2026 ■ Vol. 22 ■ no. 2

Robustness of Regularized Regression Methods Under Compound Model Misspecification: A Simulation Benchmarking Study

Onur Ramazan [a] and Yiu Wa Lui [b]

[a]The University of Hong Kong

[b]The Chinese University of Hong Kong

Abstract ■ Regularized regressions are widely used in psychological research where the fitted model is assumed to be correctly specified. Yet, psychological data routinely violates assumptions of linearity, homoscedasticity and additivity which raises questions about estimator robustness when models are misspecified. The present simulation study examined how regularized estimators—ridge, LASSO, elastic net, adaptive LASSO, SCAD and MCP—degrade under compound misspecification. Using high-dimensional data with block-wise correlation (p = 2,000) and three sample sizes (n = 500, 2,000 and 10,000), we generated outcomes under correct specification and compound misspecification incorporating nonlinearity, heteroscedasticity and omitted interactions. Performance was evaluated through predictive accuracy (out-of-sample R^2 and RMSE) and selection stability (mean Jaccard index across replicates). Results revealed that misspecification produced modest degradation in predictive accuracy with R^2 reductions of 0.006–0.017 across sample sizes. Selection stability was minimally affected by misspecification at smaller samples. At n = 10,000, non-convex methods achieved highest stability.

Keywords ■ Regularized regression, Model misspecification, Variable selection stability, High-dimensional data, Monte Carlo simulation study.

✉ oramazan@hku.hk

doi 10.20982/tqmp.22.2.p090

Acting Editor ■ Denis Cousineau (Université d'Ottawa)

Reviewers
■ One anonymous reviewer

Introduction

Prediction stands as a fundamental task across the applied psychological sciences (Fokkema et al., 2022; Yarkoni & Westfall, 2017). Industrial-organizational psychologists forecast job performance from aptitude and integrity measures (e.g., Iliescu et al., 2023; Richardson & Norgate, 2015), clinical psychologists anticipate treatment outcomes for patients with complex presentations and forensic psychologists assess recidivism risk to inform parole decisions (e.g., Janković et al., 2022; Moulden et al., 2024). Decades of research have decisively demonstrated that statistical models, on average, produce more accurate predictions than human judgment combining the same information (Meehl, 1954; Grove et al., 2000; Ægisdóttir et al., 2006). This empirical reality has driven the widespread adoption of predictive modeling in psychological research and practice with machine learning methods increasingly deployed to handle the high-dimensional, multi-source data that characterize modern psychological science (Dwyer et al., 2018; Ostojic et al., 2024; Yarkoni & Westfall, 2017).

Among the most influential developments are regularized machine learning methods: ridge (Hoerl & Kennard, 1970), LASSO (Tibshirani, 1996), and elastic net (Zou & Hastie, 2005). These techniques constrain coefficient estimation to balance model complexity against predictive accuracy and have become foundational tools across psychological domains (McNeish, 2015; Jacobucci et al., 2016; Yarkoni & Westfall, 2017). The statistical properties of these estimators are well-characterized under ideal conditions. When the data-generating process is linear, homoscedastic and additive, researchers can confidently select among methods based on their theoretical advantages. However, psychological data rarely satisfy these assumptions. Con-

structs are inherently latent, predictors contain measurement error, and true functional forms are seldom known (Hastie et al., 2009; James et al., 2021; Yarkoni & Westfall, 2017). Human behavior emerges from complex systems where nonlinearities, threshold effects and interactions are the rule rather than the exception (Mayer-Kress et al., 2006; Pelech, 2011; Turkheimer et al., 2021). Recent critiques emphasize that psychological constructs may be poorly aligned with the linear and additive models routinely applied to them (De Boeck et al., 2023; Song et al., 2025; Turkheimer et al., 2021).

The implications for regularized machine learning are profound yet underexplored. While simulation studies have rigorously compared methods under correct specification (e.g., Kipruto & Sauerbrei, 2025; Pavlou et al., 2015), their performance when the fitted model is systematically wrong remains largely unknown. How these methods degrade when nonlinearities are present but unmodeled remains unknown. Whether LASSO's aggressive selection becomes unstable under heteroscedasticity is an open question. These gaps carry substantial practical importance as applied researchers may not know the true data-generating process and would rely on methods robust to inevitable discrepancies between model and reality. Parallel developments in uncertainty quantification offer promising avenues for addressing these challenges (e.g., Jia et al., 2013; Kayanan & Wijekoon, 2019; Ter Braak, 2009; Zou & Zhang, 2009).

The present study addresses these gaps through a controlled simulation introducing compound misspecification such as nonlinearities, heteroscedasticity and omitted interactions as applied simultaneously to high-dimensional data with block-wise correlation structures characteristic of psychological measurement (Beaujean, 2017; Benson, 2017). Our evaluation framework incorporates predictive accuracy (R^2, RMSE) and selection stability measured through the mean and standard deviation Jaccard indices across replicates (Baldassarre et al., 2017). This multi-faceted evaluation allows us to ask not merely which method predicts best, but which maintains reliable uncertainty quantification and stable selection when core assumptions are violated.

The present study has been guided by the following research questions:

RQ1. How does compound model misspecification affect the predictive accuracy (out-of-sample R^2 and RMSE) of regularized regression methods across small, intermediate and large sample sizes?

RQ2. Do shrinkage-dominated methods like ridge regression exhibit greater robustness to compound misspecification in terms of predictive accuracy compared to selection-oriented methods like LASSO, adaptive LASSO, SCAD and MCP?

RQ3. How does compound misspecification impact the variable selection stability (mean Jaccard index) of sparsity-inducing methods and do non-convex methods (SCAD, MCP) differ from convex methods (LASSO, elastic net, adaptive LASSO) in their selection stability under misspecification?

Methods

Data Generation

We conducted a simulation (Beaujean, 2017; Benson, 2017) to evaluate the robustness of regularized regression methods under controlled model misspecification. All datasets were generated with 2,000 predictors and three sample sizes of 500, 2,000, and 10,000 observations by following recommendations for high-dimensional simulation studies in the psychological literature (Paxton et al., 2001). Predictors were drawn from a multivariate normal distribution with zero mean and a block-diagonal covariance matrix designed to mimic the correlated structure common in psychological measurements (Bühlmann & van de Geer, 2011). Specifically, predictors were grouped in 100 blocks of size 20 where within each block, the correlation was 0.4 while between blocks the correlation was zero which reflected the clustered nature of psychological constructs where items measuring the same latent variable tend to be moderately correlated (Yuan et al., 2019). True coefficients were constructed to reflect a dense-but-weak signal structure where most predictors exert small effects and a minority have moderate effects which was consistent with the sparsity-of-effects principle observed in many psychological applications (Baldassarre et al., 2017; Wu & Hamada, 2021). Formally, 10% of predictors received coefficients drawn from a uniform distribution between 0.2 and 0.5 with random sign, while the remaining 90% received coefficients drawn from a uniform distribution between 0.01 and 0.1 also with random sign. This specification moves beyond the strict sparsity often assumed in methodological research and better reflects the reality that many psychological variables may have non-zero but trivial predictive value (Lavelle-Hill et al., 2025). The set of predictors receiving larger coefficients was randomly selected and remained consistent across all replications. The linear predictor was computed as the matrix product of the predictor matrix and the coefficient vector. Baseline homoscedastic errors were generated as normally distributed with mean zero and variance chosen so that the signal-to-noise ratio yielded an approximate population R^2 of 0.8 in the correctly specified model specifically by setting the error standard deviation to 0.5 times the standard deviation of the linear predictor by following established practices in sim-

ulation research (Hastie et al., 2009; James et al., 2021).

Misspecification Layers

To examine performance when the fitted linear-additive model is incorrect, we created four misspecification conditions by each modifying the outcome generation while keeping the predictor matrix and true coefficients unchanged. The baseline condition which served as the reference point generated outcomes as the linear predictor plus homoscedastic normal error. The nonlinearity condition introduced nonlinear effects through quadratic terms and threshold functions by following evidence that psychological relationships often exhibit curvilinearity and abrupt transitions (Meehl, 1990; De Boeck et al., 2023). For this condition, 50 randomly selected predictors received quadratic effects with coefficients drawn from a uniform distribution between negative 0.05 and 0.05 and 20 randomly selected predictors received threshold effects where the indicator of a positive predictor value was multiplied by coefficients drawn from a uniform distribution between negative 0.2 and 0.2. The magnitudes were chosen to be modest so that the nonlinearities would not dominate the linear signal, consistent with recommendations for building realistic simulation designs (Skrondal, 2000). The heteroscedasticity condition made error variance proportional to the absolute value of the linear predictor, such that observations with larger predicted values exhibited greater variability which is a pattern frequently encountered in psychological data where larger responses are often more variable (Raudenbush & Bryk, 2002). The omitted interactions condition added pairwise interactions between 100 pairs of predictors that were correlated by virtue of belonging to the same block which reflected the reality that interactive effects among related constructs are common but often unmodeled (Aiken & West, 1991). Interaction coefficients were drawn from a uniform distribution between -0.1 to 0.1 and these interactions were deliberately omitted from the subsequently fitted models which created structural misspecification. Finally, the compound misspecification condition applied all three misspecifications simultaneously by combining nonlinear terms, heteroscedastic errors and omitted interactions which represented the most realistic scenario given that real-world data typically violate multiple assumptions concurrently (Lei et al., 2018). For every replication and sample size, we generated one dataset for the base layer with no misspecification and one dataset with the compound misspecification.

Regularization Methods

We compared penalized regression estimators that represent a spectrum of shrinkage and selection behaviors.

Ridge regression. Ridge regression applies an L_2 penalty that shrinks coefficients continuously toward zero while retaining all predictors in the final model which is a method originally developed to address multicollinearity in ordinary least squares regression (Hoerl & Kennard, 1970). More formally, the ridge estimator is defined as the solution to:

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{i,j}\ \beta_j)^2 + \lambda_2 \sum_{j=1}^{p} {\beta_j}^2 \tag{1}$$

where β are coefficients for predictors x, and λ_2 is the shrinkage strength for L_2 penalty.

Note that the L_2 penalty does not set coefficients to exactly 0 and thus, will not result in a sparse model, unlike the models that follow.

LASSO. The least absolute shrinkage and selection operator, commonly known as LASSO, applies an L_1 penalty that enables continuous shrinkage and automatic variable selection by setting some coefficients exactly to zero:

$$\min_{\beta}\ \frac{1}{n}\sum_{i=1}^{n}(y_i\ - \sum_{j=1}^{n} x_{i,\,j}\ \beta_j)^2 + \lambda_{\,1} \sum_{j=1}^{p} |\beta_j| \tag{2}$$

where λ_2 is the shrinkage strength for L_1 penalty. The ability to shrink coefficients to exactly zero makes it particularly useful for high-dimensional data where model interpretability is valued (Tibshirani, 1996).

Elastic Net. The elastic net combines both penalties via a mixing parameter a ranging from zero to one where zero corresponds to ridge and one corresponds to LASSO, with the penalty designed to handle correlated predictors more effectively than LASSO alone (Zou & Hastie, 2005).

$$\min_{\beta}\ \frac{1}{n}\sum_{i=1}^{n}(y - \sum_{j=1}^{n} x_{i,\,j}\ \beta_j)^2 + \lambda\ \alpha \sum_{j=1}^{p} |\beta_j| + \lambda\ (1-\alpha) \sum_{j=1}^{p} {\beta_j}^2 \tag{3}$$

where α is the mixing parameter, and λ is the penalization strength.

Adaptive LASSO. Adaptive LASSO uses a weighted L_1 penalty with data-dependent weights to achieve oracle properties. That is, it can perform as well as if the true underlying model were known in advance asymptotically (Zou, 2006).

$$\min_{\beta}\ \frac{1}{n}\sum_{i=1}^{n}(y_i\ - \sum_{j=1}^{p} x_{i,\,j}\ \beta_j)^2 + \lambda \sum_{j=1}^{p} \omega_j\ |\beta_j| \tag{4}$$

where ω is the adaptive weight. A common choice for ω_j is:

$$\omega_j\ = \frac{1}{\left|{\beta_j}^{init}\right|^{\gamma}} \tag{5}$$

where β_j^{init} is usually obtained through ridge regression and γ is a positive constant. In practice, it is most commonly set to 0.5, 1, or 2 (Zou, 2006). While γ can in principle be tuned via cross-validation, practitioners frequently limit the search to these same discrete values even when using CV.

So far, the methods discussed rely on convex penalized objectives (convex loss plus convex penalty) to shrink coefficients toward zero. Thanks to convexity, efficient algorithms exist such as cyclical coordinate descent (Friedman et al., 2010) and least angle regression (Efron et al., 2004) that compute the full solution path quickly which makes these approaches highly attractive for large-scale and high-dimensional data. However, standard convex penalties (e.g., LASSO and ridge) apply shrinkage in a way that does not adaptively distinguish coefficients based on their likely importance which results in the same qualitative shrinkage behavior across signals of different magnitudes. Adaptive Lasso partially addresses this by placing less shrinkage on coefficients with larger initial estimates (via data-driven weights), but even adaptive Lasso typically introduces some bias toward zero for non-zero coefficients in finite samples. Below, we introduced a different class of methods that employ non-convex (and more precisely, folded concave) penalties. While these lead to non-convex optimization problems with the attendant risk of multiple local minima and greater computational difficulty, they substantially reduce (and often nearly eliminate) bias for large non-zero coefficients, even in finite-sample settings, often approaching the ideal 'oracle' estimator that knows the true support in advance (Fan & Li, 2001).

Smoothly-clipped-absolute-deviation-penalized-regression. The smoothly-clipped-absolute-deviation-penalized-regression (abbreviated as SCAD) employs a folded-concave penalty that avoids excessive bias for large coefficients while maintaining oracle properties, using the regularization parameter a = 3.7 as recommended in the original work (Fan & Li, 2001). In more concrete terms, SCAD optimizes:

$$\min_{\beta} \frac{1}{n}\sum_{i=1}^{n}(y_i - \sum_{j=1}^{n} x_{i,j}\,\beta_j)^2 + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \tag{6}$$

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| & if\ |\beta| \le \lambda \\ \frac{a\lambda|\beta| - \beta^2 - \lambda^2}{a-1} & if \lambda < |\beta| \le a\lambda \\ \frac{\lambda^2(a+1)}{2} & if\ |\beta| > a\lambda \end{cases} \tag{7}$$

For coefficients with magnitude less than λ, SCAD applies an aggressive penalty similar to LASSO. When a coefficient's magnitude falls between λ and $a\lambda$, a quadratic penalty creates a smooth transition from a linear slope to a flat line. Once the magnitude exceeds $a\lambda$, the penalty becomes constant. This property allows SCAD to shrink small coefficients exactly to zero—producing a sparse model—while preserving strong signals by gradually reducing the penalty as the coefficient grows.

Minimax-concave-penalized-regression. The minimax-concave-penalized-regression (MCP) uses another folded-concave penalty that minimizes maximum concavity and has been shown to achieve favorable variable selection properties in high-dimensional settings (Zhang, 2010). The objective function to minimize is similar to SCAD and it only differs at the penalty function:

$$p_\lambda(\beta) = \begin{cases} \lambda|\beta| - \frac{\beta^2}{2\gamma}, & if\ |\beta| \le \gamma\lambda \\ \frac{1}{2}\gamma\lambda^2, & if\ |\beta| > \gamma\lambda \end{cases} \tag{8}$$

In contrast to SCAD, MCP applies a quadratic penalty for coefficients when their magnitude is below the threshold $\gamma\lambda$ and a constant penalty otherwise. All methods were implemented as available in R packages with `glmnet` used for ridge, LASSO, elastic net and adaptive LASSO (Friedman et al., 2010) and `ncvreg` used for SCAD and MCP (Breheny & Huang, 2011).

Model Estimation and Tuning

For each generated dataset, we randomly split the observations into a training set comprising 80% of the data and a test set comprising the remaining 20%, a common split ratio in predictive modeling studies that balances model estimation with validation sample size (Hastie et al., 2009; James et al., 2021). Predictors were standardized to have zero mean and unit variance using the training set means and standard deviations, with test set predictors transformed accordingly using the same values to prevent information leakage, which is essential for obtaining unbiased estimates of out-of-sample performance (Kuhn & Johnson, 2013). Hyperparameters were selected via 5-fold cross-validation on the training set, minimizing mean squared error, as cross-validation remains the standard approach for tuning penalized regression models (Hastie et al., 2009). For ridge and LASSO we tuned the penalty parameter λ over a grid of 100 values, following recommendations for grid density in high-dimensional settings (Friedman et al., 2010). For elastic net we performed a two-dimensional grid search over $\alpha \in 0.1, 0.3, 0.5, 0.7, 0.9$ with λ tuned simultaneously over the same grid which allows the mixing parameter to adapt to the data structure (Zou & Hastie, 2005). For adaptive LASSO, γ is chosen a priori to be 1, weights were computed from a ridge with λ chosen by cross-validation, and then a standard LASSO was fit with those weights, following the original adaptive LASSO procedure (Zou, 2006). SCAD and MCP were tuned over their λ parameter using the default grids provided in the ncvreg package, which are designed to cover the regularization path efficiently (Breheny & Huang, 2011). For SCAD, the shape parameter α is chosen

to be 3.7. For MCP, the concavity parameter γ is chosen to be 3 following the ncvreg's default. After tuning, the final model was refitted on the entire training set using the optimal hyperparameters and predictions were obtained for the test set.

Performance Evaluation

We assessed performance along three dimensions: predictive accuracy and selection stability. For predictive accuracy, we computed for each replication the out-of-sample R^2 and root mean squared error on the test set with these metrics averaged across replications for each condition, following standard practice in simulation research evaluating predictive models (Kipruto & Sauerbrei, 2025; Pavlou et al., 2015). For selection stability, we quantified the consistency of the models' variable selection using the Jaccard index, as stability is essential for interpretable science (Baldassarre et al., 2017). It is important to distinguish selection stability from selection accuracy. We did not measure how well the models recovered the true data-generating variables. Rather, we measured how consistently the models selected the same variables across different random samples. Specifically, for any two replications a and b, let S_a and S_b denote the sets of predictors assigned non-zero coefficients by the model. The pairwise Jaccard index is defined as:

$$J(S_a, S_b) = \frac{|S_a \cap S_b|}{|S_a \cup S_b|} \tag{9}$$

For each combination of sample size, data-generating layer, and sparsity-inducing method (Ridge regression was excluded as it retains all predictors and does not perform variable selection), we computed this index for all possible unique pairs across the 100 replications (yielding 4,950 pairwise comparisons per condition). We then report the mean and standard deviation of these pairwise Jaccard values. We employed parallel computing using the library `foreach` (Daniel et al., 2022) and `doParallel` (Folashade et al., 2022) to perform model fitting and to compute the performance metrics in batches.

Simulation Parameters and Replications

We performed $R = 100$ independent replications for each combination of sample size ($n = 500$, $n = 2000$, $n = 10000$) and misspecification layer. All simulations were conducted in R version 4.3.1 (R Core Team, 2026) using the packages `MASS` (for multivariate normal data), `glmnet` (Friedman et al., 2010), `ncvreg` (Breheny & Huang, 2011), and `caret` (Kuhn, 2008) for cross-validation.

Results

The present simulation study generated 100 independent replications for each combination of three sample sizes ($n = 500$, $n = 2000$, $n = 10000$) and two misspecification conditions: correct specification (Layer 1) and compound misspecification incorporating nonlinearity, heteroscedasticity, and omitted interactions (Layer 2). Performance was evaluated along three dimensions: predictive accuracy assessed by out-of-sample R^2 and root mean squared error (RMSE) and selection stability quantified by the mean Jaccard index across replicates. Results are organized by sample size to demonstrate how method performance and robustness varied as statistical information increases.

Predictive Accuracy

To evaluate predictive performance (Table 2), we calculated the root mean square error (RMSE) to quantify absolute prediction error. To provide a standardized, scale-free measure of how well the models predicted the target relative to a simple mean benchmark, we also calculated out-of-sample R^2.

At the smallest sample size ($n = 500$), predictive performance was modest across all methods with out-of-sample R^2 values ranging from 0.036 to 0.222 under correct specification. Ridge regression achieved the highest predictive performance under both correct specification (M $R^2 = 0.216$, SD = 0.064, where M denotes the mean and SD, the standard deviation) and compound misspecification (M $R^2 = 0.222$, SD = 0.065), demonstrating notable stability in its predictive performance. Elastic net followed closely (M $R^2 = 0.163$, SD = 0.078 under correct specification; M $R^2 = 0.164$, SD = 0.085 under misspecification) while LASSO exhibited somewhat lower performance (M $R^2 = 0.133$, SD = 0.076 and M $R^2 = 0.123$, SD = 0.084, respectively). Adaptive LASSO and the non-convex methods SCAD and MCP showed substantially degraded predictive performance at this sample size with R^2 values below 0.105 across conditions. RMSE values mirrored these patterns, with ridge producing the lowest prediction error (M RMSE = 5.304–5.317) and adaptive LASSO the highest (M RMSE = 5.859–5.902). Compound misspecification produced minimal changes in predictive accuracy at $n = 500$ with R^2 differences between layers not exceeding 0.010 for any method.

At the intermediate sample size ($n = 2000$), predictive accuracy improved substantially for all methods. Elastic net achieved the highest R^2 under correct specification (M = 0.535, SD = 0.038), closely followed by LASSO (M = 0.533, SD = 0.039) and adaptive LASSO (M = 0.522, SD = 0.040). Ridge regression produced lower R^2 (M = 0.409, SD = 0.034) at this sample size, reflecting its inability to benefit from increased information to the same extent as methods capable of variable selection. SCAD and MCP demonstrated R^2 values of 0.391 and 0.365, respectively, under correct specification. Compound misspecification reduced R^2 for all methods by approximately 0.006 to 0.011 points, with the largest

Table 1 ■ Simulation Algorithm Steps for Data Generation

Step	Component	Description
1	Predictor Selection	From the full set of predictors : $X = \{x_1, x_2, \ldots, x_p\}$ a. Sample a subset S_{large} of predictors have large linear effects. b. Sample a subset S_{small} of predictors have small linear effects. c. Sample a subset S_{quad} of predictors have quadratic effects. d. Sample a subset S_{thresh} of predictors have threshold effects.
2	Interaction Selection	From all possible pairs of predictors $\{x_i, x_j\}$ (where $i < j$): Sample a subset S_{inter} of pairs have interaction effects.
3	Effect Size Assignment	For each selected effect (linear, quadratic, threshold, interaction), draw the corresponding effect size (β).
4	Data Generation Loop	For each replication "rep" = 1 to R: a. Draw X_{full} from a Multivariate Normal distribution with mean zero and covariance matrix Σ. b. Compute the base linear predictor $LP_{full} = \beta X_{full}$. c. Draw the base noises $\forall j \in [1:10000]$: N(0, σ^2), where σ = 0.5 * $sd\,(LP_{full})$ d. Draw the heteroscedastic noise $\forall j \in [1:10000]$: $\tilde{}_{ind.}$ N(0,$\sigma^2{\omega_j}^2$), where $\tilde{}_{ind.}$ N(0, $\sigma^2{\omega_j}^2$), where ω_j is defined as $\frac{\lvert LP_{full,\,j}\rvert}{sd(LP_{full})}$. e. Compute the outcome $Y_{base\ full}$ and $Y_{mis\ full}$ with and without misspecification layers respectively. a. $X_n \leftarrow X_{full}$ [1:n,:] (truncate the predictor matrix). b. $Y_{base\ n} \leftarrow Y_{base\ full}$ [1:n] and $Y_{mis\ n} \leftarrow Y_{mis\ full}$ [1:n] (truncate the outcome vectors). c. Store X_n, $Y_{base\ n}$and $Y_{mis\ n}$ for the current sample size and replication.

Note. p = total number of predictors; R = number of simulation replications; Σ = covariance matrix for the predictors. The effect sizes (β) for linear, quadratic, threshold, and interaction terms are drawn from pre-specified distributions. The final outcome Y_{full} is a composite function of the linear predictor, the selected effects (with their β's), and the two noise components.

absolute decline observed for MCP ($\Delta R^2 = \breve{}0.010$) and adaptive LASSO ($\Delta R^2 = \breve{}0.009$). RMSE increased correspondingly under misspecification, with all methods showing elevations of 0.085 to 0.103.

At the largest sample size ($n = 10000$), predictive accuracy approached the population R^2 of 0.80 for most methods under correct specification. LASSO (M $R^2 = 0.745$, SD = 0.011), elastic net (M $R^2 = 0.745$, SD = 0.011) and ridge (M $R^2 = 0.743$, SD = 0.011) achieved the highest accuracy with MCP (M $R^2 = 0.735$, SD = 0.012) performing nearly equivalently. Adaptive LASSO (M $R^2 = 0.736$, SD = 0.011) and SCAD (M $R^2 = 0.735$, SD = 0.012) showed marginally lower but still strong predictive performance. Compound misspecification reduced R^2 by 0.015 to 0.017 points across all methods with the declines nearly uniform regardless of estimator type. RMSE increased by approximately 0.099 to 0.104 under misspecification, indicating that the degradation in predictive accuracy was consistent and moderate in magnitude. Notably, the relative ranking of methods remained stable across misspecification conditions at this sample size with all methods exhibiting similar sensitivity to model violations.

Although the virtual similarity of Layer 1 and Layer 2 performance might suggest negligible impact of misspecification, we explicitly tested whether this impression could stem from insufficient statistical power. Using the 100 independent replications at $N = 10000$, we conducted Welch's t-tests comparing the out-of-sample R^2 and RMSE distributions between layers for each method. The degradation in R^2 (approximately 0.015) was highly statistically significant for all six methods (all $p < .001$) with Monte Carlo standard errors of about 0.001 which supported high power to detect even minor deviations. Moreover, despite the small absolute drop in R^2 on the 0-to-1 scale, the standardized effect sizes (Cohen's d) exceeded 1.20 for every method which corresponded to a large effect by conventional benchmarks. Importantly, the modest absolute degradation is intentional as we calibrated the misspecifications (e.g., interaction coefficients in [-0.1, 0.1]) to mirror the "dense-but-weak" signal structure typical of psychological data where interactions and nonlinearities exist but rarely eclipse main effects. Thus, the regularized models

Table 2 ■ Predictive Accuracy

N	Layer	Method	R^2 Mean	R^2 SD	RMSE Mean	RMSE SD
500	1	ridge	0.216	0.064	5.304	0.401
500	2	ridge	0.222	0.065	5.317	0.53
2000	1	ridge	0.409	0.034	4.615	0.158
2000	2	ridge	0.402	0.031	4.703	0.26
10000	1	ridge	0.743	0.011	3.067	0.052
10000	2	ridge	0.728	0.013	3.166	0.081
500	1	lasso	0.133	0.076	5.575	0.419
500	2	lasso	0.123	0.084	5.64	0.542
2000	1	lasso	0.533	0.039	4.099	0.148
2000	2	lasso	0.523	0.04	4.194	0.219
10000	1	lasso	0.745	0.011	3.056	0.05
10000	2	lasso	0.73	0.012	3.153	0.079
500	1	enet	0.163	0.078	5.477	0.412
500	2	enet	0.164	0.085	5.507	0.535
2000	1	enet	0.535	0.038	4.09	0.145
2000	2	enet	0.525	0.039	4.186	0.22
10000	1	enet	0.745	0.011	3.054	0.05
10000	2	enet	0.73	0.012	3.151	0.08
500	1	adalasso	0.036	0.138	5.859	0.445
500	2	adalasso	0.033	0.139	5.902	0.523
2000	1	adalasso	0.522	0.04	4.144	0.148
2000	2	adalasso	0.513	0.045	4.237	0.209
10000	1	adalasso	0.736	0.011	3.107	0.051
10000	2	adalasso	0.721	0.013	3.205	0.08
500	1	scad	0.101	0.059	5.68	0.422
500	2	scad	0.1	0.068	5.723	0.571
2000	1	scad	0.391	0.051	4.679	0.198
2000	2	scad	0.383	0.044	4.776	0.257
10000	1	scad	0.735	0.012	3.116	0.05
10000	2	scad	0.718	0.015	3.22	0.08
500	1	mcp	0.045	0.058	5.856	0.447
500	2	mcp	0.041	0.063	5.908	0.59
2000	1	mcp	0.365	0.055	4.78	0.207
2000	2	mcp	0.354	0.048	4.883	0.257
10000	1	mcp	0.735	0.012	3.116	0.05
10000	2	mcp	0.718	0.015	3.22	0.08

are successfully capturing the dominant linear signal as the desired behavior for robust predictive tools and the statistically significant and large-effect-size drops support that the misspecification is detectable and not merely noise.

Selection Stability

Selection stability (Table 3) which was measured by the mean Jaccard index across replicates varied dramatically across methods and sample sizes. At $n = 500$, stability was uniformly low, reflecting the difficulty of reliable variable selection in small samples with high-dimensional predictors. Adaptive LASSO demonstrated the highest stability (mean Jaccard = 0.103–0.104), followed closely by elastic net (0.093–0.096) and LASSO (0.060–0.062). SCAD and MCP showed the lowest stability (0.026–0.050) with MCP exhibiting particularly high variability (SD = 0.040–0.068). Compound misspecification at this sample size produced negligible changes in stability with differences in mean Jaccard indices being roughly 0.002 - 0.003.

At $n = 2000$, selection stability improved substantially for most methods. Elastic net achieved the highest stability (M Jaccard = 0.315–0.318), followed closely by LASSO (M = 0.296–0.299) and adaptive LASSO (M = 0.280). SCAD and MCP demonstrated moderate stability (M = 0.238–0.241), substantially improved from their performance at $n = 500$ but still below the convex methods. Compound misspecification produced minimal reductions in stability with mean Jaccard indices decreasing by 0.001 to 0.003 across all methods, indicating that the consistency of selected variable sets was largely unaffected by the introduced misspecifications

at this sample size.

At $n = 10000$, selection stability reached high levels for all methods although notable differences emerged. SCAD achieved the highest stability (M Jaccard = 0.914–0.932) with MCP showing similarly high stability (M = 0.897–0.924). Elastic net also demonstrated excellent stability (M = 0.727–0.740) and was followed by LASSO (M = 0.714–0.728). Adaptive LASSO exhibited somewhat lower stability (M = 0.570–0.583) despite its strong predictive performance. Compound misspecification reduced stability modestly for all methods with declines of 0.007 to 0.027 in mean Jaccard indices. The largest absolute reductions were observed for MCP ($\Delta = \breve{}0.027$) and SCAD ($\Delta = \breve{}0.017$), suggesting that the non-convex methods were slightly more sensitive to misspecification in their selection consistency than the convex methods while achieving the highest stability overall.

Discussion

The present study examined how compound model misspecification affects the predictive accuracy and selection stability of regularized regression methods across three sample sizes. Our results provide several insights for the robustness of these estimators when the fitted linear-additive model is systematically wrong which is a common scenario that reflects the complexity of psychological data (De Boeck et al., 2023; Turkheimer et al., 2021).

Regarding our first research question, compound misspecification produced surprisingly minimal degradation in predictive accuracy at the smallest sample size ($n = 500$) with R^2 differences not exceeding 0.010 for any method. At intermediate ($n = 2000$) and large ($n = 10000$) sample sizes, misspecification reduced R^2 by approximately 0.006–0.011 and 0.015–0.017 points respectively with corresponding increases in RMSE. These findings suggest that while misspecification does impair predictive performance, the magnitude of degradation is modest and remarkably consistent across estimators at larger sample sizes. The uniform sensitivity observed at $n = 10000$ indicates that when statistical information is abundant, all methods are similarly affected by violations of linearity, homoscedasticity and additivity.

Our second research question examined whether shrinkage-dominated methods exhibit greater robustness to misspecification. At $n = 500$, ridge regression indeed demonstrated superior robustness by achieving the highest predictive accuracy under both correct specification (M $R^2 = 0.216$) and misspecification (M $R^2 = 0.222$) with performance actually improving slightly under misspecification. This finding aligns with theoretical expectations that shrinkage estimators may be less vulnerable to overfitting when the model is misspecified (Hastie et al., 2009). However, at larger sample sizes, the hypothesized robustness advantage dissipated. At $n = 2000$, elastic net and LASSO substantially outperformed ridge and by $n = 10000$, all methods converged to similar predictive accuracy with nearly identical degradation under misspecification. This pattern suggests that the theoretical advantages of shrinkage-dominated methods for robustness are primarily realized in small-sample contexts where regularization serves a dual function of controlling both model complexity and sensitivity to misspecification.

Our third research question addressed the impact of misspecification on selection stability. At smaller sample sizes ($n = 500$ and $n = 2000$), compound misspecification produced negligible effects on Jaccard indices with differences of 0.001–0.003. This finding indicates that instability in variable selection at these sample sizes is driven primarily by limited statistical information rather than by model violations. However, at $n = 10000$ non-convex methods (SCAD and MCP) showed larger absolute reductions in stability under misspecification ($\Delta = \breve{}0.017$ to $\breve{}0.027$) compared to convex methods like LASSO and elastic net ($\Delta = \breve{}0.007$ to $\breve{}0.014$). This differential sensitivity suggests that the oracle properties of non-convex penalties (Fan & Li, 2001; Zhang, 2010) may entail a trade-off. When the fitted model is incorrect, their more aggressive selection mechanisms exhibit reduced consistency. Nevertheless, SCAD and MCP still achieved the highest absolute stability values (Jaccard = 0.897–0.932), indicating that even under misspecification, they select variables more consistently than convex alternatives.

The convergence of predictive performance across methods at large sample sizes warrants further consideration. At $n = 10000$, all estimators achieved R^2 values approaching the population value of 0.80 with nearly identical degradation under misspecification. This finding suggests that the bias introduced by different penalty structures becomes less consequential as statistical information increases which is consistent with asymptotic theory for regularized estimators (Fan & Li, 2001; Zou, 2006). That is, concerns about method selection for prediction may be secondary to ensuring adequate sample size.

The differential patterns observed for predictive accuracy and selection stability highlight the importance of multi-faceted evaluation (Baldassarre et al., 2017). At $n = 10000$, adaptive LASSO demonstrated strong predictive accuracy (M $R^2 = 0.736$) but substantially lower selection stability (Jaccard = 0.583) compared to other methods. This dissociation suggests that good prediction does not guarantee reliable variable identification which is a distinction with implications for research contexts where interpretation is paramount.

Our findings extend previous simulation research that

Table 3 ■ Selection Stability

N	Layer	Method	Jaccard Stability Mean	Jaccard Stability SD
500	1	ridge	NA	NA
500	2	ridge	NA	NA
2000	1	ridge	NA	NA
2000	2	ridge	NA	NA
10000	1	ridge	NA	NA
10000	2	ridge	NA	NA
500	1	lasso	0.062	0.019
500	2	lasso	0.06	0.02
2000	1	lasso	0.299	0.014
2000	2	lasso	0.296	0.014
10000	1	lasso	0.728	0.019
10000	2	lasso	0.714	0.014
500	1	enet	0.093	0.025
500	2	enet	0.096	0.023
2000	1	enet	0.318	0.02
2000	2	enet	0.315	0.021
10000	1	enet	0.74	0.018
10000	2	enet	0.727	0.017
500	1	adalasso	0.104	0.015
500	2	adalasso	0.103	0.014
2000	1	adalasso	0.28	0.013
2000	2	adalasso	0.28	0.013
10000	1	adalasso	0.583	0.014
10000	2	adalasso	0.57	0.015
500	1	scad	0.05	0.019
500	2	scad	0.048	0.02
2000	1	scad	0.24	0.018
2000	2	scad	0.238	0.018
10000	1	scad	0.932	0.031
10000	2	scad	0.914	0.044
500	1	mcp	0.026	0.04
500	2	mcp	0.025	0.068
2000	1	mcp	0.241	0.025
2000	2	mcp	0.239	0.026
10000	1	mcp	0.924	0.037
10000	2	mcp	0.897	0.056

has focused primarily on correct specification (Kipruto & Sauerbrei, 2025; Pavlou et al., 2015). We provide evidence that regularized methods maintain much of their utility even when assumptions are violated by demonstrating that compound misspecification produces only moderate degradation in predictive accuracy across estimators. The sample-size-dependent patterns we observed emphasize the importance of considering statistical power in high-dimensional settings (Paxton et al., 2001; Skrondal, 2000).

The modest impact of misspecification on selection stability at smaller samples when combined with the emergence of differential sensitivity at large samples suggests that the relationship between model violations and selection consistency is more complex than previously recognized. While non-convex methods achieve superior absolute stability when sample size is sufficient, their slightly greater vulnerability to misspecification warrants consideration when researchers suspect substantial violations of modeling assumptions.

Implications

The findings of the present study carry several implications for the application of regularized regression methods in psychological research.

The results demonstrate that compound misspecification produces only moderate degradation in predictive accuracy across all estimators particularly at larger sample sizes. This finding suggests that concerns about model misspecification may not preclude the use of regularized methods for prediction tasks in psychological research

(Fokkema et al., 2022; Yarkoni & Westfall, 2017). Even when the assumptions of linearity, homoscedasticity and additivity are violated which are the conditions that characterize much psychological data (De Boeck et al., 2023), these methods could maintain much of their predictive capacity. Applied researchers can therefore proceed with reasonable confidence when using regularized regression in contexts where the true data-generating process is unknown.

Method selection may be contingent upon sample size and research objectives rather than assumptions regarding differential robustness to misspecification. In small-sample contexts ($n = 500$ with $p = 2000$), ridge regression provides superior predictive accuracy and should be preferred when prediction constitutes the primary analytic goal. This recommendation aligns with the theoretical properties of ridge regression as a shrinkage estimator that handles multicollinearity effectively (Hoerl & Kennard, 1970) and extends them to misspecified conditions. At intermediate sample sizes ($n = 2000$), elastic net offers an advantageous balance between predictive accuracy and selection stability which is consistent with its design for handling correlated predictors (Zou & Zhang, 2009). At large sample sizes ($n = 10000$), estimators converge in performance which may permit researchers to select methods based on secondary considerations such as interpretability or computational efficiency.

As Yarkoni and Westfall (2017) argued, the primary path to improved prediction in psychology lies in larger samples and more systematic accumulation of evidence rather than methodological sophistication alone.

When variable selection stability represents a central research objective, for instance, in theory-development contexts where identifying replicable predictors is essential (Baldassarre et al., 2017), non-convex methods warrant consideration despite their marginally greater sensitivity to misspecification. At large sample sizes, SCAD and MCP achieve substantially higher Jaccard indices than convex alternatives which indicate greater confidence that selected variables would replicate across samples. This advantage stems from the oracle properties of non-convex penalties (Fan & Li, 2001; Zhang, 2010), which reduce bias for large coefficients while shrinking small coefficients to zero. However, at smaller samples where selection stability is uniformly low, researchers should exercise considerable caution in interpreting selected variable sets regardless of method choice as limited statistical information rather than model violations is the primary constraint (Paxton et al., 2001).

The minimal impact of misspecification on selection stability at smaller samples suggests that concerns regarding model violations affecting variable selection consistency may be secondary to concerns regarding inadequate sample size. Working with high-dimensional data may need to prioritize collecting larger samples over attempting to model every potential nonlinearity or interaction as increased statistical information improves both predictive accuracy and selection stability more reliably than correcting for misspecification. This conclusion aligns with established recommendations in the simulation literature (Skrondal, 2000) and reinforces the importance of statistical power considerations in high-dimensional settings.

The dissociation between predictive accuracy and selection stability observed for adaptive LASSO at large samples which is strong prediction but comparatively low stability highlights the importance of aligning method choice with research goals. For purely predictive applications where variable identification is not required, adaptive LASSO remains a viable option. However, for research aiming to identify reliable predictors for theory development or intervention targeting, methods with higher selection stability such as SCAD or MCP may be considered despite their slightly greater computational demands (Breheny & Huang, 2011).

Finally, the convergence of predictive performance across methods at large sample sizes suggests that the field's focus on identifying optimal estimators for prediction may be somewhat misplaced. When sample size is sufficient, all regularized methods perform similarly and degradation under misspecification is uniform. This finding implies that efforts to improve prediction in psychological research might be better directed toward increasing sample sizes and improving measurement quality rather than refining estimator selection. As Yarkoni and Westfall (2017) argued, the primary path to improved prediction in psychology lies in larger samples and more systematic accumulation of evidence rather than methodological sophistication alone.

Limitations and Future Directions

Several limitations of the present study suggest directions for future research. First, we examined compound misspecification rather than isolating individual violations. While this approach enhances ecological validity, it precludes conclusions about which specific assumption violations most affect performance. Future research should manipulate each misspecification layer independently to identify differential sensitivities. Our dense-but-weak sig-

nal structure and fixed within-block correlation ($\rho = 0.4$) represent only one point in a larger parameter space. Different results might be obtained under strict sparsity, alternative correlation structures or more severe misspecification magnitudes. Future investigations should systematically vary these parameters to map the conditions under which specific methods demonstrate superior robustness. In addition, we examined a set of regularized estimators, but this set is not exhaustive. Bayesian methods, ensemble approaches and algorithms for non-linear relationships may exhibit different robustness properties and warrant investigation. Our evaluation focused on predictive accuracy and selection stability. Future work should incorporate additional criteria such as computational efficiency, calibration of prediction intervals and accuracy of coefficient recovery to provide more comprehensive assessment. All conditions maintained $p = 2,000$ predictors. Research should examine whether these patterns generalize to the $p > n$ regime common in some psychological applications (Jacobucci et al., 2016; McNeish, 2015). Finally, replication using semi-synthetic designs with real predictor data would provide valuable evidence regarding external validity.

Conclusions

The present study examined how compound misspecification affects regularized regression methods across sample sizes from 500 to 10,000. Misspecification produced modest and uniform degradation in predictive accuracy across estimators particularly at larger samples where all methods showed nearly identical R^2 reductions of 0.015–0.017 points. Ridge regression's robustness advantage is limited to small samples ($n = 500$). At larger samples, selection-oriented methods performed equivalently or better with elastic net and LASSO achieving highest accuracy at $n = 2000$ and convergence across all methods at $n = 10,000$. Misspecification minimally affected selection stability at smaller samples where statistical information is the primary constraint. At large samples, non-convex methods (SCAD, MCP) achieve highest absolute stability but exhibit slightly greater sensitivity to misspecification than convex alternatives. Applied researchers could use regularized methods with reasonable confidence under misspecification. Method selection should be guided by sample size such as ridge for small-sample prediction, elastic net for intermediate samples and non-convex methods when selection stability is paramount at large samples.

References

Ægisdóttir, S., White, M. J., Spengler, P. M., Maugherman, A. S., Anderson, L. A., Cook, R. S., Nichols, C. N., Lampropoulos, G. K., Walker, B. S., Cohen, G., & Rush, J. D. (2006). The meta-analysis of clinical judgment project: Fifty-six years of accumulated research on clinical versus statistical prediction. *The Counseling Psychologist*, *34*(3), 341–382. doi: 10.1177/0011000005285875.

Aiken, L. S., & West, S. G. (1991). *Multiple regression: Testing and interpreting interactions* [Sage Publications, Inc]. https://psycnet.apa.org/record/1991-97932-000

Baldassarre, L., Pontil, M., & Mourão-Miranda, J. (2017). Sparsity is better with stability: Combining accuracy and stability for model selection in brain decoding. *Frontiers in Neuroscience*, *11*, 1–15. doi: 10.3389/fnins.2017.00062.

Beaujean, A. A. (2017). Simulating data for clinical research: A tutorial. *Journal of Psychoeducational Assessment*, *36*(1), 7–20. doi: 10.1177/0734282917690302.

Benson, N. F. (2017). Introduction to a special issue on simulation studies as a means of informing psychoeducational testing and assessment. *Journal of Psychoeducational Assessment*, *36*(1), 3–6. doi: 10.1177/0734282917728236.

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *The Annals of Applied Statistics*, *5*(1), 232–253. doi: 10.1214/10-aoas388.

Bühlmann, P., & van de Geer, S. (2011). *Statistics for high-dimensional data: Methods, theory and applications* (1st ed.). Springer Berlin. doi: 10.1007/978-3-642-20192-9.

Daniel, F., Ooi, H., Calaway, R., Microsoft Corporation, & Weston, S. (2022). *Foreach: Provides foreach looping construct* [R Package]. Retrieved August 4, 2026, from https://doi.org/10.32614/CRAN.package.foreach

De Boeck, P., Pek, J., Walton, K., Wegener, D. T., Turner, B. M., Andersen, B. L., Beauchaine, T. P., Lecavalier, L., Myung, J. I., & Petty, R. E. (2023). Questioning psychological constructs: Current issues and proposed changes. *Psychological Inquiry*, *34*(4), 239–257. doi: 10.1080/1047840x.2023.2274429.

Dwyer, D. B., Falkai, P., & Koutsouleris, N. (2018). Machine learning approaches for clinical psychology and psychiatry. *Annual Review of Clinical Psychology*, *14*(1), 91–118. doi: 10.1146/annurev-clinpsy-032816-045037.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *The Annals of Statistics*, *32*(2), 407–499. doi: 10.1214/009053604000000067.

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi: 10.1198/016214501753382273.

Fokkema, M., Iliescu, D., Greiff, S., & Ziegler, M. (2022). Machine learning and prediction in psychological assess-

ment. *European Journal of Psychological Assessment*, *38*(3), 165–175. doi: 10.1027/1015-5759/a000714.

Folashade, D., Microsoft Corporation, W. S., & Tenenbaum, D. (2022). *Doparallel: Foreach parallel adaptor for the 'parallel' package* [R Package; Statistical Software]. Retrieved August 4, 2026, from https://doi.org/10.32614/CRAN.package.doParallel

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi: 10.18637/jss.v033.i01.

Grove, W. M., Zald, D. H., Lebow, B. S., Snitz, B. E., & Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment*, *12*(1), 19–30. doi: 10.1037/1040-3590.12.1.19.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.). Springer. doi: 10.1007/978-0-387-84858-7.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67. doi: 10.1080/00401706.1970.10488634.

Iliescu, D., Ion, A., Ilie, A., Ispas, D., & Butucescu, A. (2023). The incremental validity of personality over time in predicting job performance, voluntary turnover, and career success in high-stakes contexts - a longitudinal study. *Personality and Individual Differences*, *213*, 1–23. doi: 10.1016/j.paid.2023.112288.

Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling*, *23*(4), 555–566. doi: 10.1080/10705511.2016.1154793.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2021). *An introduction to statistical learning: With applications in r* (2nd ed.). Springer. doi: 10.1007/978-1-0716-1418-1.

Janković, M., Van Boxtel, G., & Bogaerts, S. (2022). Violent recidivism and adverse childhood experiences in forensic psychiatric patients with impaired intellectual functioning. *International Journal of Offender Therapy and Comparative Criminology*, *68*(13-14), 1357–1379. doi: 10.1177/0306624x221133013.

Jia, J., Rohe, K., & Yu, . B. (2013). The lasso under poisson-like heteroscedasticity. *Statistica Sinica*, *23*(1), 99–118. doi: 10.5705/ss.2010.254.

Kayanan, M., & Wijekoon, P. (2019). Performance of lasso and elastic net estimators in misspecified linear regression model. *Ceylon Journal of Science*, *48*(3), 293. doi: 10.4038/cjs.v48i3.7654.

Kipruto, E., & Sauerbrei, W. (2025). Evaluating prediction performance: A simulation study comparing penalized and classical variable selection methods in low-dimensional data. *Applied Sciences*, *15*(13), 7443. doi: 10.3390/app15137443.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. doi: 10.18637/jss.v028.i05.

Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling*. Springer. doi: 10.1007/978-1-4614-6849-3.

Lavelle-Hill, R., Smith, G., Deininger, H., & Murayama, K. (2025). An explainable artificial intelligence handbook for psychologists: Methods, opportunities, and challenges. *Psychological Methods*, *Advance online publication*. doi: 10.1037/met0000772.

Lei, J., G'Sell, M., Rinaldo, A., Tibshirani, R. J., & Wasserman, L. (2018). Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, *113*(523), 1094–1111. doi: 10.1080/01621459.2017.1307116.

Mayer-Kress, G., Liu, Y., & Newell, K. M. (2006). Complex systems and human movement. *Complexity*, *12*(2), 40–51. doi: 10.1002/cplx.20151.

McNeish, D. M. (2015). Using lasso for predictor selection and to assuage overfitting: A method long overlooked in behavioral sciences. *Multivariate Behavioral Research*, *50*(5), 471–484. doi: 10.1080/00273171.2015.1036965.

Meehl, P. E. (1954). *Clinical versus statistical prediction: A theoretical analysis and a review of the evidence*. University of Minnesota Press. doi: 10.1037/11281-000.

Meehl, P. E. (1990). Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, *66*(1), 195–244. doi: 10.2466/pr0.1990.66.1.195.

Moulden, H. M., Matthew, S. A., & Chaimowitz, G. (2024). Enhancing pre-treatment motivation improves forensic mental health outcomes. *International Journal of Offender Therapy and Comparative Criminology*, *70*(2-3), 150–166. doi: 10.1177/0306624x241228229.

Ostojic, D., Lalousis, P. A., Donohoe, G., & Morris, D. W. (2024). The challenges of using machine learning models in psychiatric research and clinical practice. *European Neuropsychopharmacology*, *88*, 53–65. doi: 10.1016/j.euroneuro.2024.08.005.

Pavlou, M., Ambler, G., Seaman, S., De Iorio, M., & Omar, R. Z. (2015). Review and evaluation of penalised regression methods for risk prediction in low-dimensional data with few events. *Statistics in Medicine*, *35*(7), 1159–1177. doi: 10.1002/sim.6782.

Paxton, P., Curran, P. J., Bollen, K. A., Kirby, J., & Chen, F. (2001). Monte carlo experiments: Design and implementation. *Structural Equation Modeling*, *8*(2), 287–312. doi: 10.1207/s15328007sem0802_7.

Pelech, W. (2011). Complex systems and human behavior. *Social Work Education*, *30*(8), 1023–1024. doi: 10.1080/02615479.2011.554224.

R Core Team. (2026). *R: A language and environment for statistical computing. r foundation for statistical computing*. https://www.R-project.org

Raudenbush, S. W., & Bryk, A. S. (2002). *Hierarchical linear models: Applications and data analysis methods* [Sage Publications, Inc]. https://us.sagepub.com/en-us/nam/hierarchical-linear-models/book9230

Richardson, K., & Norgate, S. H. (2015). Does iq really predict job performance? *Applied Developmental Science*, *19*(3), 153–169. doi: 10.1080/10888691.2014.983635.

Skrondal, A. (2000). Design and analysis of monte carlo experiments: Attacking the conventional wisdom. *Multivariate Behavioral Research*, *35*(2), 137–167. doi: 10.1207/s15327906mbr3502_1.

Song, J., Pan, F., & Liu, H. (2025). Income, psychological security, and subjective well-being in urban china: A machine learning analysis with shap interpretation. *BMC Psychology*, *13*(1). doi: 10.1186/s40359-025-03591-2.

Ter Braak, C. J. (2009). Regression by l1 regularization of smart contrasts and sums (roscas) beats pls and elastic net in latent variable model. *Journal of Chemometrics*, *23*(5), 217–228. doi: 10.1002/cem.1213.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288. doi: 10.1111/j.2517-6161.1996.tb02080.x.

Turkheimer, F. E., Rosas, F. E., Dipasquale, O., Martins, D., Fagerholm, E. D., Expert, P., Váša, F., Lord, L., & Leech, R. (2021). A complex systems perspective on neuroimaging studies of behavior and its disorders. *The Neuroscientist*, *28*(4), 382–399. doi: 10.1177/1073858421994784.

Wu, C. F., & Hamada, M. S. (2021). *Experiments: Planning, analysis, and optimization*. John Wiley & Sons. doi: 10.1002/9781119470007.

Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*(6), 1100–1122. doi: 10.1177/1745691617693393.

Yuan, S., De Roover, K., Dufner, M., Denissen, J. J., & Van Deun, K. (2019). Revealing subgroups that differ in common and distinctive variation in multi-block data: Clusterwise sparse simultaneous component analysis. *Social Science Computer Review*, *39*(5), 802–820. doi: 10.1177/0894439319888449.

Zhang, C. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. doi: 10.1214/09-aos729.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, *101*(476), 1418–1429. doi: 10.1198/016214506000000735.

Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *67*(2), 301–320. doi: 10.1111/j.1467-9868.2005.00503.x.

Zou, H., & Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *The Annals of Statistics*, *37*(4), 1733–1751. doi: 10.1214/08-aos625.

Open practices

The *Open Data* badge was earned because the data of the experiment(s) are available on github.com/onurramazan/Ramazan-Lui/blob/main/RScriptforDataSimulation

The *Open Material* badge was earned because supplementary material(s) are available on github.com/onurramazan/Ramazan-Lui/blob/main/RScriptforDataAnalysis

Citation

Ramazan, O., & Lui, Y. W. (2026). Robustness of regularized regression methods under compound model misspecification: A simulation benchmarking study. *The Quantitative Methods for Psychology*, *22*(2), 90–102. doi: 10.20982/tqmp.22.2.p090.

Received: 14/02/2026 ~ Accepted: 31/07/2026