

SEM-Based Out-of-Sample Predictions

Mark de Rooij, Julian D. Karch, Marjolein Fokkema, Zsuzsa Bakk, Bunga Citra Pratiwi and Henk Kelderman

Leiden University

ABSTRACT

Predictive modeling is becoming more popular in psychological science. Machine learning techniques have been used to develop prediction rules based on items of psychological tests. However, this approach does not take into account that these items are noisy indicators of the constructs they intend to measure. Structural equation modeling does take this into account. Several authors have concluded that it is impossible to make out-of-sample predictions based on a reflective structural equation model. We show that it is possible to make such predictions and we develop R-code to do so. With two empirical examples, we show that SEM-based prediction can outperform prediction based on linear regression models. With three simulation studies, we further investigate the SEM-based prediction rule and its robustness in comparison with predictions using regularized linear regression. We conclude that the new SEM-based prediction rule is robust against violation of the normality assumption but sensitive to model misspecification.

KEYWORDS

Cross-validation; machine learning; prediction rule; regression; structural equation modeling

1. Introduction

Since the turn of the century, there has been an increased awareness of the difference between explanatory statistical analysis and predictive statistical analysis. Much of the discussion originated with the paper by Breiman (2001b) and thereafter by Shmueli (2010). Shmueli and colleagues (Shmueli, 2010; Shmueli et al., 2016; Shmueli & Koppius, 2011) emphasized the importance of predictive modeling beyond building purely predictive models: Evaluation of predictive performance can help to improve existing theory or to drive new theory, to develop and evaluate measures, to improve construct operationalization, or to compare theories. Yarkoni and Westfall (2017) emphasize the importance of predictive modeling in the context of the *reproducibility crisis* in psychology (Open Science Collaboration, 2015).

The goal of predictive modeling is to predict the output value of new cases by applying the model parameters estimated from one data sample to generate predictions for cases outside of that sample. That is, based on the estimated parameters a prediction rule is defined, which is a mathematical function in which values of predictor variables can be inserted and a predicted value of the outcome variable rolls out. Fundamental to a proper predictive procedure is the ability to predict measurable information on new cases, that are *out-of-sample*. To assess predictive performance, the data are often partitioned into a training or calibration set and a

test set. This can be done repeatedly, as in *K*-fold cross-validation (de Rooij & Weeda, 2020). The biggest threat to out-of-sample predictive performance is to fit the model too tightly on the training set, which is known as overfitting. Overfitting is inherently related to the *bias-variance trade-off* (Hastie et al., 2009), where more complex models tend to have less bias but more variance and simpler models tend to have more bias and less variance. Complexity is often defined in terms of the number of parameters that need to be estimated, such that models with more parameters are more complex. Bias also arises if a model does not allow for the true functional form of relationships, for example, a linear model cannot completely accommodate non-linear relationships, and this results in bias. If two models both can represent the true functional forms, both models will have roughly equal bias, but the simpler model will have better predictive performance because it needs less parameters to be estimated and therefore will have lower variance. We will see examples of the latter in our simulation experiments.

Recently, this framework of developing prediction rules has been applied to data from psychological tests. Seeboth et al. (2018) hypothesized, for example, that personality questionnaire items could increase prediction strength in comparison to big five scores (i.e., scale scores). Single questionnaire items can be understood as *personality markers*, analogous to genetic markers in medical research, related to health outcomes. In such investigations, the items measuring

psychological constructs are taken as predictors for some outcome or response variable. Seeboth et al. (2018) use life outcomes, such as sleeping enough, optimism, general health, or depression, as response variables.

More specifically, Seeboth et al. (2018) measured big-five personality characteristics. In total there are 50 personality items, measuring Openness, Conscientiousness, Extraversion, Agreeableness, and Neuroticism. In more mathematical terms they have observations x_{ip} for $i = 1, \dots, n$ and $p = 1, \dots, P$ where $P=50$. The regression model based on the *item scores* as predictors is defined by

$$y_i = \beta_0 + \sum_{p \in \mathcal{P}} \beta_p x_{ip} + \epsilon_i,$$

where $\mathcal{P}$ is the set of all 50 predictor items. The regression model based on the *scale scores* is defined as

$$y_i = \beta_0 + \sum_{s=1}^5 \beta_s \left(\sum_{p \in \mathcal{P}_s} x_{ip} \right) + \epsilon_i$$

where first the items belonging to one of the five personality scales, indicated by the set $\mathcal{P}_s$, are summed. The regression models are estimated using ordinary least squares or penalized estimation methods, such as LASSO (Tibshirani, 1996), ridge (Hoerl & Kennard, 1970), or the elastic net algorithm (Zou & Hastie, 2005). The penalized regression models, sometimes called *regularized* regression models, actively trade-off bias and variance, that is, by penalizing the coefficients the model becomes biased but less variable diminishing prediction error.

One of the assumptions of linear regression models is, however, that predictor variables are fixed or measured without error (Fox, 2016). In the examples mentioned in the previous paragraph, the predictor variables x_{ip} are noisy indicators for the latent variables. Structural Equation Models (SEM) were developed to deal with measurement error. For that purpose, measurement models were combined with structural (i.e., regression) models into one overarching model to deal with measurement error in the predictor variables in an optimal way. As a simplified example, consider the situation where we have three items measuring, say, conscientiousness (first latent variable) and three items measuring, say, neuroticism (second latent variable) plus one outcome variable, for example, depression.

The structural equation model for this situation is graphically depicted in Figure 1. In the left-hand side of Figure 1, the outcome variable is a scale score which might be obtained by a (weighted) sum of itemscores. Note that in some cases the outcome variable is also measured through a set of indicators, which we will then denote by y_{ir} for $r = 1, \dots, R$ (in the univariate case $R=1$ and we only have a single response variable). Depression can, for example, be measured using a set of items. The structural equation model can be simply adapted to such a situation. On the right-hand side of Figure 1, we show a more complicated model where now also the response variable is a latent variable measured by several indicators.

The first three items are indicators of the first latent variable (ξ_1) and the others for the second latent variable (ξ_2). A linear relationship is assumed between latent variables and their indicators, which is quantified through the factor loadings (λ_{pq}). The two latent variables are assumed to be predictors of the outcome variable Y (Figure 1a) or ξ_y (Figure 1b) with regression weight β_q . Note that these two models are so-called *reflective models*, where all manifest variables are endogenous.

These types of structural equation models are widely used for explanation in psychology and other scientific fields. However, it is not entirely clear how we can make out-of-sample predictions, that is, how to predict the output value for new cases (i.e., the y_{ir} for new subject i) for whom we have scores on the items/predictors (i.e., the x_{ip} for new subject i) using the model estimated from the current sample. Shmueli et al (2016) call that the biggest elephant in the room: that predictive performance in SEM means being able to predict new data at the item-level, that is we like create a rule for the y_{ir} .

In this paper, we focus on *operative* prediction, which entails using items of exogenous constructs (i.e., the x_{ip}) to predict items of endogenous constructs (i.e., the y_{ir}), and doing so by using the full measurement and structural properties posited by the estimated model. Operative prediction, therefore, requires manifest items as predictors and outcomes. Evermann and Tate (2016) write

“In the context of predictive modeling, structural equation models can present unique challenges. In predictive modeling, the values

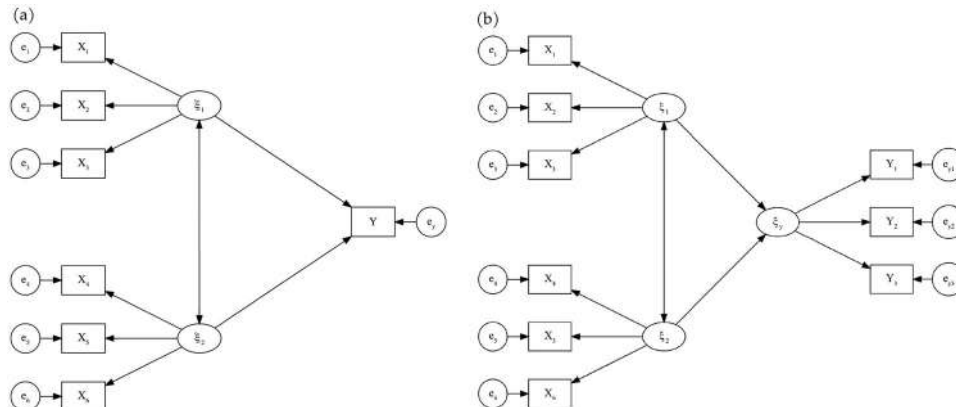


Figure 1. Two fully reflective structural equation models for which it is thought prediction is impossible. On the left, a reflective model with observed outcome variable. On the right, a reflective model with a latent outcome variable measured by a set of indicators.

for a new case are predicted from the predictor variables for that case. However, many structural equation models in the management disciplines are specified as fully reflective [...] with all manifest variables as endogenous and only latent variables as exogenous. Hence, no manifest predictors exist from which values for new cases can be predicted [...]. In the terminology of Shmueli et al. (2016), operative prediction is not possible from a fully reflective model."

Also in Hair et al. (2017) and Sarstedt et al. (2016), it is concluded that predictions based on a reflective structural equation model are not feasible and they recommend to use Partial Least Squares approaches instead. This premise is a belief of (some) researchers in the field of PLS, but it is (as we will show) not completely factual.

In this paper, we will show that it is possible to make out-of-sample predictions based on a fitted reflective structural equation model. We will develop our theory in the next section, where we also present an R function to make predictions based on a fitted SEM using Lavaan (Rosseel, 2012). We will confine ourselves to the situation that all variables are continuous and assumed to be normally distributed. In Section 3, we will show how to use the theory and corresponding R function to make out-of-sample predictions in two empirical data sets. Furthermore, we show how the quality of the SEM-based prediction rule can be evaluated using repeated K -fold cross-validation. We end this paper with some discussion.

2. Covariance Based Structural Equation Modeling

We assume that the variables can be partitioned into two sets, one set with the predictor variables and one with the response or outcome variables. In SEM it is assumed that the variables follow a joint multivariate normal distribution where the predictor variables have means μ_{x_p} (collected in the vector μ_x) and the response variables have means μ_{y_r} (collected in the vector μ_y). The covariance matrix of the predictors and responses (Σ) can be partitioned as follows

$$\Sigma = \begin{bmatrix} \Sigma_{xx} & \Sigma_{xy} \\ \Sigma_{yx} & \Sigma_{yy} \end{bmatrix}$$

where Σ_{yy} is the $R \times R$ covariance matrix of the responses, Σ_{xx} is the $P \times P$ covariance matrix of the predictors, and Σ_{xy} contains the covariances between predictors and responses, and is of size $P \times R$.

2.1. Estimation

First, let us define the parameter vector θ including the means, factor loadings, and regression weights. To obtain an estimate of θ often maximum likelihood estimation is used, although also other techniques, such as generalized least squares, weighted least squares, or unweighted least squares can be used. Let such a loss function, negative likelihood or least squares, be denoted by $\mathcal{L}(\mathbf{X}, \mathbf{Y}, \theta)$, where $\mathbf{X}$ and $\mathbf{Y}$ are data matrices, then the parameter estimates are given by

$$\hat{\theta} = \operatorname{argmin}_{\theta} \mathcal{L}(\mathbf{X}, \mathbf{Y}, \theta),$$

that is, the arguments that minimize the loss function.

2.2. Predictions for a New Observation

Suppose, we have fitted our SEM to an empirical data set to obtain the estimated means, factor loadings, and regression weights. Based on the estimated parameters ($\hat{\theta}$), the estimated model-implied mean vector ($\mu(\hat{\theta})$) and covariance matrix ($\Sigma(\hat{\theta})$) can be computed. In many software applications, such as Lavaan, functions are provided to extract those from the fitted model.

From a predictive modeling perspective, SEM fitting functions estimate the joint distribution of the predictors and the responses. Within the statistical learning framework, such a method that estimates the joint distribution is called a generative model (Ng & Jordan, 2001). Predictions from generative models are obtained via the model-implied conditional distribution of the responses given the predictors, known as the *predictive distribution*. Suppose, we have a new observation with predictor values $\mathbf{x}_0$. For a fitted SEM, the predictive distribution is a (multivariate) normal distribution with mean vector $\mu_{y|x_0}(\hat{\theta})$ and covariance matrix $\Sigma_{y|x_0}(\hat{\theta})$ as follows

$$\begin{aligned} \hat{\mu}_{y|x_0} &= \hat{\mu}_y + \hat{\Sigma}_{xy} \hat{\Sigma}_{xx}^{-1} (\mathbf{x}_0 - \hat{\mu}_x); \\ \hat{\Sigma}_{y|x_0} &= \hat{\Sigma}_{yy} - \hat{\Sigma}_{xy} \hat{\Sigma}_{xx}^{-1} \hat{\Sigma}_{xy}, \end{aligned} \quad (1)$$

where we have dropped the dependence of all estimates on the estimated parameters $\hat{\theta}$ for notational unclutteredness (i.e., we used $\hat{\Sigma} = \Sigma(\hat{\theta})$ and similar for the means). This predictive distribution (Equations 1) follows directly from the rules for conditioning a normal distribution, see, for example, Eaton (1983, Section 3.4) or Bishop (2006, Section 2.3.1).

The most popular approach to obtain a point prediction is given by the mean of the predictive distribution. For a normal distribution, the SEM-based prediction rule is $\hat{\mathbf{y}} = \hat{\mu}_{y|x_0}$, which can be rewritten as

$$\hat{\mathbf{y}} = \hat{\mu}_y + \hat{\Gamma}(\mathbf{x}_0 - \hat{\mu}_x) = \underbrace{\hat{\mu}_y - \hat{\Gamma} \hat{\mu}_x}_{=\alpha} + \hat{\Gamma} \mathbf{x}_0 = \hat{\alpha} + \hat{\Gamma} \mathbf{x}_0,$$

where

$$\hat{\Gamma} = \hat{\Sigma}_{xy} \hat{\Sigma}_{xx}^{-1}.$$

Thus, the SEM-based prediction for each response y_r is a multiple linear regression model with intercept $\hat{\alpha}_r$ and regression weight vector $\hat{\gamma}_r$, where $\hat{\gamma}_r$ denotes the r th row of $\hat{\Gamma}$. The crucial difference is that the intercept and regression weights are estimated differently compared to ordinary (least squares) regression, namely by taking into account the structure defined in the SEM.

Another perspective on the relationship between SEM-based predictions and multiple regression is revealed by the fact that SEM-based predictions contain ordinary least squares multiple linear regression predictions as a special case. If one uses any saturated SEM, the model-implied covariance is identical to the sample covariance, and as a result, the SEM-based predictions are identical to ordinary least squares predictions. Therefore, SEM-based prediction

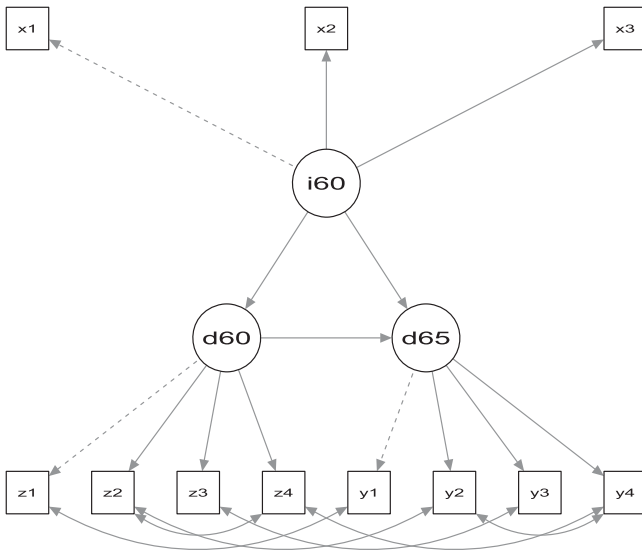


Figure 2. SEM for Political Democracy Data.

may have more bias but less variance than multiple linear regression.

This also reveals another perspective on the potential utility of SEM-based predictions. While multiple linear regression does not make any assumptions about the covariance matrix, SEM-based predictions allow harnessing the capability of SEM to restrict the covariance matrix based on prior knowledge, such as which items belong to the same scale. Incorporating this knowledge might improve the estimate of the covariance matrix and, consequently, the predictions.

2.3. R Code

We created the R function `predicty.lavaan` that computes predicted values for a set of response variables given a set of predictor values for a new person and an output object from the `sem` function in Lavaan. The function can be found on the github page of the first author (<https://github.com/mjderooij/SEMpredict>). With the following two lines of code, a researcher can fit a structural equation model to a training data set and make predictions for a test data set.

```
fit = sem(mymodel , data = train , meanstructure = TRUE)
yhat = predicty . lavaan (fit , newdata = test,
xnames = xn , ynames = yn)
```

In the prediction function, the variables denoted by `xnames` are treated as predictor variables, and the variables denoted by `ynames` as responses.

3. Empirical Examples

3.1. Political Democracy Data

Our first data set concerns the Political Democracy data, a data set well-known and available in the Lavaan package.

The Political Democracy data (see Bollen, 1989, p. 12, 17, 36, 228, and 321) contains various measures of

political democracy and industrialization in developing countries. There are three latent variables, Industrialization measured in 1960 (*i60*) for which there are three indicators available (denoted by x_1, x_2, x_3 in the data set), political democracy in 1960 (*d60*) with four indicators (denoted by z_1, z_2, z_3, z_4), and political democracy in 1965 (*d65*) also with four indicators (denoted by y_1, y_2, y_3, y_4). The data set has $n = 75$ observations. We consider the political democracy items in 1965 as the responses and the other variables as the predictors. The usual SEM fitted to these data (as in the Lavaan tutorial) is graphically represented in Figure 2. Fitting the SEM model to the complete data, we find the following fit indices: $\chi^2 = 38.1$ with 35 df, CFI = 0.995, RMSEA = 0.035, and SRMR = 0.041, all fit the Hu and Bentler (1999) criteria (CFI 0.95, RMSEA 0.06, SMRR 0.08).

We will show how to use the prediction rule in a repeated K -fold cross-validation setting and compare the predictive performance of SEM against that of multivariate linear regression. In K -fold cross-validation, the data is partitioned in K parts, where cyclically one part is left out as the test set, and the other $K-1$ parts are used as the training set. In the end, for every observation in the data set, there is a predicted value. The predicted values can be compared to the observed values using a measure of prediction error. This K -fold cross-validation can be used to compare the predictive performance of multiple approaches, where the approach that has the smallest prediction error is considered best. In the partitioning of the data, there is randomness which might influence which approach is considered best. To avoid this randomness, often the K -fold cross-validation is repeated a number, say 100, of times. This will give each approach 100 values of prediction error, that show which approaches on average predicts best, the variability of prediction errors over repetitions, and the percentage of times that a certain approach is considered best. The 100 prediction error values for each approach can be visualized using boxplots. For more details about repeated K -fold cross-validation see de Rooij and Weeda (2020).

For this data set, we consider two approaches: the SEM-based approach to prediction and an approach based on multivariate multiple linear regression. For the SEM-based approach, we fit on the training set the model as defined in Figure 2. For prediction in the test sample we use variables x_1, x_2, x_3 and z_1, z_2, z_3, z_4 as predictors and y_1, y_2, y_3, y_4 as outcomes. In the multivariate linear regression model we also use x_1, x_2, x_3 and z_1, z_2, z_3, z_4 as predictors and y_1, y_2, y_3, y_4 as outcomes. As a measure of predictive performance, we use the Root Mean Squared Error of prediction (RMSEp), defined as

$$\text{RMSEp} = \sqrt{\frac{1}{n \times R} \sum_{i=1}^n \sum_{r=1}^R (y_{ir} - \hat{y}_{ir})^2}, \quad (2)$$

where $R = 4$ for this data set.

The results of the 100 times 10-fold cross-validation are shown in the boxplots of Figure 3, which clearly shows that for this data set the SEM-based prediction rule outperforms the rule from multivariate linear regression. The average



prediction error is smaller for the SEM-based approach and the SEM-based approach wins in all 100 repetitions. The difference in prediction error between the two approaches is small but consistent.

3.2. CERQ Data

Garnefski and Kraaij (2007) describe a questionnaire for the assessment of cognitive emotion regulation, the CERQ. The CERQ measures nine conceptually different subscales of emotion regulation: self-blame, other-blame, rumination, catastrophizing, putting into perspective positive refocusing, positive reappraisal, acceptance, and planning. Each subscale is measured by four five-point Likert items. Garnefski and Kraaij (2007) have two criterion variables measured at follow-up, one year later: depressive and anxiety symptoms as measured by the SCL-90. We will use the depression subscale which is measured by 16 items, each with possible scores between 1 and 5. The data that we use have 240 observations.

With the SEM approach, we can either use the scale score (i.e., the sum of the 16 item scores) of depression as the outcome variable or the 16 item scores (corresponding to Figures 1a and b, respectively). Using the scale score follows the approach taken by Seeboth et al. (2018) who also have a univariate outcome. A scale score, however, assumes that the factor loadings for the items belonging to depression are all equal. Furthermore, such an approach does not give insight into which items are good or bad predicted. In the following two subsections, we first use the scale score as an outcome and afterward the item scores as outcomes.

3.2.1. Prediction of Depression Scale Scores

In this subsection, we use the scale score of depression, with a theoretical range from 16 to 80, as the outcome variable. For the SEM-based approach, this corresponds to Figure 1a. Fitting the SEM model to the complete data, we find the

following fit statistics: $\chi^2 = 1196.3$ with 585 df, CFI = 0.867, RMSEA = 0.066, and SRMR = 0.072. Although only the SRMR fits the Hu and Bentler's (1999) criteria (CFI = 0.95, RMSEA = 0.06, SMRR = 0.08), possibly due to the small sample size (Hu & Bentler, 1999), we expect this model to contain enough prior information about the structure of the data for our purpose.

First, we fit an SEM on the data with nine latent variables, each having four indicators and one observed response variable (approach 1). Second, we first compute scale scores on the predictor side and use these nine scores in a linear regression model (approach 2). Finally, we fitted lasso regression (approach 3), elastic net regression with lasso-ridge mixing parameter of 0.1 till 0.9 (approaches 4 till 12), ridge regression (approach 13), and ordinary least squares regression (approach 14) using the 36 indicators as predictor variables with the depression score as an outcome. Like in the previous section we use 100 times repeated 10-fold cross-validation. For the regularization methods (approaches 3–13), we use a nested cross-validation scheme, where we determine the optimal penalty parameter using another round of 10-fold cross-validation in the training set.

The 100 RMSEP measure for each of these fourteen approaches can be seen in Figure 4. It is clear that the SEM-based prediction rule and the rule using the scale scores outperform the regularized regression models. Comparing approaches 1 and 2 we can conclude that, for these data, these give highly similar prediction error measures. If we compare from cycle to cycle of the 100 repetitions, the SEM-based prediction rule performs better in 83 of the 100 cases.

3.2.2. Prediction of Depression Item Scores

Instead of predicting the depression scale score, we can also predict at the item level. That is, for each of the 16 items measuring depression, we can obtain predicted values. For the SEM-based approach, this corresponds to Figure 1b.

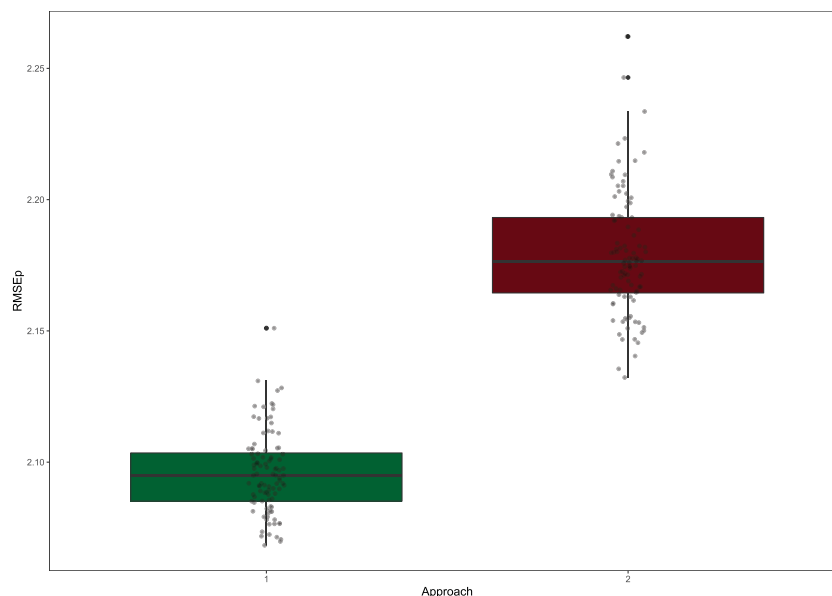


Figure 3. Cross validated prediction error from the SEM-based approach (Green, 1) and from multivariate multiple linear regression (Red, 2).

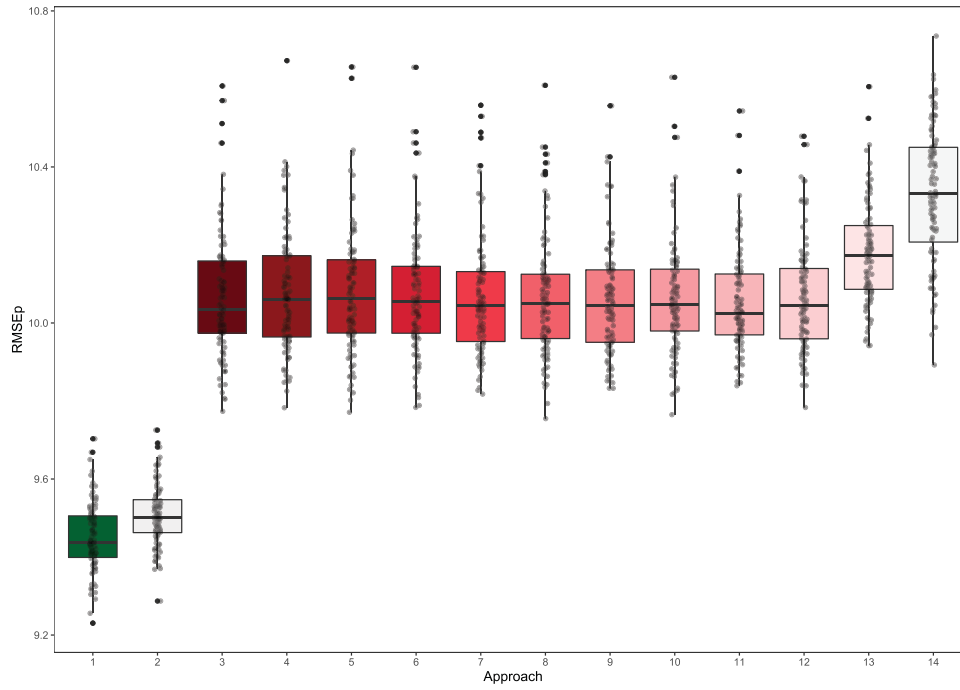


Figure 4. Cross validated prediction error from 14 different approaches on the CERQ data with predictions of the depression scale. Approach 1 corresponds to the SEM-based prediction rule.

Fitting the SEM model to the complete data, we find the following fit statistics: $\chi^2 = 2250.5$ with 1229 df, CFI = 0.863, RMSEA = 0.059, and SRMR = 0.069, where SRMR and RMSEA satisfy the Hu and Bentler criteria.

We compare the same approaches as in the previous section, where now all regression models become multivariate regression models with 16 outcome variables.

In this case, we can define two types of prediction error measures: a global one and an item level one. The global RMSEP is defined by Equation (2) where we sum over all 16 items. An item level prediction error for item r ($r = 1, \dots, 16$) is given by

$$\text{RMSEP}(r) = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_{ir} - \hat{y}_{ir})^2}. \quad (3)$$

The results of 100 times repeated 10-fold cross-validation for the global RMSEP can be seen in Figure 5, which shows a pattern that is highly similar to the pattern observed in the prediction of the scale scores. SEM-based prediction outperforms all other ways of making predictions.

In Appendix A we show the results on item level. Overall, the pattern is very similar to the global RMSEP measure where the SEM-based predictions are best for all items.

4. Simulation Studies

In the empirical data examples, we have seen that the SEM-based prediction rule can outperform prediction rules obtained from elastic net regression or linear least squares regression. In this section, we will perform three simulation studies to further compare SEM-based predictions with those of elastic net regression.

4.1. Design

In the first simulation study, we generate data following the assumptions of a structural equation model. This is the most favorable situation for the SEM-based prediction. As the basis for data generation, we follow the empirical analysis of the CERQ data, with item scores as outcome variables (as in Figure 1b). In more detail, in the structural model, we have in total ten latent variables of which nine are considered on the predictor side (denoted by ξ_x) and one on the response side (ξ_y). The nine latent variables are drawn from a multivariate normal distribution with zero means, variances equal to one, and correlations equal to the estimated correlations between the factor scores in the empirical example (Section 3.2.2). In the measurement model, each of the nine latent variables has four indicator items, the single response latent variable has sixteen indicators. So, in total there are $9 \times 4 + 1 \times 16 = 52$ observed (generated) variables.

We generate training and test data where we manipulate the strength of the measurement model, the strength of the structural model, and the sample size of the training set. The test data set has a constant size, that is, 1,000, and has the same characteristics as the training data set. In the simulation study, we manipulated the following factors

M: Measurement Model is a factor with two levels, weak and strong, and defines the strength of the relationships between the nine latent variables (ξ_x) and their indicators. In the weak condition, the factor loadings are drawn from a uniform distribution ranging from 0.20 to 0.50; in the strong condition, they are drawn from a uniform distribution ranging from 0.50 to 0.80. The residual variances of the items are $1 - \lambda^2$, where λ is the factor loading. The item means are drawn from a uniform distribution with a minimum of 1.50 and a maximum of 3.00.

S: Structural Model is a factor with two levels, weak and strong. In the weak condition, the regression weights are drawn from a

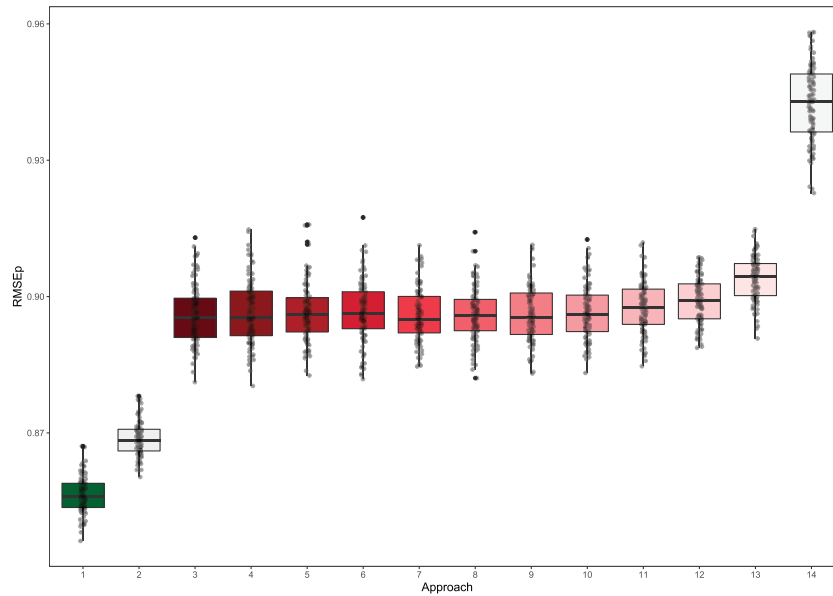


Figure 5. Cross validated prediction error from 14 different approaches on the CERQ data with predictions of the depression items. Approach 1 corresponds to the SEM-based prediction rule.

uniform distribution ranging from 0.15 to 0.25 leading to on average 25% explained variance in ξ_y ; in the strong condition, they are drawn from a uniform distribution ranging from 0.25 to 0.40 leading to on average 68% explained variance. The signs of the regression weights are the same as the signs of the weights in the empirical example.

N: Sample size is a factor with four levels. The training sample sizes are 100, 200, 500, and 1,000 reflecting a broad range of possibilities. Note that these are the sizes of the training set, the test set has a fixed size of 1,000 observations. In the training set, we find the optimal prediction rule for both the structural equation model as well as the elastic net model. For the elastic net model, this involves tuning both the penalty parameter as well as the parameter that balances between the lasso and ridge penalty. We do this in a 10-fold cross-validation regime. With the optimal parameters, we define the prediction rule for the test set.

The factor loadings for the items of the latent response variable are sampled from a uniform distribution in the range 0.20–0.70, the item means are drawn from a uniform distribution in the range 1.15–2.15 (based on the empirical example). In the first simulation study, we fit the correct structural equation model to the generated data. Therefore, this is a study where all the assumptions of the structural equation model are satisfied.

In the second simulation study, we change the distribution of the nine latent variables (ξ_x). First, we draw them from a normal distribution and subsequently take the exponent and then compute z-scores from the latter. In this way, we obtain highly skewed latent variables. The skewness of the latent variables translates into the skewness of the indicators. In this second simulation study, we investigate the robustness of our prediction rule with respect to a violation of the normality assumption. Note that the elastic net regression model does not assume a particular distribution of the predictor variables, while the structural equation model does.

In the data generation model for the third simulation study, we add direct effects from the item indicators to the

latent response variable. In our fitted structural equation model, these effects are not present. In this simulation study, we, therefore, investigate the robustness of our prediction rule with respect to model misspecification. For each of the ξ_x there is one indicator that has a direct effect on ξ_y . As in the first simulation study, we investigate a weak and strong measurement model. The structural effects between the latent variables are only weak (like in simulation study 1 drawn from a uniform distribution with a minimum of 0.15 and a maximum of 0.25). For the direct effects, we have two levels weak and strong. In the weak condition, the regression weights are sampled from a uniform distribution with a range of 0.10–0.15 and a randomly sampled sign. In the strong condition, the regression weights are sampled from a uniform distribution with a range of 0.15–0.25 and a randomly sampled sign. The weak condition leads to on average 40% explained variance in ξ_y , whereas the strong condition leads to 59% explained variance.

4.2. Analysis

In each of the three simulation studies, for each of the 16 conditions, we do 100 replications. On each generated data set we fit a structural equation model and a multivariate elastic net regression. The multivariate elastic net model has two regularization parameters: the usual penalty parameter determining the strength of the penalty and the parameters that balance between the lasso and ridge penalty. In a 10-fold cross-validation scheme, we determine the optimal pair of parameters in the training set and with this optimal pair make predictions on the test set.

Based on the fitted models in each replication, we define the prediction rule which we subsequently apply to a test data set. We evaluate the predictive performance using the global RMSEP measure, as defined in Equation (2). We will graphically represent the results by boxplots that summarize

the 100 RMSEP measures from the repetitions. Furthermore, we will use a three between one within the analysis of the variance procedure with prediction error as the response variable (RMSEP). The ANOVA results are only used to pinpoint which effects are large, where we consider effects with $\eta_p^2 > 0.10$ to be large. In the discussion, we will focus on the large effects involving the prediction model (i.e., SEM-based or Elastic Net-based). ANOVA tables are given in [Appendix B](#).

4.3. Results Simulation Study 1

In [Figure 6](#), we present the results of the first simulation study, where the fitted structural equation model equals the data generation model. ANOVA results ([Appendix B](#)) show that large effect sizes occur for the main effect of the model, the three two-way interaction effects between prediction model and measurement model, structural model, and sample size, and the three-way interaction effect of sample size, measurement model and prediction model. The main conclusion from this simulation study results' is that there is no condition in which the prediction error from the elastic net is smaller on average than that of SEM. The interactions point out that differences between prediction errors of elastic net and SEM are largest for smaller sample sizes, stronger measurement models, and stronger structural models.

4.4. Results Simulation Study 2

In [Figure 7](#), we present the results of the second simulation study, where the normality assumption of the structural equation model does not hold.

The ANOVA results ([Appendix B](#)) show that similar to Simulation study 1, large effect sizes occur for the main effect of the model, three two-way interaction effects between the prediction model with measurement model, structural model, and sample size, and the three-way interaction effect of sample size, measurement model and prediction model. The results from this simulation study show that in all but one condition the SEM-based prediction error is smaller on average than that of the elastic net. Only in the condition with a small sample size ($N=100$) and weak measurement model the average prediction error elastic net is smaller than the prediction error of SEM. The interactions point out that differences between prediction errors of elastic net and SEM are largest for smaller sample sizes, stronger measurement models, and stronger structural models. Overall, the results are very similar to Simulation study 1, indicating that the SEM-based prediction rule is robust with respect to violations of the normality assumption.

4.5. Results Simulation Study 3

In [Figure 8](#), we present the results of the third simulation study, where the fitted structural equation model is

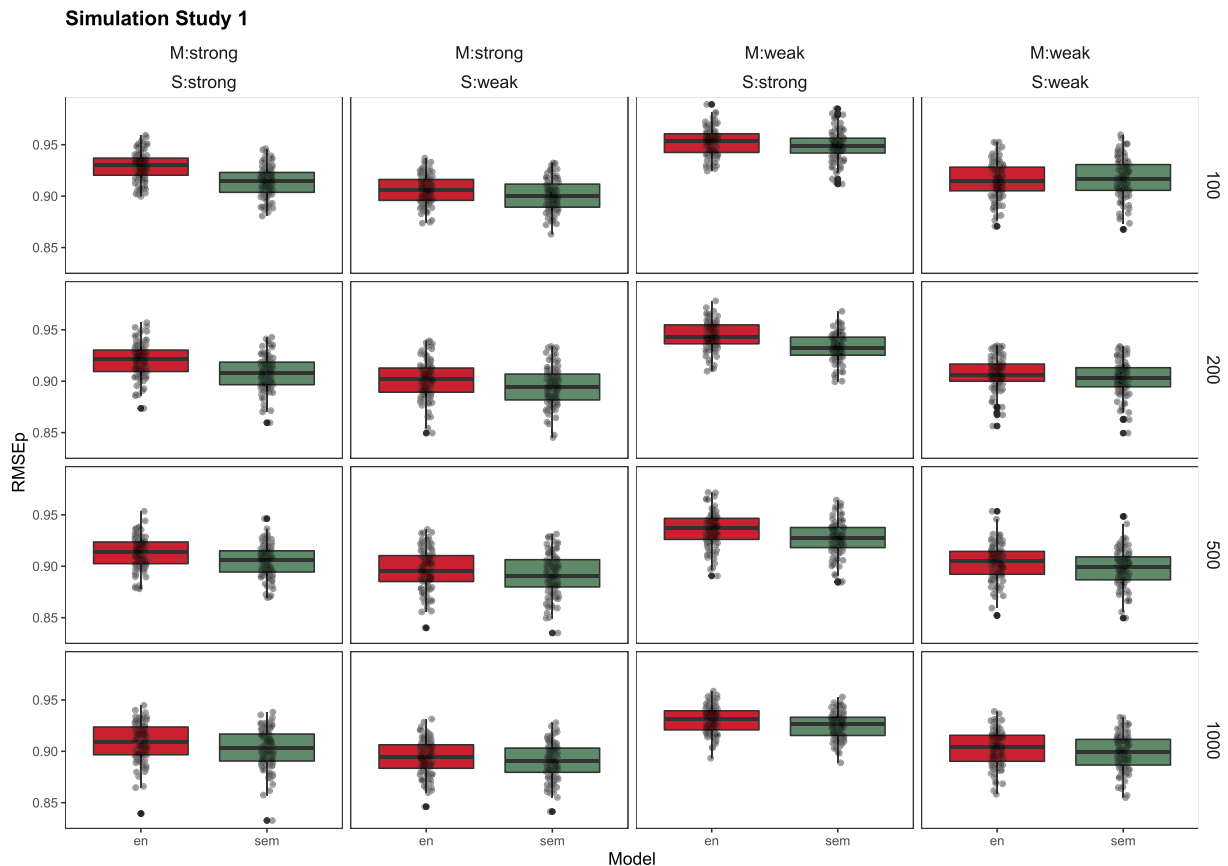


Figure 6. Results from simulation study 1 comparing the prediction error from the SEM-based rule (sem) and a rule obtained from elastic net regression (en). In simulation study 1, the data generation process equals the structural equation model. The rows represent the different sample sizes. The columns are indexed by measurement model (M) and structural model (S) and whether these effects are weak or strong in the data generation process.

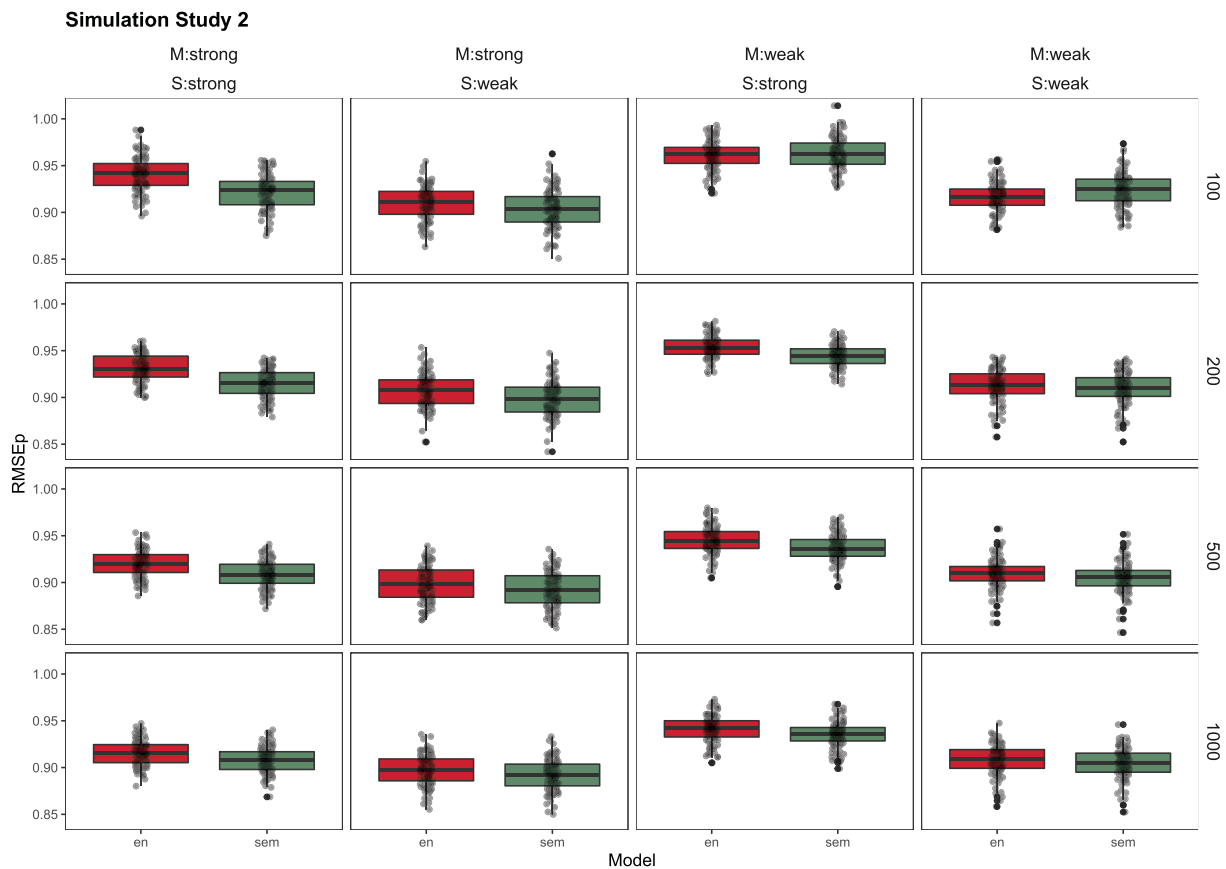


Figure 7. Results from simulation study 2 comparing the prediction error from the SEM-based rule (sem) and a rule obtained from elastic net regression (en). In simulation study 2, the latent variables have a very skew distribution which translates in a skew distribution of the items. Therefore, the assumptions of the structural equation model are violated. The rows represent the different sample sizes. The columns are indexed by measurement model (M) and structural model (S) and whether each is weak or strong.

misspecified. ANOVA results (Appendix B) show that there are large effect sizes for the main effect of the prediction model and the two-way interactions of sample size and structural model with the prediction model.

On average overall conditions, elastic net predicts better than SEM. When sample sizes are small ($N=100, 200$), SEM-based predictions are more accurate than elastic net predictions, whereas elastic net predictions are better for the larger sample sizes. With weak direct effects, SEM predictions are on average better, although this difference is small, whereas in the case of strong direct effects elastic net predictions are better. The results of this simulation study show that the SEM-based prediction rule is sensitive with respect to model misspecification except in the case of small samples.

5. Conclusion

Predictive modeling is becoming more popular in many scientific fields. The goal of predictive modeling is to predict the output value of new cases by applying the model parameters estimated from one data sample to generate predictions for individual cases outside of that sample.

The main result of this paper is that we showed that it is possible to make predictions for new cases based on a fully reflective covariance-based SEM. This is in contrast with the current status in the literature that seems to imply this is

impossible (Evermann & Tate, 2016; Hair et al., 2017; Sarstedt et al., 2016; Shmueli et al., 2016). We derived a general prediction rule for SEM, developed R-code to be used with fitted Lavaan objects, and showed that the SEM-based prediction rule is often more accurate than both standard linear regression and the commonly recommended elastic net machine learning approach.

In particular, SEM-based predictions were more accurate than regression-based predictions for the first example. This result can be understood from the bias-variance trade-off perspective. As the regression model equals a saturated structural equation model, the structural equation model has less parameters. A model with less parameters has generally more bias but less variance. This is especially true for small sample sizes, as with large sample sizes the variance reduces. In the first example, there are only 75 observations and therefore it pays off to make the model simpler (constrain parameters) to reduce the variance.

Our second data set is more in line with recent psychological research. Like Seeboth et al. (2018), we find in our second example that using the items as predictors performs better than using the scale scores. However, unlike Seeboth et al. (2018), we find that making an assumption about the structure of the items, instead of a data-driven way of deriving a prediction rule pays off. There are two main differences between our example with the CERQ data and the results presented in Seeboth et al. (2018): (1) We have a

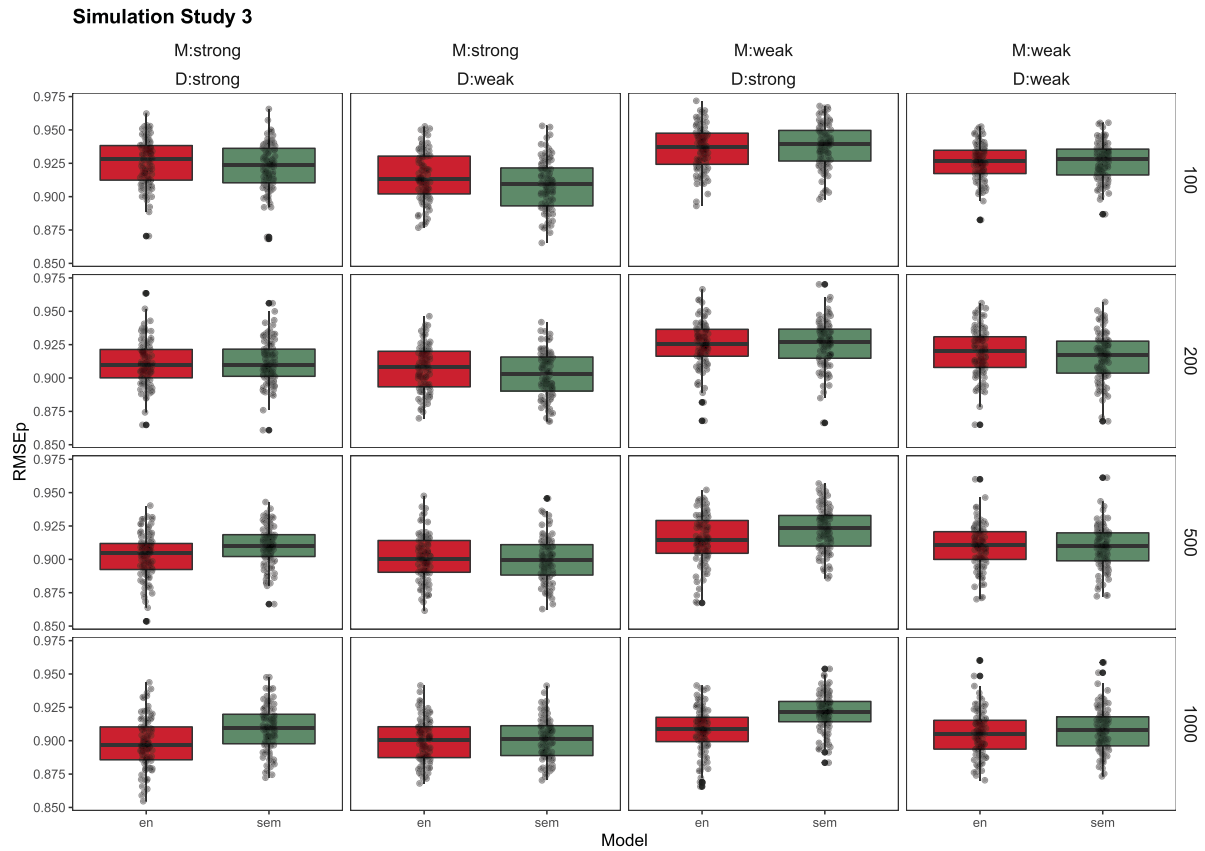


Figure 8. Results from simulation study 3 comparing the prediction error from two rules: The SEM-based rule (sem) and a rule obtained from elastic net regression (en). In simulation study 3, there are direct effects from the indicators on the predictor side of the model to the latent response variable. Therefore, the structural equation model is misspecified. The rows represent the different sample sizes. The columns are indexed by measurement model (M) and direct effects (D) and whether each is weak or strong.

sample size of 240, whereas the sample size in their study was 8719; (2) We have four items for each of the nine constructs, whereas they have 10 items for each of the five constructs. The large sample size in their study makes the role of the variance in the bias-variance trade-off really small, whereas in our relatively small sample size the variance plays a large role in the predictive accuracy.

We performed three simulation studies to further investigate the performance of our SEM-based rule in comparison with an elastic net-based rule. The data generation for the simulation studies was guided by the second empirical example. In the first simulation study, the data were generated following an SEM and the same SEM was fitted to the data. In all conditions, the SEM-based predictions were more accurate than the elastic net-based predictions. In the second simulation study, we investigated the robustness of the SEM-based prediction rule with respect to deviations from the normality assumption. We found that the SEM-based prediction rule is robust with respect to deviations from the normality assumption. Results were very similar to the results of the first simulation study, and overall the SEM-based prediction rule outperformed the elastic net-based rule. In the third simulation study, we investigated the effect of model misspecification. That is, in the generated data we added direct effects of the predictors to the latent response variable. These effects were not present in the fitted structural equation model. In this case, the elastic

net outperformed the SEM-based predictions, although in small sample sizes the SEM-based predictions were still better.

Cross-validation for covariance-based structural equation models has previously been considered by Browne (2000), Browne and Cudeck (1989), and Cudeck and Browne (1983). These papers, however, use the covariance matrix in the training and test set to define measures of prediction error. Therefore, these methods cannot be used for assessing predictive performance for individual cases. Furthermore, although the predictive performance of several SEMs can be compared using such an approach, it does not allow for the comparison of SEM-based rules with for example predictions from regression models or more general machine learning methods like Random Forests (Breiman, 2001a). With the prediction rule we developed, out-of-sample predictions can be made in the usual cross-validation sense. Therefore, with this rule, it becomes possible to actually compare the predictive performance of SEM with other approaches, such as Partial Least Squares predictive modeling or other statistical learning approaches like Random Forests (Breiman, 2001a). The SEM-based prediction rule may make realistic assumptions about the structure of the data. Techniques like Random Forests are completely data-driven and are expected to work well with reasonably large sample sizes. The SEM-based approach makes assumptions and with these assumptions, if tenable, we obtain predictive

power. Browne (2000) concludes that a proper SEM model will cross-validate better than a saturated model (which is equivalent to multiple regression) when the sample size is small to medium and the model specification error of the SEM model is not too large. These conclusions concur with our conclusions.

We developed R-code to be used together with SEMs fitted using the Lavaan package (Rosseel, 2012). The Lavaan package itself also has a predict function, called lavPredict. The documentation of this function warns that the function cannot be used to predict values of dependent variables, given the values of independent values (in the regression sense). Instead, the function gives predicted values of the latent variables. In contrast, our predicty.lavaan function can be used to predict values of the dependent variables, given the values of independent values in the regression sense. Therefore, it can be viewed as a useful addition to the Lavaan package.

Although predictions can be made for any fitted SEM, of course, a researcher has to think about whether such a prediction makes sense. If the fitted model is, for example, a pure measurement model (say a two-factor model), it might not be sensible to predict the indicators of one factor from those of the other factor. There should be some natural division in predictors and responses. In the applications we showed, and also for the applications discussed in Seeboth et al. (2018) it does make sense to develop an SEM-based prediction rule.

In all our analyses, we used maximum likelihood estimation. When certain assumptions are not mine, other estimation methods are considered more appropriate. Least-squares fitting approaches, for example, might be more robust to departures from normality. Besides maximum likelihood, weighted and unweighted least squares, two-step approaches (Croon, 2002; Skrondal & Laake, 2001) where the estimation of the measurement model and structural model is separated might also be considered. Different estimation methods result in different estimated implied covariance matrices, thereby directly influencing our prediction rule. It would be of interest to consider different estimation approaches and their effect on the accuracy of the SEM-based prediction rule.

We did not discuss model selection for structural equation models, but our prediction rule in combination with K -fold cross-validation can be used to compare the predictive performance of several structural equation models. This would give yet another measure to compare various SEMs with each other, but one that signals overfitting.

Recently, there have been developments with regularization within the SEM context (Arruda & Bentler, 2017; Huang et al., 2017; Jacobucci et al., 2016; Jacobucci & Grimm, 2018). Often the penalty parameters for these regularized SEM are chosen based on information criteria. In the appropriate context, we can choose the optimal value of the penalty parameter based on out-of-sample predictions using our prediction rule, as is the standard approach in regularized regression models.

We confined ourselves to continuous variables assumed to be normally distributed. For categorical variables the prediction rule that we developed needs adaptation. The model-implied joint distribution is not the multivariate normal distribution anymore, and hence deriving the conditional distribution as we did in Equation (1) is not straightforward. Further research is needed for models with categorical variables.

In sum, in the SEM-based prediction approach we buy predictive power by making assumptions, that is, assumptions about the structure of our variables and how they are related. With these assumptions, we add bias but decrease variance as shown in the examples. The payoff is expected with regular sample sizes often seen in psychology, such as in our second example.

References

- Arruda, E. H., & Bentler, P. M. (2017). A regularized GLS for structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 24, 657–665.
- Bishop, C. M. (2006). *Pattern recognition and machine learning* (1st ed.). Springer.
- Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.
- Breiman, L. (2001a). Random forests. *Machine Learning*, 45, 5–32.
- Breiman, L. (2001b). Statistical modeling: The two cultures. *Statistical Science*, 16, 199–231.

- Browne, M. W. (2000). Cross-validation methods. *Journal of Mathematical Psychology*, 44, 108–132.
- Browne, M. W., & Cudeck, R. (1989). Single sample cross-validation indices for covariance structures. *Multivariate Behavioral Research*, 24, 445–455.
- Croon, M. (2002). Using predicted latent scores in general latent structure models. In G. Marcoulides & I. Moustaki (Eds.), *Latent variable and latent structure models* (pp. 207–236). Psychology Press.
- Cudeck, R., & Browne, M. W. (1983). Cross-validation of covariance structures. *Multivariate Behavioral Research*, 18, 147–167.
- de Rooij, M., & Weeda, W. (2020). Cross-validation: A method every psychologist should know. *Advances in Methods and Practices in Psychological Science*, 3, 248–263.
- Eaton, M. L. (1983). *Multivariate statistics: A vector space approach*. John Wiley & Sons.
- Evermann, J., & Tate, M. (2016). Assessing the predictive performance of structural equation model estimators. *Journal of Business Research*, 69, 4565–4582.
- Fox, J. (2016). *Applied regression analysis & generalized linear models*. Sage Publications, Inc.
- Garnefski, N., & Kraaij, V. (2007). The cognitive emotion regulation questionnaire: Psychometric features and prospective relationships with depression and anxiety in adults. *European Journal of Psychological Assessment*, 23, 141–149.
- Hair, J. F., Matthews, L. M., Matthews, R. L., & Sarstedt, M. (2017). PLS-SEM or CB-SEM: Updated guidelines on which method to use. *International Journal of Multivariate Data Analysis*, 1, 107–123.
- Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). Springer.
- Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12, 55–67.
- Hu, L. T., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, 6, 1–15.
- Huang, P.-H., Chen, H., & Weng, L.-J. (2017). A penalized likelihood method for structural equation modeling. *Psychometrika*, 82, 329–354.
- Jacobucci, R., & Grimm, K. J. (2018). Comparison of frequentist and Bayesian regularization in structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 25, 639–649.
- Jacobucci, R., Grimm, K. J., & McArdle, J. J. (2016). Regularized structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal*, 23, 555–566.
- Ng, A. Y., & Jordan, M. I. (2001). On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. *Advances in neural information processing systems*, 14, 841–848.
- Open Science Collaboration (2015). Estimating the reproducibility of psychological science. *Science*, 349, aac4716(1)–aac4716(8).
- Rosseel, Y. (2012). Lavaan: An R package for structural equation modeling. *Journal of Statistical Software*, 48, 1–36.
- Sarstedt, M., Hair, J. F., Ringle, C. M., Thiele, K. O., & Gudergan, S. P. (2016). Estimation issues with PLS and CBSEM: Where the bias lies! *Journal of Business Research*, 69, 3998–4010.
- Seeboth, A., Möttus, R., & Kandler, C. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, 32, 186–201.
- Shmueli, G. (2010). To explain or to predict. *Statistical Science*, 25, 289–310.
- Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, 35, 553–572.
- Shmueli, G., Ray, S., Estrada, J. M. V., & Chatla, S. B. (2016). The elephant in the room: Predictive performance of PLS models. *Journal of Business Research*, 69, 4552–4564.
- Skrondal, A., & Laake, P. (2001). Regression among factor scores. *Psychometrika*, 66, 563–575.
- Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58, 267–288.
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, 12, 1100–1122.
- Zou, H., & Hastie, T. (2005). Regularization and variable selection via the elastic-net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67, 301–320.

Appendix A: Prediction errors on item level for CERQ data

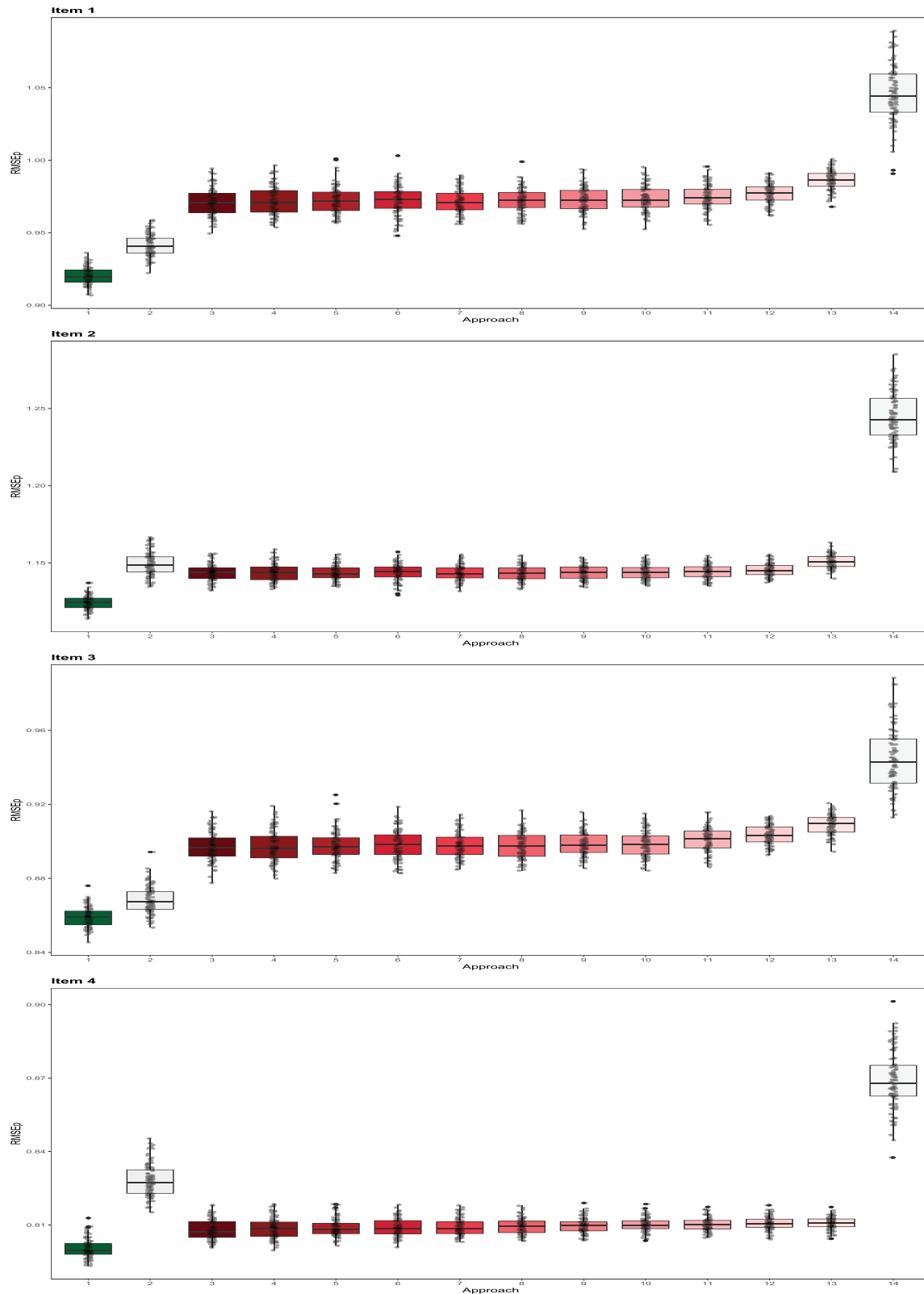


Figure 9. Cross validated predictions for Items 1 till 4 from 14 different approaches on the CERQ data.

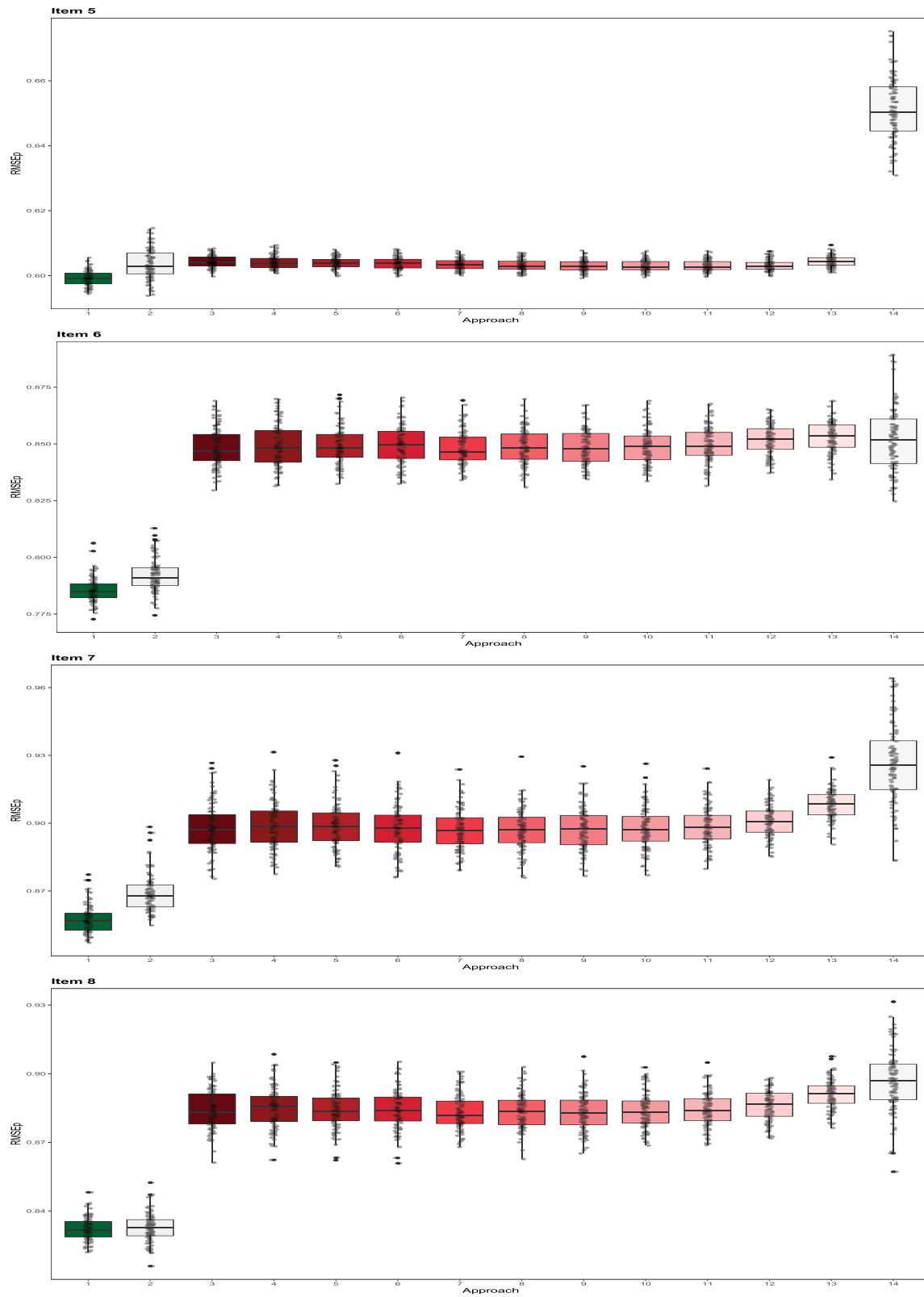


Figure 10. Cross validated predictions for Items 5 till 8 from 14 different approaches on the CERQ data.

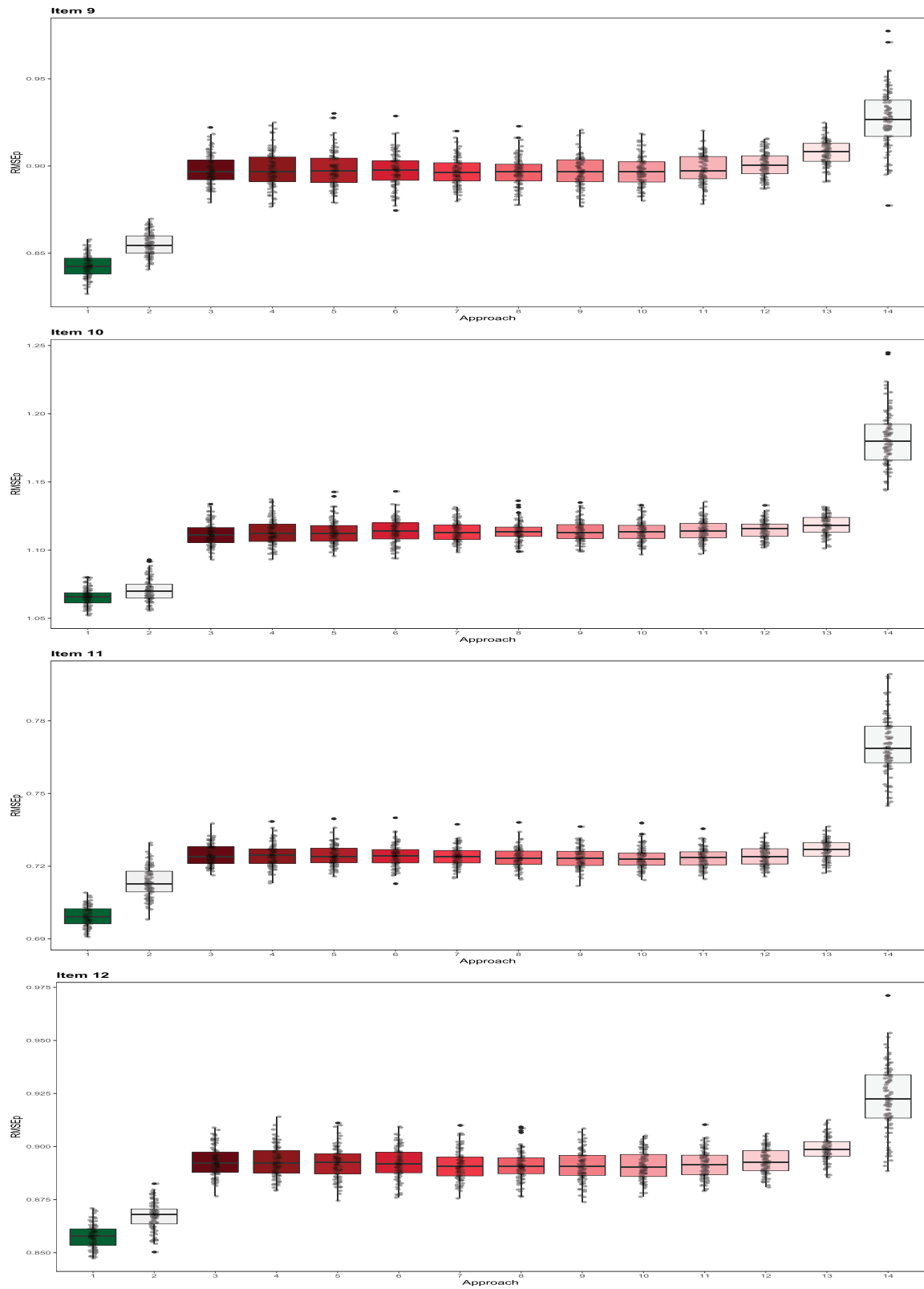


Figure 11. Cross validated predictions for Items 9 till 12 from 14 different approaches on the CERQ data.

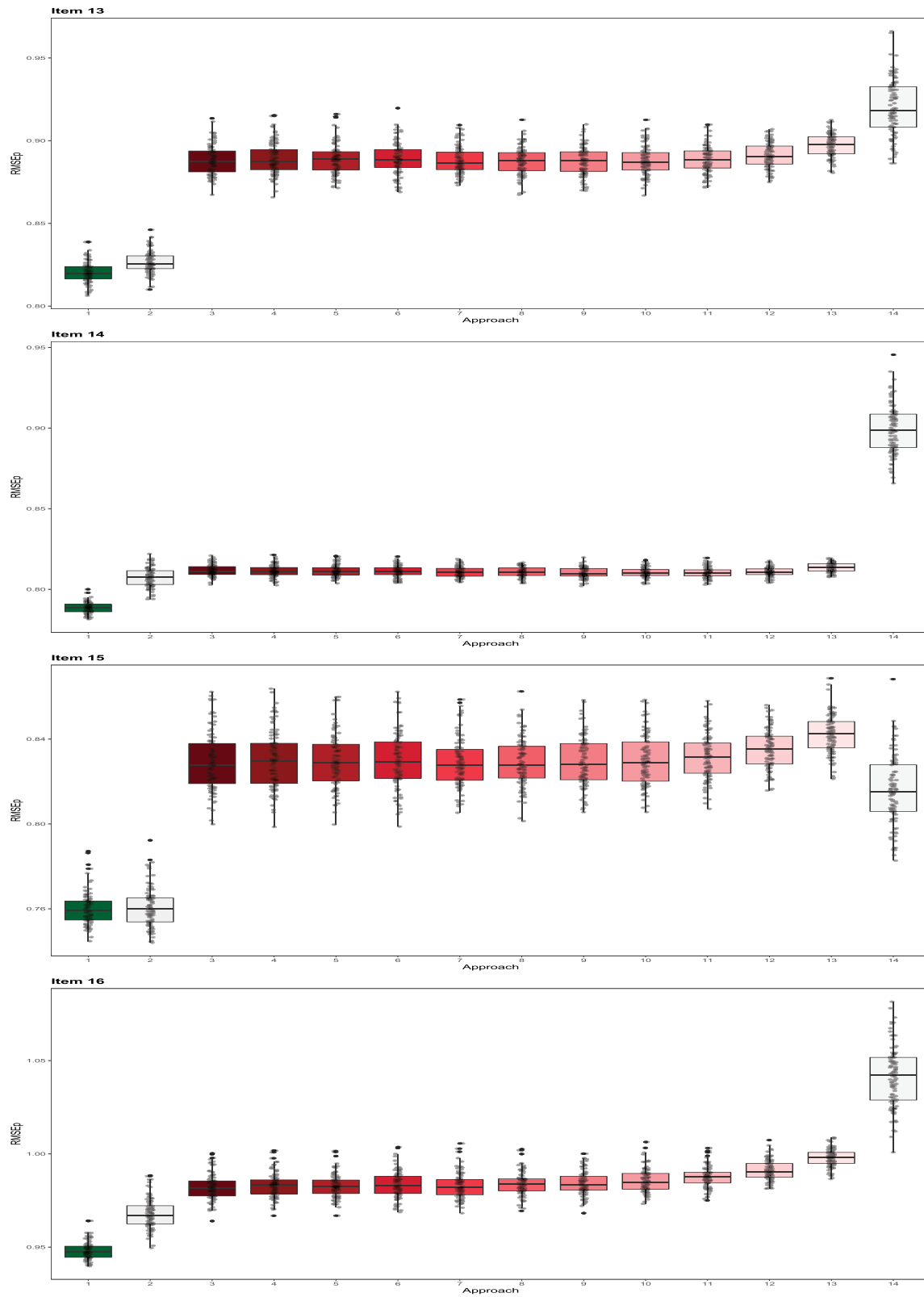


Figure 12. Cross validated predictions for Items 12 till 16 from 14 different approaches on the CERQ data.

Appendix B: Anova results from simulations

Table 1. ANOVA results simulation study 1.

	Effect	DFn	DFd	<i>F</i>	<i>p</i>	<i>p</i> < .05	η_p^2
1	N	3.00	1,584.00	80.05	0.00	*	0.13
2	M	1.00	1,584.00	433.56	0.00	*	0.21
3	S	1.00	1,584.00	825.22	0.00	*	0.34
4	model	1.00	1,584.00	5,579.66	0.00	*	0.78
5	N:M	3.00	1,584.00	3.31	0.02	*	0.01
6	N:S	3.00	1,584.00	2.53	0.06		0.01
7	M:S	1.00	1,584.00	93.41	0.00	*	0.06
8	N:model	3.00	1,584.00	93.10	0.00	*	0.15
9	M:model	1.00	1,584.00	289.08	0.00	*	0.15
10	S:model	1.00	1,584.00	850.42	0.00	*	0.35
11	N:M:S	3.00	1,584.00	0.38	0.77		0.00
12	N:M:model	3.00	1,584.00	181.27	0.00	*	0.26
13	N:S:model	3.00	1,584.00	55.03	0.00	*	0.09
14	M:S:model	1.00	1,584.00	9.44	0.00	*	0.01
15	N:M:S:model	3.00	1,584.00	11.32	0.00	*	0.02

DFn are the degrees of freedom for the numerator, DFd for the denominator. The effect size is indicated by η_p^2 .

Table 2. ANOVA results simulation study 2.

	Effect	DFn	DFd	<i>F</i>	<i>p</i>	<i>p</i> < .05	η_p^2
1	N	3.00	1,584.00	94.48	0.00	*	0.15
2	M	1.00	1,584.00	585.25	0.00	*	0.27
3	S	1.00	1,584.00	1,233.33	0.00	*	0.44
4	model	1.00	1,584.00	2,702.64	0.00	*	0.63
5	N:M	3.00	1,584.00	1.46	0.22		0.00
6	N:S	3.00	1,584.00	4.90	0.00	*	0.01
7	M:S	1.00	1,584.00	102.53	0.00	*	0.06
8	N:model	3.00	1,584.00	85.17	0.00	*	0.14
9	M:model	1.00	1,584.00	624.80	0.00	*	0.28
10	S:model	1.00	1,584.00	513.39	0.00	*	0.24
11	N:M:S	3.00	1,584.00	0.27	0.85		0.00
12	N:M:model	3.00	1,584.00	182.65	0.00	*	0.26
13	N:S:model	3.00	1,584.00	33.71	0.00	*	0.06
14	M:S:model	1.00	1,584.00	33.49	0.00	*	0.02
15	N:M:S:model	3.00	1,584.00	7.71	0.00	*	0.01

DFn are the degrees of freedom for the numerator, DFd for the denominator. The effect size is indicated by η_p^2 .

Table 3. ANOVA results simulation study 3.

	Effect	DFn	DFd	<i>F</i>	<i>p</i>	<i>p</i> < .05	η_p^2
1	N	3.00	1,584.00	100.26	0.00	*	0.16
2	M	1.00	1,584.00	197.63	0.00	*	0.11
3	D	1.00	1,584.00	97.67	0.00	*	0.06
4	model	1.00	1,584.00	174.84	0.00	*	0.10
5	N:M	3.00	1,584.00	2.01	0.11		0.00
6	N:D	3.00	1,584.00	2.50	0.06		0.01
7	M:D	1.00	1,584.00	1.78	0.18		0.00
8	N:model	3.00	1,584.00	306.43	0.00	*	0.37
9	M:model	1.00	1,584.00	82.75	0.00	*	0.05
10	D:model	1.00	1,584.00	597.42	0.00	*	0.27
11	N:M:D	3.00	1,584.00	0.75	0.52		0.00
12	N:M:model	3.00	1,584.00	24.57	0.00	*	0.04
13	N:D:model	3.00	1,584.00	54.84	0.00	*	0.09
14	M:D:model	1.00	1,584.00	1.90	0.17		0.00
15	N:M:D:model	3.00	1,584.00	0.83	0.48		0.00

DFn are the degrees of freedom for the numerator, DFd for the denominator. The effect size is indicated by η_p^2 .