Proceedings* (pp. 23–37). Berlin, Germany: Springer.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, *29*, 1189–1232.

Hong, M., Jacobucci, R., & Lubke, G. (2020). Deductive data mining. *Psychological Methods*. Advance online publication. doi:10.1037/met0000252

Hong, M. R., & Cheng, Y. (2019). Clarifying the effect of test speededness. *Applied Psychological Measurement*, *43*, 611–623.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, *23*, 555–566.

Jie, M. A., 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.

Just, M. A., Pan, L., Cherkassky, V. L., McMakin, D. L., Cha, C., Nock, M. K., & Brent, D. (2017). Machine learning of neural representations of suicide and emotion concepts identifies suicidal youth. *Nature Human Behaviour*, *1*, 911–919.

Loken, E., & Gelman, A. (2017). Measurement error and the replication crisis. *Science*, *355*, 584–585.

Lord, F. M., & Novick, M. R. (1968). *Statistical theories of mental test scores*. Reading, MA. Addison-Wesley.

Maul, A. (2017). Rethinking traditional methods of survey validation. *Measurement: Interdisciplinary Research and Perspectives*, *15*, 51–69.

McNeish, D. (2018). Thanks coefficient alpha, we'll take it from here. *Psychological Methods*, *23*, 412–433.

Morey, R. D., & Lakens, D. (2016). *Why most of psychology is statistically unfalsifiable*. Unpublished manuscript. doi:10.5281/zenodo.838685

Patton, J. M., Cheng, Y., Hong, M., & Diao, Q. (2019). Detection and treatment of careless responses to improve item parameter estimation. *Journal of Educational and Behavioral Statistics*, *44*, 309–341.

Qiu, L., Chan, S. H. M., & Chan, D. (2018). Big data in social and psychological science: Theoretical and methodological issues. *Journal of Computational Social Science*, *1*, 59–66.

Rhemtulla, M., van Bork, R., & Borsboom, D. (2020). Worse than measurement error: Consequences of inappropriate latent variable measurement models. *Psychological Methods*, *25*(1), 30–45. Advance online publication. doi:10.1037/met0000220

R Core Team. (2017). R: A language and environment for statistical computing (Version 3.4.0) [Computer software]. Retrieved from https://www.r-project.org/index.html

Ridgeway, G. (2017). gbm: Generalized boosted regression models (Version 2.1.3) [Computer software]. Retrieved from https://CRAN.R-project.org/package=gbm

Umbach, N., Naumann, K., Brandt, H., & Kelava, A. (2017). Fitting nonlinear structural equation models in R with package nlsem. *Journal of Statistical Software*, *77*, 1–20.

van der Ploeg, T., Austin, P. C., & Steyerberg, E. W. (2014). Modern modelling techniques are data hungry: A simulation study for predicting dichotomous endpoints. *BMC Medical Research Methodology*, *14*, 1–13.

Van De Schoot, R., Winter, S. D., Ryan, O., Zondervan-Zwijnenburg, M., & Depaoli, S. (2017). A systematic review of Bayesian articles in psychology: The last 25 years. *Psychological Methods*, *22*, 217–239.

van Kesteren, E. J., & Oberski, D. L. (2019). Exploratory mediation analysis with many potential mediators. *Structural Equation Modeling: A Multidisciplinary Journal*, *26*, 710–723.

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science*, *7*, 632–638.

Walsh, C. G., Ribeiro, J. D., & Franklin, J. C. (2017). Predicting risk of suicide attempts over time through machine learning. *Clinical Psychological Science*, *5*, 457–469.

Williams, L. J., & O'Boyle, E. H., Jr. (2008). Measurement models for linking latent variables and indicators: A review of human resource management research using parcels. *Human Resource Management Review*, *18*, 233–242. doi:10.1016/j.hrmr.2008.07.002

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122. doi:10.1177/1745691617693393