An out-of-sample perspective on the assessment of incremental predictive validity

Pratiwi, B.C.; Dusseldorp, E.M.L.; Rooij, M.J. de

Behaviormetrika
https://doi.org/10.1007/s41237-024-00224-7

ORIGINAL PAPER

An out-of-sample perspective on the assessment of incremental predictive validity

Bunga Citra Pratiwi[1] · **Elise Dusseldorp**[1] · **Mark de Rooij**[1]

Received: 29 January 2023 / Accepted: 5 January 2024

Abstract

In a psychometric analysis of a new psychological test, we often assess the predictive validity of a new target test over and above a baseline test, known as the incremental predictive validity. Usually, the incremental predictive validity is evaluated using within-sample statistics. Recently, it was argued to use out-of-sample assessment to prevent overfitting and non-replicable findings. In this paper, we elaborate on how to assess incremental predictive validity out-of-sample. In such an approach, we estimate prediction rules in one sample, and evaluate incremental predictive validity in another sample. Using a simulation study, we investigate whether an out-of-sample assessment results in different findings than a within-sample evaluation, taking into account the reliability of the baseline and a target test, and other factors (i.e., sample size). Results show that there is a difference between the in-sample and out-of-sample assessment, especially in small samples. However, the reliability of the two tests has no influence on this difference. In addition, we explore the effects of ridge estimation, ordinary least squares, and SIMEX, three different methods for estimating a prediction rule, on incremental predictive validity. The results show that using SIMEX leads to a bad assessment of incremental predictive validity. Ordinary least squares and ridge estimation result in almost the same incremental predictive validity estimates with a little advantage for ridge regression. In an empirical application, we show how to assess incremental predictive validity in practice and we compare that to the usual assessment.

Keywords Incremental predictive validity · Mean squared error · Reliability · Cross-validation · Prediction

Communicated by Kentaro Hayashi.

✉ Bunga Citra Pratiwi
bunga.pratiwi@fsw.leidenuniv.nl

[1] Institute of Psychology, Methodology and Statistics Department, Leiden University, Leiden, The Netherlands

Published online: 15 February 2024

1 Introduction

Psychometric evaluations of psychological tests focus on reliability and validity. Reliability can be conceptualized in multiple ways, for example, as an assessment of the internal consistency of a test measured by Cronbach's alpha, inter-rater reliability, and test-retest reliability. All conceptualizations attempt to measure the extent to which a test is free from error. Validity also has many facets, one being predictive validity, which is conceptualized as the degree to which a test predicts a criterion of interest. Reliability and predictive validity are connected, that is, less reliable tests predict worse (Spearman 1904). In classical test theory (CTT), the observed test scores, X, are assumed to be comprised of true scores (T) and random measurement error (E) (Lord and Novick 1968), that is, $X = T + E$. The error is assumed to be independent of the true score. The reliability of the test score (ρ) is then defined by

$$\rho = \frac{\sigma_T^2}{\sigma_X^2} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_E^2} \tag{1}$$

where σ_T^2 denotes the variance of the true scores, σ_X^2 that of the observed scores, and σ_E^2 the variance of the error. In practice ρ needs to be estimated, for example with Cronbach's alpha, the greatest lower bound, or the test-retest correlation (Evers et al. 2010b). We denote the estimated reliability of test X by r_{XX}. The usual way to assess predictive validity is by the (squared) correlation coefficient between the test score and the criterion. The observed correlation is influenced by the reliability of both the test score as well as the criterion score. Spearman (1904) defined the correction for attenuation as a way to assess the true predictive validity, correcting for the measurement error in both the test and criterion score, that is,

$$r_{XY} = \frac{r_{XY}^*}{\sqrt{r_{XX} r_{YY}}}, \tag{2}$$

where r_{XY} is the estimated true predictive validity, r_{XY}^* is the observed correlation (i.e., observed predictive validity) and r_{XX} and r_{YY} denote the estimated reliabilities of X and Y, respectively. When the criterion is free from error (i.e., $r_{YY} = 1$), the predictive validity decreases at the rate of the reliability index ($\sqrt{r_{XX}}$) of the test (Lord and Novick 1968).

The situation becomes more complicated when a new test is developed and the goal is to assess the predictive validity of this new test over and above the usual test. We will use the terminology *baseline test* for the existing test and *target test* for the test of interest. We would like to assess the *incremental predictive validity* (IV, Sechrest 1963; Schmidt and Hunter 1998; Westfall and Yarkoni 2016) of the target test. A typical example is the evaluation of a student entrance test to higher education. Niessen et al. (2016) studied the incremental predictive validity of a selection test for the bachelor program in psychology. The main question is whether this test has incremental predictive validity over and above the high-school grade point average. The criterion is the academic achievement of a student, operationalized as the

grade point average in the first year of the bachelor program. Both the test score as well as the grade point average are not 100% reliable.

For the assessment of incremental predictive validity, researchers typically use hierarchical multiple regression (Hunsley and Meyer 2003), which starts by fitting two regression models:

$$\text{Model 1} : Y = b_0 + b_1 X_1 + \epsilon, \tag{3}$$

$$\text{Model 2} : Y = b_0 + b_1 X_1 + b_2 X_2 + \epsilon, \tag{4}$$

where X_1 refers to the baseline test (i.e, high-school GPA in the example above), and X_2 refers to the test to be validated, the target test. For both models, an assessment of goodness of fit, such as the explained variance or mean squared error can be calculated. Incremental predictive validity is achieved if Model 2 significantly increases the model fit compared to Model 1. When X_1 and X_2 are uncorrelated, the incremental predictive validity can be shown to decrease with respect to its reliability. However, if the tests are correlated and contain measurement error, the direction, and size of the attenuation in the regression parameters, and therefore the goodness of fit, becomes unclear (Carroll et al. 2006). More advanced models are needed to correct for the attenuation, such as the SIMEX procedure (details below; Cook and Stefanski 1994).

Recently, several papers (Breiman 2001; Shmueli 2010; Yarkoni and Westfall 2017) distinguished between explanatory and predictive modeling and motivated a greater emphasis on predictive modeling. Yarkoni and Westfall (2017) argued that predictive modeling might help psychological research in dealing with the replicability crisis, because non-replicable findings are the result of overfitting. A key ingredient of predictive modeling is the evaluation of statistical models in a new sample of observations, so-called out-of-sample evaluation. More specifically, the parameters of the statistical model, such as those in Eq. (4), are estimated in one sample of observations. Then fixing those parameters gives a *prediction rule*, for example when $\hat{b}_0 = 0.3$, $\hat{b}_1 = 0.1$, and $\hat{b}_2 = 0.5$, the prediction rule becomes

$$\hat{Y} = 0.3 + 0.1 X_1 + 0.5 X_2.$$

Determining this prediction rule is equivalent to finding an explicit equation to combine the predictors. Meehl (1954) argued that finding a prediction rule can best be done by applying statistical techniques (see also Grove et al. 2000). Observed values of X_1 and X_2 from a different sample can be inserted in this rule to obtain predicted criterion scores. The predicted criterion scores can be compared to the observed criterion score using a (squared) correlation or the mean squared error. Out-of-sample evaluation protects against overfitting.

Within the context of a psychometric analysis of the reliability and validity of a psychological test, this predictive framework seems to be valuable for the evaluation of predictive and incremental predictive validity. It is however, unknown how reliability influences (incremental) predictive validity when we evaluate it out-of-sample and how incremental predictive validity differs between different statistical techniques. This paper fills that gap, by first showing how incremental predictive

validity can be assessed out-of-sample, and then investigating what the influence is of reliability of both the baseline test (X_1) and the target test (X_2) on the incremental predictive validity. We further assess what happens with the out-of-sample incremental predictive validity when a correction model, such as SIMEX, is applied and we contrast that with penalized regression models targeted at optimizing out-of-sample predictions such as ridge regression (for details see next section; Hoerl and Kennard 1970; Darlington 1978). SIMEX and ridge regression do not use the ordinary least squares (OLS) criterion to estimate the regression weights but another manner, leading to other estimates and thus a different prediction rule, and consequently leading to another assessment of incremental predictive validity. It is worthy of note that the difference between these three methods is on the incremental predictive validity, which is based on the *difference* in predictive performance between models 1 and 2, which is unknown within the context of predictive modeling. This is because typically when an estimation method is proposed or compared (such as ridge vs OLS), the focus is on comparing the method's predictive performance of one model and not on the difference between two models.

This paper is organized as follows. In the next section, we describe the theory of out-of-sample assessment of prediction rules and we discuss the three estimation methods for estimating prediction rules. We continue with a section which contains preliminary analytical results. What follows is a section that describes a simulation study investigating the effects of within or out-of-sample assessment of incremental predictive validity in relation to the reliability of the baseline and target test, the sample size, the correlation of the two tests and the overall effect size. Furthermore, we investigate the three estimators of the prediction rules. Afterward, we discuss an empirical application, where the incremental predictive validity is assessed for three student selection tests. We end this paper with a discussion.

2 Out-of-sample assessment of incremental predictive validity

Out-of-sample predictive performance rests on the trade-off between bias and variance. Simply stated, the more complex a statistical model is (i.e., the more parameters it includes), the better it will fit the data (less bias), but the more variable its predictions will be (more variance). On the other hand, a simpler model will fit worse (more bias), but its predictions will be less variable (less variance). In a formal sense, when we fit a statistical model to a sample of data and investigate its predictive performance, the expected prediction error decomposes into (squared) bias and variance of the fitted model (Hastie et al. 2009; Yarkoni and Westfall 2017; Chapman et al. 2016; McNeish 2015). In our context, Model 2 including the new test under investigation has more variance but less bias than Model 1 because Model 2 has more predictors, hence more parameters to estimate.

Trading off bias and variance is ideally done by estimating a statistical model in one sample and evaluating the predictive performance in another independent sample. However, usually we only have a single sample available. An exception is in statistical Monte Carlo simulation studies, where generating independent test or validation sets from the same (or a different) population is easily done.

In practical data analysis, where we only have a single sample available, cross-validation is used (Mosier 1951; Stone 1974; Browne 2000; De Rooij and Weeda 2020). In so-called K-fold cross-validation, we partition the data set into K independent parts and iteratively use each part as the validation or test set and the other $K-1$ as training or calibration set.

A standard way of assessing incremental predictive validity is by judging the change in explained variance, ΔR^2. Explained variance has a one-to-one relationship to the squared error when we assess in-sample. Out-of-sample, however, this link is broken and R^2 is not a good measure for predictive performance. De Rooij and Weeda (2020) give the following example. Model 1 might predict $\hat{Y} = \{1, 2, 3, 4, 5\}$, whereas Model 2 might predict $\hat{Y} = \{7, 8, 9, 10, 11\}$, when the actual observations in the validation set are $Y = \{6, 7, 8, 9, 10\}$. It is clear that for both sets of predictions, the correlation with the observed outcome equals 1, although the predictions from Model 2 are much better than those of Model 1 (see Appendix A for a simulated example which shows this result). Furthermore, Van Loon et al. (2020) show that out-of-sample R^2 equals 1 for a statistical model only having an intercept and evaluated with K-fold cross-validation when K equals the sample size, that is, predictive performance would be optimal according to this measure and no psychological test could improve the predictive performance.

A better measure of predictive performance is given by the mean squared error of prediction, that is

$$\text{MSEP} = \frac{1}{N_\text{v}} \sum (Y - \hat{Y})^2, \tag{5}$$

where the sum is over the N_v observations in the validation set. For the two sets of predictions in the previous paragraph, the MSEP is 25 for the first model and 1 for the second. For incremental predictive validity, we use the change in MSEP

$$\Delta\text{MSEP} = \text{MSEP}_1 - \text{MSEP}_2 \tag{6}$$

where MSEP_1 is a measure of prediction error with only the baseline test and MSEP_2 that of the model including our target test. In contrast to in-sample evaluation, the measure of incremental predictive validity can become negative indicating that predictions become worse when including the new test. Positive ΔMSEP indicates incremental predictive validity.

In the training set, we need to estimate the coefficients for Model 1 and 2 (Eqs. 3 and 4). In the training set of size N, we have observed values for Y, X_1 and X_2. The observed values will be denoted by y_i, x_{i1}, and x_{i2} for $i = 1, \ldots, N$. The estimated coefficients become part of the prediction rule, therefore, the way of estimation determines the prediction rule, the MSEP, and the ΔMSEP. Three different estimators will be investigated.

The first method of estimation is *ordinary least squares* (OLS). For Model 2 (i.e., Eq. 4), the estimates are obtained by minimizing the usual least squares function

$$L_{\text{ols}}(b_0, b_1, b_2) = \sum_{i=i}^{N} (y_i - b_0 - x_{i1}b_1 - x_{i2}b_2)^2. \tag{7}$$

A similar loss function is used when estimating the regression weights for model 1 (i.e., Eq. 3).

The second method of estimation is *ridge regression*, where the squared regression weights are penalized (Hoerl and Kennard 1970; Darlington 1978). By penalizing the regression weights, bias is increased but variance is reduced. For Model 2 (i.e., Eq. 4), the estimates are obtained by minimizing

$$L_{\text{ridge}}(b_0, b_1, b_2) = \sum_{i=i}^{N} (y_i - b_0 - x_{i1}b_1 - x_{i2}b_2)^2 + \gamma(b_1^2 + b_2^2), \tag{8}$$

where γ is a tuning parameter, giving more or less weight to the penalty. The optimal value of γ is often found by cross-validation, that is, a sequence of values is determined and for each value K-fold cross-validation is performed. The value that leads to the lowest prediction error is chosen as the optimal value. There are also other penalized least squares methods, such as lasso (Tibshirani 1996) and elastic-net (Zou and Hastie 2005), but these are more oriented toward variable selection as these penalties tend to shrink regression weights to zero. The ridge penalty often leads to better predictive performance, without selecting variables.

The third method of estimation is SIMEX (Cook and Stefanski 1994). The goal of this method is to obtain estimated coefficients from supposedly error-free predictors. In contrast to the first two methods, the SIMEX method does not use a loss function to be minimized. Instead, the method is composed of the following steps:

1. In the first step, B additional data sets are generated for every value $\lambda_1, \lambda_2 \in \{0.5, 1, 1.5, 2\}$ with increasing levels of measurement error. Therefore, draw a vector $E_j \sim N(0, \sigma^2_{E_j})$ and compute $Z_j(\lambda_j) = X_j + \lambda_j E_j = T_j + E_j + \lambda E_j$ for both predictors ($j = 1, 2$). The variance of these inflated measurement errors of $Z_j(\lambda_j)$ equals $(1 + \lambda_j^2)\sigma^2_{E_j}$.
2. In the second step, regression weights are estimated using OLS for each generated data set with Y as criterion and $Z_1(\lambda_1)$ and $Z_2(\lambda_2)$ as predictor variables. As the λ's increase, the predictors become less reliable, and the estimates would become increasingly biased. This relationship between the λ's and the biases of the parameter estimates is the basis for extrapolation.
3. In the third step, for each value of λ_1 and λ_2, the regression estimates are averaged over the B solutions.
4. A function (e.g., linear, quadratic, or non-linear) is fitted to the averaged estimates against the λs.
5. Using this estimated function, an extrapolation is performed to the case of no measurement error ($\lambda_1 = \lambda_2 = -1$), which is our SIMEX estimate of the regression weights (b_0, b_1, and b_2).

3 Preliminary analytical results

For simple situations, we can expect an ordering of MSEP_j between ridge, OLS, and SIMEX by demonstrating the order of the mean squared error (MSE) of the estimated coefficients. This demonstration is shown for Model 1 and Model 2 separately.

Consider Model 1 where the baseline test is the only predictor. Let b denote the true regression parameter that exist in the population and estimator $\hat{b}_{(s)} = s\hat{b}$, where s is a positive constant and $\hat{b}$ is the OLS estimator, which is unbiased. We can write $\text{MSE}[\hat{b}_{(s)}]$ as

$$\begin{aligned}
\text{MSE}[\hat{b}_{(s)}] &= \text{E}[(\hat{b}_{(s)} - b)^2] \\
&= \text{E}[(s\hat{b} - b)^2] \\
&= \text{E}[(s\hat{b} - sb + sb - b)^2] \\
&= \text{E}[(s\hat{b} - sb)^2 + 2(s\hat{b} - sb)(sb - b) + (sb - b)^2] \\
&= \text{E}[(s\hat{b} - sb)^2] + \text{E}[2(s\hat{b} - sb)(sb - b)] + (sb - b)^2 \\
&= \text{E}[(s\hat{b} - sb)^2] + 2s\text{E}[(\hat{b} - b)](sb - b) + (sb - b)^2.
\end{aligned} \tag{9}$$

Since $\hat{b}$ is the OLS estimator, as a result, $E[(\hat{b} - b)] = 0$. In addition, we can simplify $(sb - b)^2 = b^2(s - 1)^2$. Thus, $\text{MSE}[\hat{b}_{(s)}]$ can be written as

$$\text{MSE}[\hat{b}_{(s)}] = s^2\text{E}[(\hat{b} - b)^2] + b^2(s - 1)^2. \tag{10}$$

Denote $\text{E}[(\hat{b} - b)^2] = \text{var}(\hat{b})$. $\text{MSE}[\hat{b}_{(s)}]$ is minimal when

$$s = \frac{b^2}{\text{var}(\hat{b}) + b^2}, \tag{11}$$

which is a value smaller than 1. Thus, the optimal s is not when $s = 1$, which would make $\hat{b}_{(s)}$ equal to $\hat{b}$, but somewhere between 0 and 1. With $0 < s < 1$, $\hat{b}$ is shrunken toward zero, which is analogous to how regression weights are treated in ridge regression. This result has been previously shown by Darlington (1978). In addition, Van Houwelingen and Le Cessie (1990) showed that the same value of s also minimizes the mean squared error of prediction.

Suppose that the baseline test contains measurement error. A sensible remedy would be to correct for $\hat{b}$ because it is not an unbiased estimate of b. A correction for $\hat{b}$ using SIMEX is equivalent to setting s to the inverse of the reliability of the baseline test ($s = 1/\rho_1$) (Carroll et al. 2006; Cook and Stefanski 1994). Since the reliability of a baseline test containing error is $\rho_1 \in (0, 1)$, then $s > 1$, which means that $\hat{b}_{(s>1)} > \hat{b}$. For $s > 1$, $\text{MSE}[\hat{b}_{(s=1)}] < \text{MSE}[\hat{b}_{(s>1)}]$. Therefore, we can assume an ordering of $\text{MSE}[\hat{b}_{(0<s<1)}] < \text{MSE}[\hat{b}] < \text{MSE}[\hat{b}_{(s>1)}]$ (ridge, OLS, SIMEX, respectively).[1]

[1] See Appendix A for a simulated example.

For Model 2, where both baseline test and target test are predictors, an order for the MSE of the three estimators can be shown when the tests are uncorrelated. Let $\mathbf{X}$ be an orthonormal matrix containing scores on a baseline and a target test and $\mathbf{y}$ be a vector containing scores on an outcome variable. Let $\mathbf{b}$ be a vector of the true regression coefficients b_1 and b_2. In matrix notation, the OLS estimator of $\mathbf{b}$ is

$$\hat{\mathbf{b}} = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}. \tag{12}$$

For an estimator of a parameter, its MSE can be decomposed to its variance and squared bias. Thus, MSE[$\hat{\mathbf{b}}$] can be defined as follows:

$$\text{MSE}[\hat{\mathbf{b}}] = \text{var}(\hat{\mathbf{b}}) + (\text{bias}(\hat{\mathbf{b}}))^2. \tag{13}$$

Since $\hat{\mathbf{b}}$ is an unbiased estimator of $\mathbf{b}$, MSE[$\hat{\mathbf{b}}$] reduces to its variance. The variance can be defined as $\text{var}(\hat{\mathbf{b}}) = \sigma^2 \text{tr}((\mathbf{X}^\top \mathbf{X})^{-1})$ (Hoerl and Kennard 1970), where σ^2 is the variance of the residuals. Thus, MSE[$\hat{\mathbf{b}}$] for orthonormal $\mathbf{X}$ is $2\sigma^2$.

The ridge estimator $\hat{\mathbf{b}}_{(\kappa)}$ (Hoerl and Kennard 1970) is defined as

$$\hat{\mathbf{b}}_{(\kappa)} = (\mathbf{X}^\top \mathbf{X} + \kappa \mathbf{I})^{-1}\mathbf{X}^\top \mathbf{y}. \tag{14}$$

In the case of orthonormal matrix $\mathbf{X}$, van Wieringen (2021) showed that the ridge estimator reduces to

$$\begin{aligned}
\hat{\mathbf{b}}_{(\kappa)} &= (\mathbf{X}^\top \mathbf{X} + \kappa \mathbf{I})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (\mathbf{I} + \kappa \mathbf{I})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (1+\kappa)^{-1}\mathbf{I}\mathbf{X}^\top \mathbf{y} \\
&= (1+\kappa)^{-1}(\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y} \\
&= (1+\kappa)^{-1}\hat{\mathbf{b}}.
\end{aligned} \tag{15}$$

It is clear that the ridge estimator shrinks the OLS estimator by $(1+\kappa)^{-1}$. If we take the expectation on both sides, the ridge estimator is shown to be biased.

$$\begin{aligned}
\text{E}[\hat{\mathbf{b}}_{(\kappa)}] = \text{E}[(1+\kappa)^{-1}\hat{\mathbf{b}}] &= (1+\kappa)^{-1}\text{E}[\hat{\mathbf{b}}] \\
&= (1+\kappa)^{-1}\mathbf{b}.
\end{aligned} \tag{16}$$

In the case of orthonormal $\mathbf{X}$, the mean squared error of the ridge estimator is

$$\begin{aligned}
\text{MSE}[\hat{\mathbf{b}}_{(\kappa)}] &= \text{E}[(\hat{\mathbf{b}}_{(\kappa)} - \text{E}[\hat{\mathbf{b}}_{(\kappa)}])^2] + (\text{E}[\hat{\mathbf{b}}_{(\kappa)}] - \mathbf{b})^2 \\
&= \text{E}[((1+\kappa)^{-1}\hat{\mathbf{b}} - (1+\kappa)^{-1}\mathbf{b})^2] + ((1+\kappa)^{-1}\mathbf{b} - \mathbf{b})^2 \\
&= (1+\kappa)^{-2}\text{E}[(\hat{\mathbf{b}} - \mathbf{b})^2] + \mathbf{b}\mathbf{b}^\top((1+\kappa)^{-1} - 1)^2 \\
&= (1+\kappa)^{-2}\sigma^2 \text{tr}((\mathbf{X}^\top \mathbf{X})^{-1}) + \mathbf{b}\mathbf{b}^\top((1+\kappa)^{-1} - 1)^2,
\end{aligned} \tag{17}$$

and because $\mathbf{X}^\top \mathbf{X} = \mathbf{I}$, MSE[$\hat{\mathbf{b}}_{(\kappa)}$] becomes

$$\text{MSE}[\hat{\mathbf{b}}_{(\kappa)}] = (1+\kappa)^{-2}2\sigma^2 + \mathbf{b}\mathbf{b}^\top((1+\kappa)^{-1} - 1)^2, \tag{18}$$

which is minimal when

$$\kappa = \frac{2\sigma^2}{\mathbf{b}\mathbf{b}^\top}. \tag{19}$$

MSE reaches its minimum not when $\kappa = 0$ ($\hat{\mathbf{b}}_{(\kappa)} = \hat{\mathbf{b}}$) but when $\kappa > 0$ ($\hat{\mathbf{b}}_{(\kappa)} < \hat{\mathbf{b}}$). Therefore, for Model 2, it is clear that $\text{MSE}[\hat{\mathbf{b}}_{(\kappa>0)}] < \text{MSE}[\hat{\mathbf{b}}]$. In the case where $\mathbf{X}$ contains measurement error, which means a correction is warranted, the correction done in SIMEX will increase $\hat{\mathbf{b}}$ because $\mathbf{X}$ remains uncorrelated. To increase $\hat{\mathbf{b}}$, κ must be between -1 and 0. When $-1 < \kappa < 0$, using (18), it should be clear that $\text{MSE}[\hat{\mathbf{b}}] < \text{MSE}[\hat{\mathbf{b}}_{(-1<\kappa<0)}]$. Therefore, we can assume the particular ordering of the MSE when baseline and target tests are not correlated is $\text{MSE}[\hat{\mathbf{b}}_{(\kappa>0)}] < \text{MSE}[\hat{\mathbf{b}}] < \text{MSE}[\hat{\mathbf{b}}_{(-1<\kappa<0)}]$ (ridge, OLS, and SIMEX respectively). The expected ordering becomes challenging to prove analytically when baseline and target tests are correlated. When tests are correlated, the constant κ will depend on their correlation and the true regression coefficients of each test in the model (Darlington 1968). Moreover, if the tests contain measurement error, the reliability of both tests will also play a role (Carroll et al. 2006).

Above, we showed analytical derivations for Model 1 and Model 2. We showed that the mean squared error will be smaller for the ridge estimator compared to the OLS estimator, and that in turn the mean squared error of the OLS estimator will be smaller than the SIMEX estimator. This is true for both Model 1 and Model 2 (see Eqs. 11 and 19). Incremental predictive validity is, however, defined as the difference between the mean squared errors of prediction of Model 1 and Model 2. That is, we are interested in the change in the mean squared error of prediction between models 1 and 2, when both are estimated by ridge, OLS, or SIMEX. For this difference, no analytical results are available for the different estimators. Therefore, we resort to Monte Carlo Simulations.

4 Simulation study

In the simulation study, we considered two test scores, one for the baseline test and one for the target test. The simulation study was divided into two parts. In Part I, we focused on two questions. the first question is: what is the effect of using within or out-of-sample assessment of incremental predictive validity? The second question is: what are the effects of the reliabilities of the baseline and target test on incremental predictive validity for the within sample and the out-of-sample approach and is this effect different for the two approaches? For both questions, we took sample size into account and the collinearity between the true tests scores, the overall predictability of the criterion (R^2), and the ratio of the effect of the baseline and target test. In this first part, we only compared results from prediction rules estimated with ordinary least squares.

In Part II, we focused on the assessment of incremental predictive validity from an out-of-sample perspective. Here, we compared out-of-sample

incremental predictive validity of prediction rules estimated by OLS, SIMEX, and ridge. We were interested in whether the differences in out-of-sample incremental predictive validity across estimation methods depended on sample size, the collinearity between the true tests scores, the predictability of the criterion (R^2), and the ratio of the effect of the two tests.

4.1 Data generation and simulation factors

Criterion Y is generated by

$$Y = b_1 T_1 + b_2 T_2 + \epsilon,$$

where true test scores T_1 and T_2 were sampled from a bivariate normal distribution with means $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariances $\Sigma = \begin{pmatrix} 1 & r_{12} \\ r_{12} & 1 \end{pmatrix}$, and the error term ϵ is drawn from a standard normal distribution with a standard deviation of $\sqrt{1 - R^2}$. We used four conditions of the ratio of the regression coefficients $(b_1 : b_2) \in (2:1, 1:1, 1:2, 1:0)$, four conditions of overall effect $(R^2 \in (0.1, 0.2, 0.3, 0.4))$ and five degrees of collinearity $(r_{12} \in (0.1, 0.3, 0.5, 0.7, 0.9))$. The first three conditions of $b_1 : b_2$ reflect conditions when there is incremental predictive validity. The last condition of $b_1 : b_2$ represents a situation where there is no incremental predictive validity in the population. Appendix B gives the exact calculations of the regression coefficients given overall effect and degree of collinearity.

In practice, instead of observing the true test scores, we observe scores that contain random measurement error. The observed test scores were generated by adding random normally distributed error to the true scores with mean 0 and variance

$$\sigma^2_{E_j} = \frac{1 - \rho_j}{\rho_j}, \tag{20}$$

where ρ_j is the reliability of test j, defined as in Eq. (1). We vary the reliabilities of both the baseline and target test. Reliabilities 0.6 to 0.9 represent the different cut-off values of standard criteria of acceptability (Evers et al. 2010a). Reliabilities of 0.5 and 1 reflect extreme cases of reliability.

Calibration data were generated with sample sizes of 50, 100, 200, 500, and 1000. These conditions reflect typical sample sizes that may be found in validation studies of various psychological tests. Validation data were generated using the same model, but with a sample size of 10,000. A large number of the validation sample provides the true error of the prediction rule (Varma and Simon 2006). Furthermore, having a uniform size for the validation sample enables a fair comparison of the predictive accuracies between prediction rules. Table 1 gives a summary of the design factors for our simulation study. In each of the 14,400 conditions, we generated 500 calibration and validation sets.

Table 1 Summary of simulation design factors

Description	Factors	Levels	*n* levels
Reliability of baseline test	ρ_1	0.5, 0.6, 0.7, 0.8, 0.9, 1	6
Reliability of target test	ρ_2	0.5, 0.6, 0.7, 0.8, 0.9, 1	6
Calibration sample	N_{cal}	50, 100, 200, 500, 1000	5
Ratio of the effect size	$b_1 : b_2$	1:0, 2:1, 1:1, 1:2	4
Degree of collinearity	r_{12}	0.1, 0.3, 0.5, 0.7, 0.9	5
Predictability of the criterion	R^2	0.1, 0.2, 0.3, 0.4	4
Total conditions		$6 \times 6 \times 5 \times 4 \times 5 \times 4$	14,400

4.2 Outcome measure of incremental predictive validity

For each generated calibration data set, we fitted the two regression models defined in (3) and (4) using all three estimators (OLS, ridge, and SIMEX), resulting in two prediction rules for each estimation method.

In Part I of the simulation studies, we only investigated the estimated prediction rules from OLS. To compare the within-sample assessment of IV with the out-of-sample assessment, we used either the calibration data set (again) as the validation data set or the validation data set. Inserting the values of the test scores in these prediction rules gives predicted criterion scores. For both models, we can compute the MSEP and their difference. When the calibration set is used as validation set, the MSEPs are equal to the usual MSEs and they are prone to overfitting. When the validation data set is used, the assessment should be protected against overfitting.

In Part II, where we only assess IV out-of-sample, we also take the SIMEX and ridge rules into account. Our interest lies in the difference between these three prediction rule estimators on the IV and possible interactions with other design factors.

To evaluate our simulation results, we performed two mixed analysis of variance (ANOVA) tests using SPSS version 27 (IBM corp 2020). The outcome variable in these mixed-ANOVA's is the ΔMSEP. To inspect the size of the effects in the mixed-ANOVA, we used the partial η^2. We used the guidelines of Cohen et al. (2013), that is $\eta^2 = 0.01$ denotes a small effect, $\eta^2 = 0.06$ a medium effect, and $\eta^2 = 0.14$ a large effect.

The simulation was done using R version 4.2.1 (R Core Team 2022). To implement the SIMEX method, we used the **simex** package (Lederer et al. 2017) and for the ridge method, the **parcor** package (Kraemer et al. 2009) was used. In the SIMEX implementation, we used the default options of the main function to run the algorithm: quadratic extrapolation function, 100 bootstrap samples, and a predetermined set of λ values as previously mentioned. In the ridge method implementation, we used a tenfold cross-validation process to choose the optimal shrinkage parameter (γ) from a predetermined set of values of this parameter provided in the function.

4.3 Simulation results

4.3.1 Part I: effect of within or out-of-sample assessment of IV

To investigate the effect of using a within or out-of-sample assessment of IV and the effect of reliabilities on IV for each approach, a mixed-ANOVA analysis was performed with the assessment approach (i.e., within or out-of-sample) as the within-subjects factor and the data design factors as between-subjects factors. We included all main effects till three-way interactions. Mauchly's test of sphericity was violated in this analysis. Therefore, all the significance tests are based on the Greenhouse–Geisser correction. The assumptions of normality of the residuals of the ANOVA were assessed by checking the raw skewness values of the residuals in each of the 14,400 conditions. We found that the majority of the skewness were between -2 and 2 which is considered as not severely skewed (Kim 2013; Kline 2015; Hair Jr et al. 2021). Therefore, we proceeded with using the results from this mixed-ANOVA. However, because there were some conditions that fell outside of the range, we also performed robust ANOVAs on the effects that we interpret to ensure that the significance tests were not affected.

In the regular mixed-ANOVA, we found a significant main effect of the approach to analyze IV with a large effect size $F(17, 199{,}737) = 469{,}590.395$, $p < 0.001$, $\eta^2 = 0.061$. This effect was found to only depend on the size of the calibration sample with a medium to large effect $F(4, 7{,}199{,}737) = 103{,}035.608$, $p < 0.001$, $\eta^2 = 0.054$. The interaction effect between approach and other design factors (ρ_1, ρ_2, $b_1 : b_2$, r_{12}, and R^2) had effect sizes below 0.01, and therefore not considered influential. Note that the main effect of IV approach and the interaction effect between IV approach and calibration sample size were also significant in the robust ANOVAs (see Table 5 in Appendix C).

Figure 1 shows aggregated results of IV for the within or out-of-sample approach against N_{cal} collapsed over all other factors. As seen in Fig. 1, there is a discrepancy between the within- and out-of-sample assessment, that is, IV is larger when assessed within sample, a sign of overfitting. This difference diminishes as the size of the calibration sample increases.

Effects of reliabilities on IV for both approaches As the difference in IV between approaches only depended on N_{cal}, we conclude that the effects of the reliabilities (ρ_1 and ρ_2) were the same for both approaches. Therefore, we further inspected the between-subjects effects of the reliabilities. In general, all design factors (except the size of the calibration sample) were found to have significant and large between-subjects main effects on IV (see Table 2). For an overview of the effects, we listed the between-subjects effects that were medium to large in Table 2. Note that these effects were also found to be significant in the robust ANOVAs (see Table 5 in Appendix C). As can be seen in Table 2, the size of the calibration sample did not appear, which suggests that its between-subjects effects were not influential because their effect sizes were below .01. The factors that were influential were $b_1 : b_2$, r_{12}, R^2, and the reliabilities of the tests (ρ_1 and ρ_2). In the following, we focus the interpretation on medium to large two-way interaction effects on IV that involve the reliabilities of the tests.

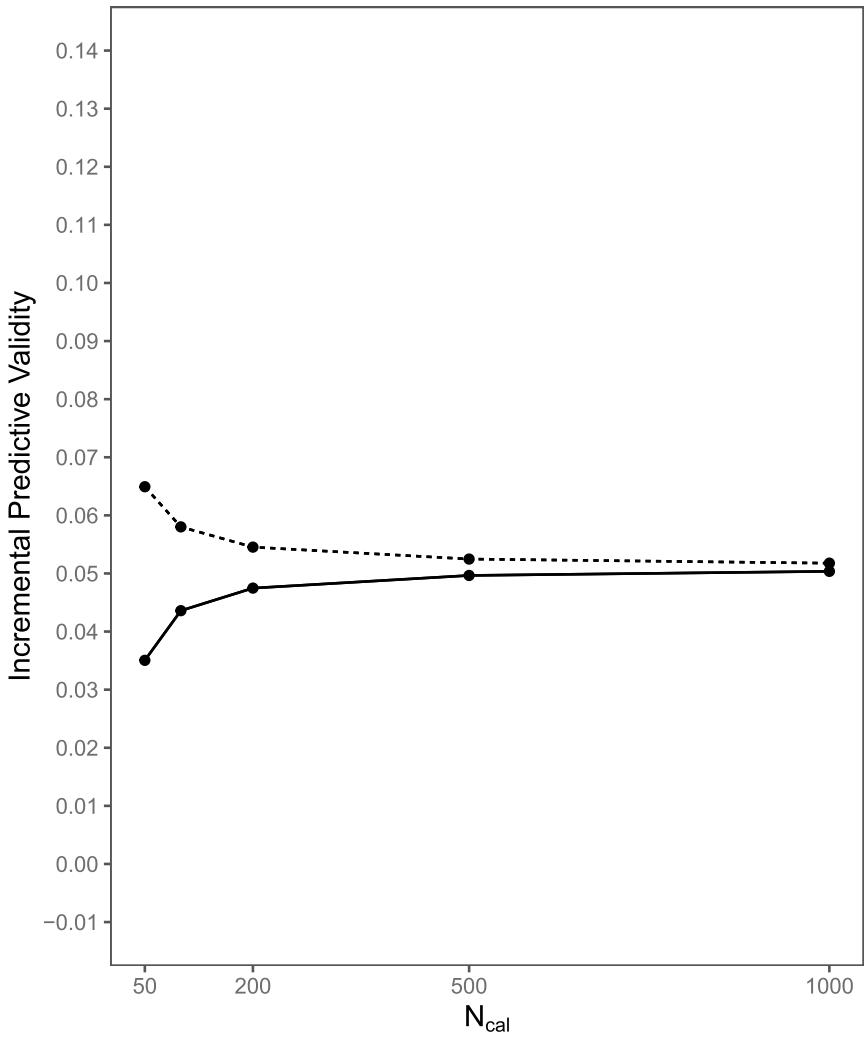


Fig. 1 Incremental predictive validity (ΔMSEP) against N_{cal} defined using the within-sample assessment (bold line) and the out-of-sample assessment (dotted line)

Figure 2 shows several multi-panel line plots in which we display aggregated results of IV for each approach against the reliability of the target test (ρ_2) and against the reliability of the baseline test (ρ_1). The panels in each plot represent one other design factor. Notice that for all plots the two lines representing the within and out-of-sample approaches to assess IV are parallel, showing the main effect of the approach to assess IV.

The reliability of the target test is shown to have a positive relationship with its IV, such that the IV increased as the test became more reliable (Fig. 2a, b). It can also be seen that this effect became stronger as the predictability of the criterion (R^2) increased (see Fig. 2a). In Fig. 2b, the effect of the reliability of the target test (ρ_2) on IV weakened as the ratio between b_1 and b_2 increased. For the condition where there is no IV ($b_1 : b_2 = 1 : 0$), within-sample IV was always positive, whereas out-of-sample IV was negative for very weak reliability (0.5 and 0.6) but could be slightly positive when tests had high reliability ($\rho_2 > 0.7$).

The effect of the reliability of the baseline test on the IV of the target test was dependent on the predictability of the criterion (Fig. 2c) and collinearity between

Table 2 Between subjects effects table on IV (averaged over approaches) using prediction rules from OLS

	SS	*df*	*F*	η^2	Robust ANOVA
$(b_1 : b_2)$	15,110.20	3	5,818,679.37	0.708	+
R^2	7529.46	3	2,899,465.87	0.547	+
$(b_1 : b_2) \times r_{12}$	3635.56	12	349,997.61	0.368	+
$(b_1 : b_2) \times R^2$	3023.37	9	388,082.64	0.327	+
ρ_2	2742.66	5	633,691.51	0.306	+
ρ_1	2230.81	5	515,429.15	0.264	+
r_{12}	1817.24	4	524,840.93	0.226	+
$\rho_2 \times (b_1 : b_2)$	1013.55	15	78,060.28	0.140	+
$\rho_1 \times r_{12}$	861.33	20	49,752.31	0.121	+
$\rho_2 \times R^2$	547.76	15	42,186.49	0.081	+
$\rho_1 \times R^2$	446.90	15	34,418.59	0.067	+
$r_{12} \times R^2$	363.83	12	35,026.43	0.055	+
Error	6232.20	7,199,737			

Robust ANOVA = (+) significant ($p < 0.05$) and (−) not significant ($p > 0.05$)

SS sum of squares, *df* degrees of freedom, η^2 = partial eta-squared

the true scores of the tests (Fig. 2d). Figure 2c shows the relationship between IV and the reliability of the baseline test (ρ_1) aggregated on the predictability of the criterion. We see that as the reliability of the baseline test becomes higher the IV of the target test decreased. In addition, this effect became stronger as the criterion was increasingly predictable. In Fig. 2d, we can see that as the true test scores became highly correlated, the effect of the reliability of the baseline test (ρ_1) became stronger. With higher reliability, the IV of the target test decreased, more so when collinearity increased.

4.3.2 Part II: effect of estimation methods on out-of-sample assessment of IV

The second goal of the simulation study is to compare the out-of-sample assessment of IV for the different estimators of the prediction rule. Before we discuss this comparison, we show how the estimates in the prediction rules differ between the three methods.

Figure 3 shows aggregated results of the coefficients of Model 2 across 500 repetitions for the condition when the true regression coefficients were equal, the size of calibration sample was moderate ($N_{\text{cal}} = 200$), the effect of the full model was medium ($R^2 = 0.2$), and with low collinearity in the true test scores ($r_{12} = 0.1$). Figure 3a shows the effect of the reliability of the target test (ρ_2) on the estimates when the baseline test is fully reliable ($\rho_1 = 1$), whereas Fig. 3b shows the effect of the reliability of the baseline test (ρ_1) when the target test if fully reliable ($\rho_2 = 1$).

As shown in Fig. 3a, the estimated intercepts were not affected by the reliability of the target test and they were close to the true values. The reliability of the target test influences its corresponding coefficient $\hat{b}_2$, such that, as reliability

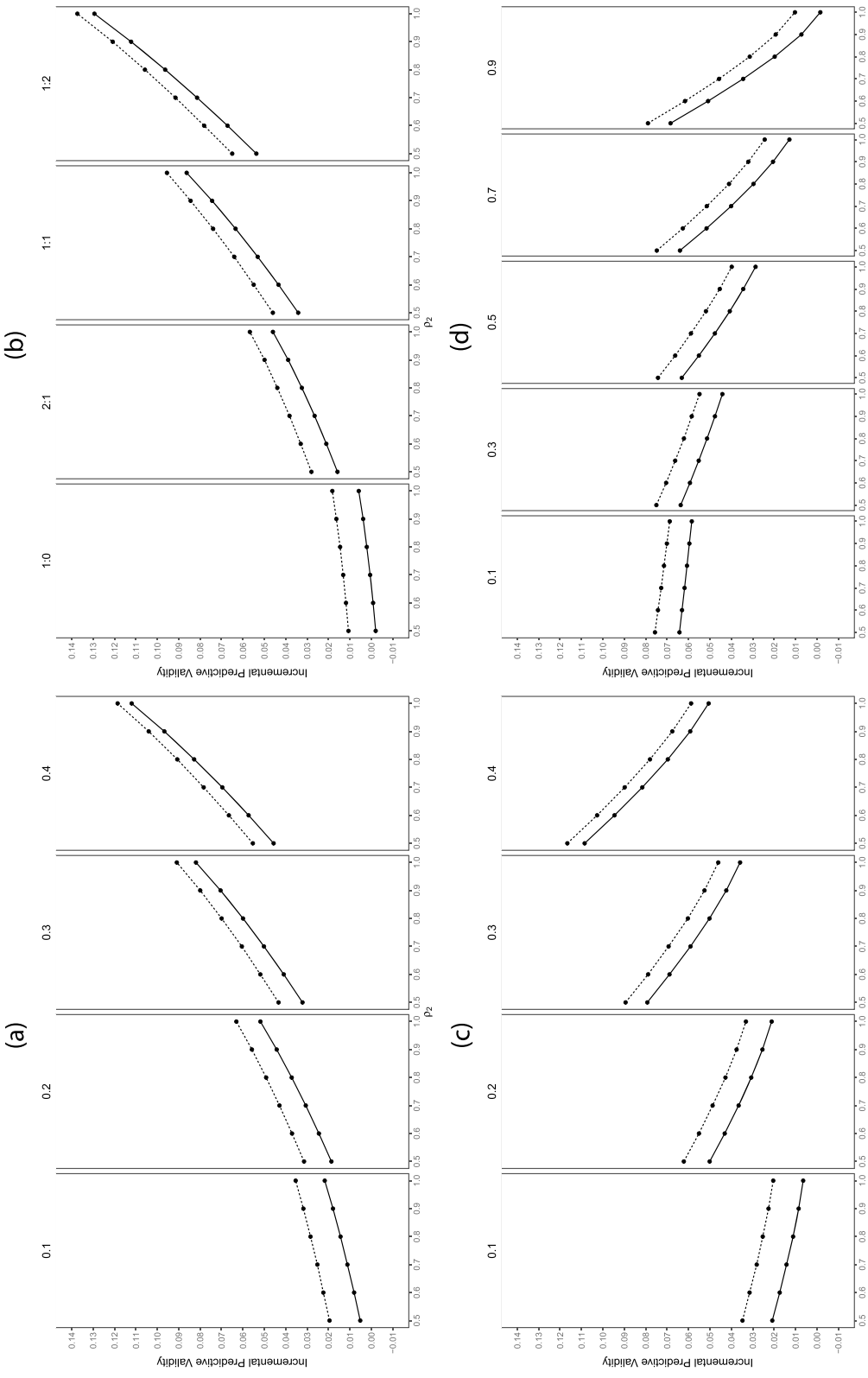


Fig. 2 Incremental predictive validity as function of the reliability. In the upper row, incremental predictive validity against the reliability of the target test (ρ_2) as a function of (**a**) the effect size of the full model (R^2) and (**b**) the ratio of the two effect sizes($b_1 : b_2$). In the lower row, incremental predictive validity against the reliability of the baseline test (ρ_1) as a function of (**c**) the effect size the full model (R^2) and (**d**) the collinearity between the two tests (r_{12}). The two lines represent IV assessed within sample (bold line) and out-of-sample (dotted line)

Fig. 3 Mean estimated coefficients of Model 2 (i.e., $\hat{b}_0$, $\hat{b}_1$, and $\hat{b}_2$) against (**a**) the reliability of the target test (ρ_2; with $\rho_1 = 1$) and (**b**) the reliability of the baseline test (ρ_1; with $\rho_2 = 1$) as a function of three estimation methods (dotted = OLS, dashed = SIMEX, and bold = ridge), for the condition when $R^2 = 0.2$, $N_{\text{cal}} = 200$, $b_1 : b_2 = 1 : 1$, and $r_{12} = 0.1$. Red lines represent true values of the coefficients in the population ▶

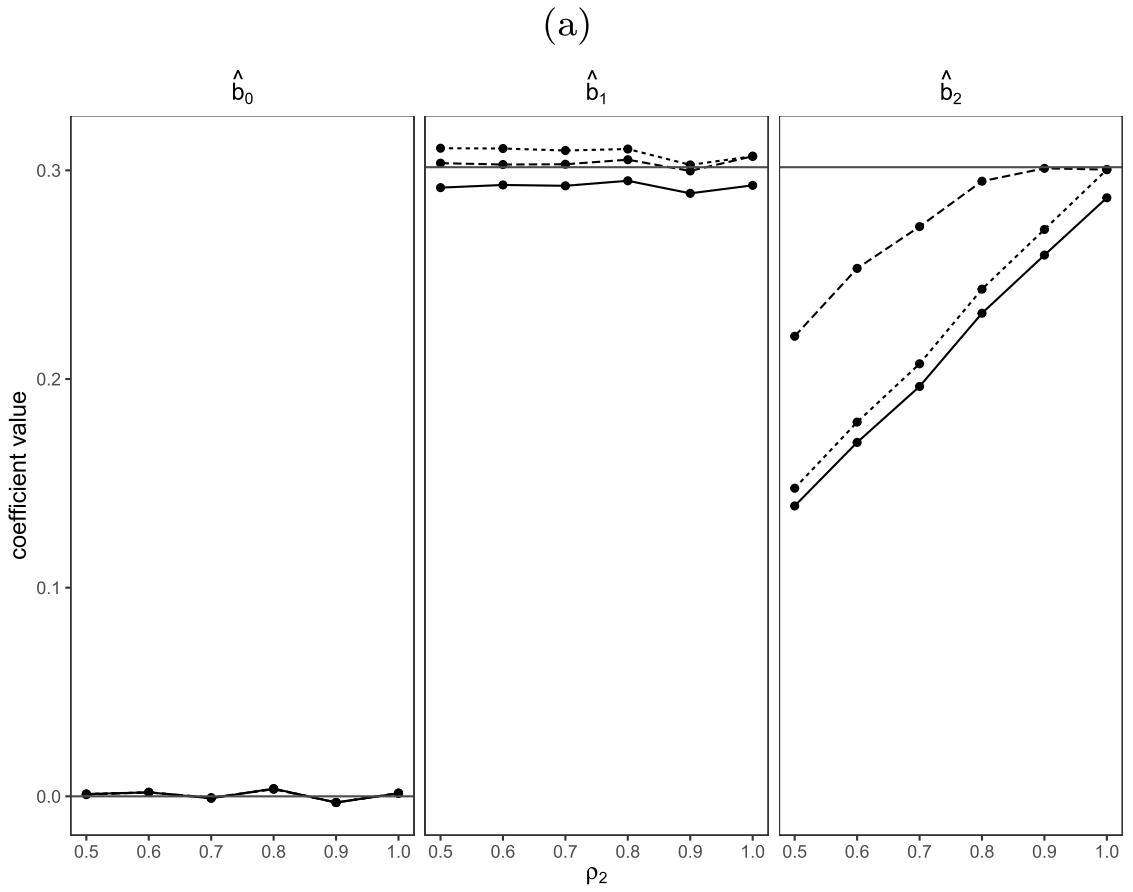
increased, the estimated coefficients approximate the true values. Estimates of the SIMEX method were on average closer to the true values compared to ridge and OLS estimates. Estimates $\hat{b}_1$ and $\hat{b}_2$ from the ridge method were always smaller than the estimates from OLS (and SIMEX), indicating the shrinkage effect on the coefficients.

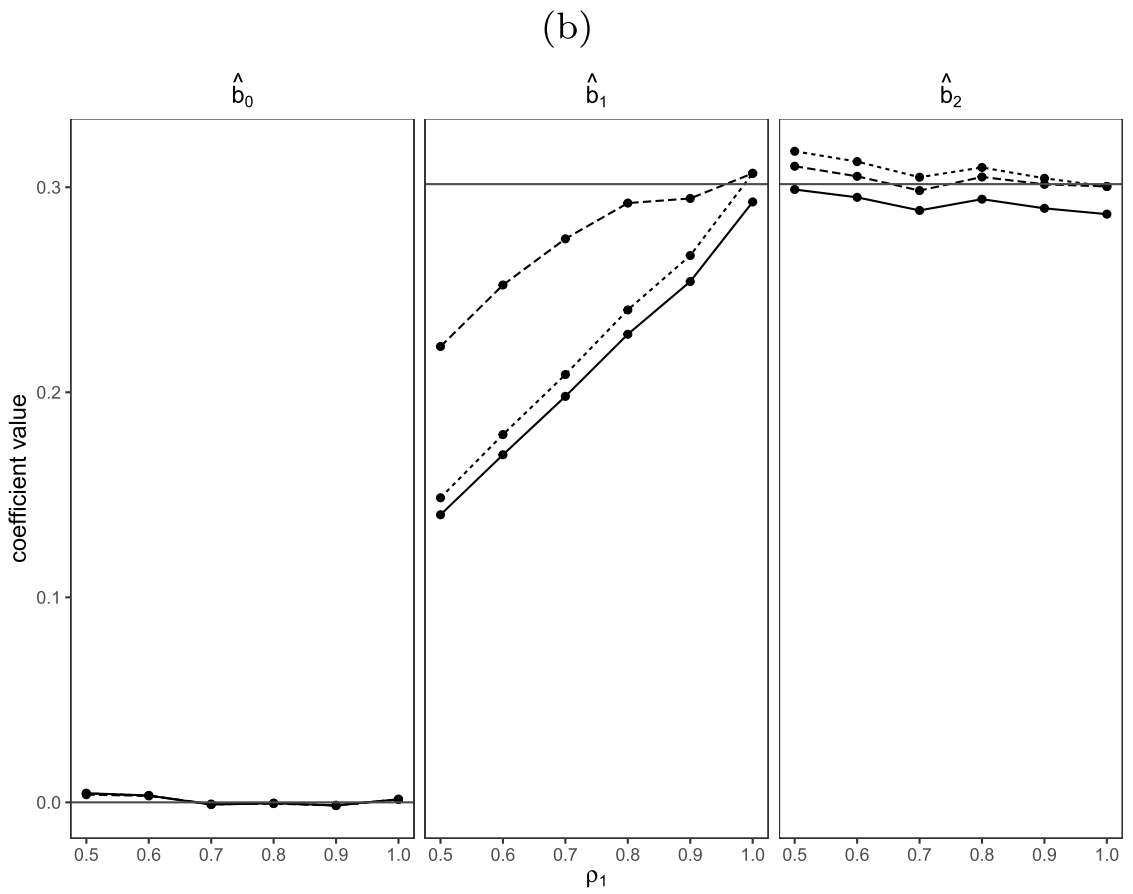
In Fig. 3b, it is shown that the reliability of the baseline test only influenced the b_1 estimates. When $\rho_1 < 0.9$, estimated $\hat{b}_1$ from the SIMEX correction were on average smaller than its true value. Estimates of $\hat{b}_2$ were on average very close to the true value and were not affected by the reliability of the baseline test. In this figure, we can also see that the ridge estimates of $\hat{b}_1$ and $\hat{b}_2$ were consistently smaller than the estimates from OLS and SIMEX.

To compare out-of-sample IV between the three methods, a mixed ANOVA analysis was performed to investigate the effect of varying estimation methods on IV. In the analysis, methods (Methods) are the within-subjects factor with out-of-sample IV as the dependent variable. The analysis included main effects of all the factors until three-way interactions between Methods and two data design factors. Mauchly's test of sphericity was violated in this analysis, therefore the significance of the effects was tested based on the Greenhouse–Geisser correction. The assumption of the normality of the residuals of the ANOVA was assessed by checking the skewness of the residuals for each outcome in each of the 14,400 conditions. We found that most of the conditions were within the range −2 and 2, which is considered not severely skewed (Kim 2013; Kline 2015; Hair Jr et al. 2021). Thus, we proceeded with using the results from this mixed-ANOVA. However, as there were conditions that fell outside of the acceptable range of skewness, we performed several robust ANOVAs to check whether this affected the results of the significance tests.

In the mixed-ANOVA, we found a significant main effect of Methods on IV $F(1.04, 7{,}489{,}711.875) = 233{,}724.04$, $p < 0.001$, $\eta^2 = 0.031$. Additionally, we found that the effect of Methods moderately depended on the size of the calibration sample $F(4.161, 7{,}489{,}711.875) = 105{,}253.12$, $p < 0.001$, $\eta^2 = 0.055$. The interaction between Methods and the reliability of the target test was small to medium $F(5.201, 7{,}489{,}711.875) = 71{,}839.617$, $p < 0.001$, $\eta^2 = 0.048$. In addition, the reliability of the baseline test was found to have a small effect on how the methods effect IV $F(5.201, 7{,}489{,}711.875) = 23{,}987.41$, $p < 0.001$, $\eta^2 = 0.016$. The above effects were also tested using robust ANOVAs (see Table 6 in Appendix C) and were found to be significant.

As seen from the results, most design factors influenced the effect of Methods on IV; however, collinearity between the tests (r_{12}) and predictability of the criterion (R^2) had negligible impact, as the effect sizes of their interaction effects with Methods were below 0.01. Next we interpret the interaction effects between Methods and design factors with effect sizes above 0.01.


(a)
$\hat{b}_0$
$\hat{b}_1$
$\hat{b}_2$
coefficient value
0.0
0.1
0.2
0.3
0.5
0.6
0.7
0.8
0.9
1.0
ρ_2


(b)
$\hat{b}_0$
$\hat{b}_1$
$\hat{b}_2$
coefficient value
0.0
0.1
0.2
0.3
0.5
0.6
0.7
0.8
0.9
1.0
ρ_1

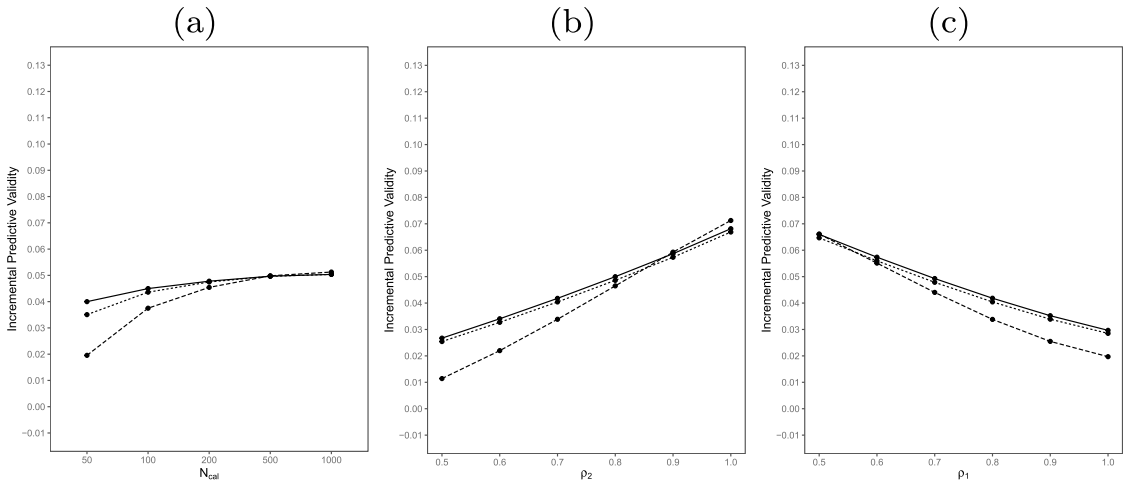


Fig. 4 Incremental predictive validity (ΔMSEP) against (**a**) calibration sample size (N_{cal}), (**b**) against reliability of the target test (ρ_2), and (**c**)against reliability of the baseline test (ρ_1) as a function of the Methods (dotted = OLS, dashed = SIMEX, and bold = ridge)

Figure 4 shows the interaction effects between Methods and N_{cal}, ρ_2, and ρ_1, averaged over 500 replications and collapsed over the other design factors. We see larger differences in IV between estimation methods when the calibration sample was small (Fig. 4a) small reliability of the target test (Fig. 4b) and high reliability of the baseline test (Fig. 4c). Overall, the results show, although not substantial, ridge slightly enhanced the IV of the target test, and using the SIMEX lowered IV compared to OLS in most cases.

5 Empirical illustration

In this section, we illustrate how to evaluate out-of-sample incremental predictive validity in practice and compare it to the classical approach. We use data from Niessen et al. (2016), who studied the predictive validities of three psychological tests in predicting higher education performance for the psychology bachelor program at a Dutch University. These tests were created to select students.

The first test, called PSYCHOLOGY, is a trial studying test that mimics an exam for a course in the first year of the bachelor's program. The other two tests were specific skills test for English and Mathematics (further denoted as ENGLISH and MATH). Our analysis focuses on investigating the incremental predictive validity of these tests in predicting academic performance over and above high school grades. Academic performance is quantified as the grade point average at the end of the first year (FYGPA) with scores running from one to ten. This criterion was calculated using the average of all the course grades taken in the first year and was reported to have a reliability of 0.89 in a follow-up study (Niessen et al. 2018).

The baseline test, High School GPA (HSGPA), was computed by the authors by averaging reported course grades from high school (Niessen et al. 2016). In a follow-up study, also including this cohort, HSGPA was reported to have a reliability of 0.73 based on calculated intraclass correlations (Niessen et al. 2018). The trial

studying test (PSYCHOLOGY) consisted of 40 items with scores running from 0 to 40. This target test was reported to have a reliability of 0.81 based on Cronbach's alpha (Niessen et al. 2016). The scores for ENGLSH and MATH run from 0 to 20 and 0 to 30, respectively. The reliability of ENGLISH is 0.70 and of MATH 0.76 (Niessen et al. 2016).

In this reanalysis, we used a sub-sample ($n = 200$) of this data set based on one cohort which comprises students that completed the highest Dutch secondary level of education before entering university. We excluded one erroneous observation[2] from this sub-sample. In our data set, HSGPA is positively correlated with each of the three new tests, correlation with PSYCHOLOGY equals 0.45 ($p < 0.001$), with ENGLISH 0.30 ($p < 0.001$), and with MATH 0.36 ($p < 0.001$).

To assess incremental predictive validity for, say, the PSYCHOLOGY test, two statistical models need to be compared:

$$\text{Model 1} : \text{FYGPA} = b_0 + b_1\text{HSGPA} + \epsilon,$$
$$\text{Model 2} : \text{FYGPA} = b_0 + b_1\text{HSGPA} + b_2\text{PSYCHOLOGY} + \epsilon.$$

In the classical, within-sample approach, we fit these two models and examine the change in MSE, ΔR^2, and its F-statistic with corresponding p-value to assess the incremental predictive validity. A residuals check for both models that were estimated using OLS was done and showed no severe violations of assumptions.

In the out-of-sample approach, we only considered prediction rules estimated using OLS and the ridge, as the simulations showed that the SIMEX method works poorly to assess incremental predictive validity. Because we only have a single sample, K-fold cross-validation is needed. To estimate the MSEP of both models, a tenfold cross-validation process was employed for the OLS method and a nested tenfold cross-validation for the ridge method. Nested cross-validation is suggested for methods that require choosing a penalty parameter (Varma and Simon 2006) as in our case for ridge regression. In every cycle of the 10-fold cross-validation, we trained the model on 9 parts and validate on 1. During the training, we need to find the optimal penalty parameter, which needs cross-validation again. So, in the training set, consisting of 9 parts of the data, we used tenfold cross-validation to obtain an optimal penalty parameter. Then the regression model with this optimal penalty parameter was fitted again on 9 parts. Based on the estimated parameters, a prediction rule was created and used for predicting criterion scores in the left out validation part. The final estimate of the prediction error is the average prediction errors of all folds.

Both cross-validation processes were repeated 100 times as recommended by Harrell (2015) and De Rooij and Weeda (2020). For every round of cross-validation, IV was estimated by subtracting the average squared prediction error of Model 2 from Model 1 (as in Eq. 6). The 100 repetitions provided 100 measures of incremental predictive validity for each estimation method. As in De Rooij and Weeda (2020), we use boxplots to visualize incremental predictive validity.

[2] Value of one for first-year grade point average was available while grades on first-year courses were not.

Table 3 Assessment of incremental predictive validity for each of the three tests

	ΔMSE	ΔR^2	F	p
Psychology	0.060	0.057	18.800	< 0.001
English	0.005	0.005	1.532	0.217
Math	0.011	0.011	3.236	0.074

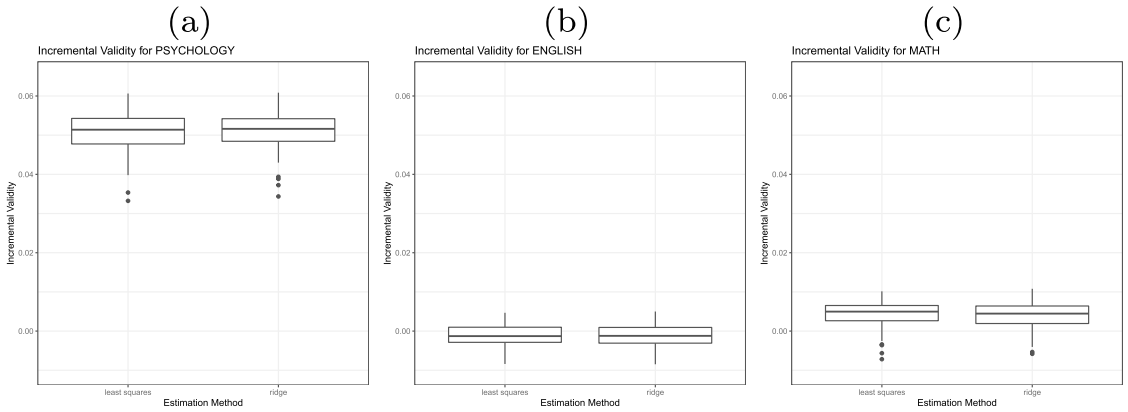


Fig. 5 Incremental predictive validity (ΔMSEP) for (**a**) PSYCHOLOGY, (**b**) ENGLISH, and (**c**) MATH. Each as a function of Estimation methods (least squares and ridge)

5.1 Results

Let us start with the classical, within-sample approach of assessing incremental predictive validity. For each of the three new tests, we fitted two models. The first model only has HSGPA as predictor, the second has HSGPA and the new test as predictors. Table 3 shows the change in MSE, the change in R^2, the incremental F-test (each with $df = (1, 197)$), and its associated p-value. The table shows that there is incremental predictive validity for PSYCHOLOGY. For MATH, the p-value is 0.07, showing it is close to significant, whereas for ENGLISH there is no evidence of incremental predictive validity.

For the out-of-sample approach, the incremental predictive validity of the three tests was estimated by the ridge method and OLS. The results are shown in Fig. 5. For these three tests, the estimation method does not make a large difference. For the PSYCHOLOGY test, the average incremental predictive validity (rounded to the third decimal) equals ΔMSEP = 0.051, for ENGLISH ΔMSEP = −0.001, and for MATH ΔMSEP = 0.004. The boxplots show that there is some variability around these averages. For PSYCHOLOGY, in every of the 100 repetitions of the repeated cross validation ΔMSEP is positive for both estimation methods. For ENGLISH, only 33 (or 35) of the repetitions lead to incremental predictive validity using least squares (ridge), and for MATH in 89 out of the 100 repetitions we found incremental predictive validity. Also from the boxplots we can tell that for both ENGLISH and MATH there is no real incremental predictive validity, where for ENGLISH the box includes zero, for MATH the whiskers include zero.

For PSYCHOLOGY on the other hand, it seems there is incremental predictive validity. However, before jumping to final conclusions we need to see how much better the predictions really become. The root mean squared error of prediction is an easier method to interpret, as it is on the same scale as the original measurements. The average RMSEP for Model 1 equals 0.8358 and for Model 2 0.8048. That means that by including the PSYCHOLOGY test, the individual predictions for the First Year Mean Grade become on average 0.031 better. As grades for courses often use only a single decimal, one might wonder whether this difference is worth the trouble.

6 Discussion

The main interest of incremental predictive validity research is to see whether adding a new test increases the accuracy of predictions of a given criterion. Usually researchers evaluate the incremental predictive value within the sample, that is, they use the same sample to estimate the prediction models and to evaluate the outcome. Such a procedure might be prone to overfitting. Therefore, out-of-sample evaluation of incremental predictive validity might be better.

In this paper, we investigated such an out-of-sample assessment of incremental predictive validity. First, we performed a simulation study to compare the within-sample approach to the out-of-sample approach to assess incremental predictive validity. In this simulation study, we also investigated the relationships between the reliabilities of the baseline and target test on incremental predictive validity for both approaches. Finally, we compared the out-of-sample assessment of incremental predictive validity using three different estimation methods and showed how to implement the out-of-sample approach to assess incremental predictive validity in practice.

Overall, the results of the simulation study showed that there was a difference between using the within- and out-of-sample approach to assess incremental predictive validity. The within-sample assessment overestimated the out-of-sample assessment for small sample sizes. In smaller samples, out-of-sample predictions from models 1 and 2 have more bias (more error) than those from larger samples. Therefore, the IV of the target test is smaller than what it could have been when estimated with a larger sample. With larger samples, the out-of-sample predictions are assumed to be more accurate. Thus, the IV of the target test is estimated to be larger than that of smaller samples.

We also found that the reliabilities of the baseline test and the target test affected the incremental predictive validity of the target test. However, these effects did not differ between the within- and out-of-sample approaches. This is actually good news, because all our knowledge on the relationship between reliability and predictive or incremental predictive validity is based on within sample reasoning. This knowledge is now also applicable for out-of-sample assessment of incremental predictive validity. We suspect that only sampling error played a role in the difference between in- and out-of-sample approaches of IV as the reliabilities of the baseline and the target test were equal in both calibration and validation samples.

The simulation study also showed that out-of-sample incremental predictive validity depends on how we estimate the prediction rule. In most cases, using the SIMEX method for estimating the prediction rules leads to a poor assessment of incremental predictive validity compared to OLS, especially when the test(s) had weak reliabilities. This is partly in line with Carroll et al. (2006) who already suggested using the original rules estimated from the error-prone tests. A plausible reason is that the test scores used for predicting in new samples are error-prone. Therefore, if the estimated regression weights are corrected for measurement error, there is a mismatch between the estimated coefficients with the test scores in the new sample, which can lead to more error in the predictions. In addition, the SIMEX method requires additional parameters to estimate: reliability of the tests and an extrapolation function, which adds more variance that may lead to more prediction error. Note that in the simulation study, reliability of the tests were known and therefore not estimated (in practice they are estimated), but even then SIMEX performed worst for the majority of cases.

Furthermore, the simulation results showed that the ridge method slightly enhanced incremental predictive validity compared to the OLS and SIMEX in small samples. In other words, the results showed that the way the weights of the prediction rule are estimated does matter. This finding is in contrast with the suggestion in Wainer (1976) to standardize predictors and use unit weights, but in line with Sackett et al. (2017) who showed that unit weighting in combination with standardization is not recommended for estimating incremental predictive validity.

In large samples, the assessed incremental predictive validity was the same for ridge and OLS rules. A plausible explanation for this finding may lie in the ratio between sample size (N) and the number of predictors (P). Although using biased estimates in prediction rules has been noted to be beneficial in multiple regression (Darlington 1978), as N/P increases, the penalty parameter γ may approach zero. As a consequence, the rules obtained from the ridge method become equivalent to the rules from OLS (McNeish 2015). This is also a plausible explanation for finding equal estimates of the incremental predictive validity using the OLS and ridge method in the application example.

In the application example, we showed how to assess incremental predictive validity using the out-of-sample approach. Therefore, we tested the incremental predictive validity of three new student selection tests. In a classical approach, one showed incremental predictive validity, for another test (ENGLISH), there was no evidence of incremental predictive validity, and for the third test (MATH), there is small evidence of incremental predictive validity (i.e., a marginally significant effect). In the out-of-sample approach, only the PSYCHOLOGY test showed incremental predictive validity. However, if we were to use this test, the predictions of academic achievement would only be 0.031 points better. Conceptually, by finding the decrease in prediction error alone one could confirm the decision to utilize the test for selection in addition to high-school grades. This does not imply that we should refrain from assessing other factors that might contribute to test utility (Hunsley and Meyer 2003). The decision to include the trial studying test in the selection procedure requires a balancing act across a variety of factors. For example, the institute should rate the improvement of 0.031 points in the prediction of academic

achievement against the labor costs of test maintenance, incurred cost of administering the test, and the effort of the applicants or test takers.

In short, we argued that incremental predictive validity should also be assessed from an out-of-sample perspective. Such a perspective protects against overfitting and provides more detail in what we actually gain (0.031 points *versus* 6% extra explained variance). Especially for smaller samples, the out-of-sample approach seems valuable. Furthermore, we investigated the link between the reliability of the baseline and the target test to incremental predictive validity and found that this link is the same irrespective of the within- or out-of-sample approach to assess incremental predictive validity.

Appendix A

In this section, we performed a small simulation experiment to demonstrate how different prediction rules can have identical R^2 but different mean squared errors when evaluated using out-of-sample data. We generated an outcome Y using the following model

$$Y = b_0 + b_1 T + \epsilon$$

where T contain true scores draw from a standard normal distribution, $b_0 = 0$, $b_1 = 0.2$, and ϵ, denoting the error drawn from a normal distribution with zero mean and variance of $\sqrt{1 - b_1^2}$.

Instead of observing T, we often observe $X = T + E$. In this example the reliability of X is set to $\rho = 0.9$, using 20, we can calculate σ_E^2. The population size was set to 10,000 from which we repeatedly draw calibration samples of size 100 and fit the following regression model using the OLS method:

$$Y^{cal} = b_0 + b_1 X^{cal} + \epsilon.$$

We obtain the estimated coefficients (i.e., $\hat{b}_0^{ols}$ and $\hat{b}_1^{ols}$) of the prediction rule and used these coefficients to compute the predicted values on a validation sample of 1000 that was also drawn from the same population. This process was repeated 100 times. Furthermore, we created different prediction rules by crudely applying a shrinkage factor on the weight of OLS $\hat{b}_1^{ols}$. The predicted values in the validation sample are given by

$$\hat{Y}^{val} = \hat{b}_0 + \hat{b}_1^* X^{val}$$

with

$$\hat{b}_1^* = s * \hat{b}_1^{ols}$$

where s is the shrinkage factor ranging from 0 to 1.1 in increments of 0.1. Note that when $s = 1$, we simply apply the prediction rule from OLS, $\hat{b}_1^* = \hat{b}_1^{ols}$. When $s < 1$,

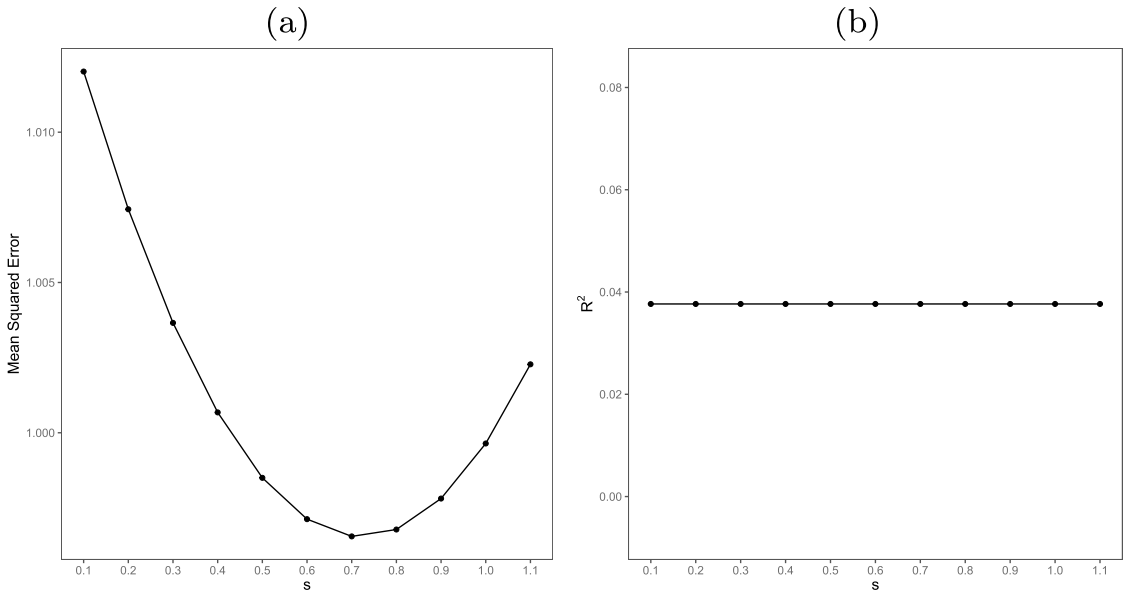


Fig. 6 Out-of-sample (**a**) Mean squared error (MSEP) and (**b**) R^2 against various shrinkage values. As $s < 1$ more shrinkage is applied with $s = 1$ equal to the solution from OLS

$\hat{b}_1^* < \hat{b}_1^{ols}$. Fixing $s = 1.1$ resembles performing a correction on the estimated coefficient when $\rho = 0.9$. Thus, $\hat{b}_1^* > \hat{b}_1^{ols}$.

Figure 6 show the aggregated results over 100 repetitions. In this figure, the R^2 calculated in the validation sample remains identical for each shrinkage factor, but the mean squared errors vary as a function of this factor. Notice also in Fig. 6a, the solution with the smallest average mean squared error is at $s = 0.7$, suggesting a benefit from shrinking the coefficient.

Appendix B

In this section, we describe the process of obtaining the regression coefficients b_1 and b_2 in the simulation study. In the simulation study, the standard deviations of true scores T_1 and T_2, and outcome variable were fixed to one ($\sigma_{T_1} = \sigma_{T_2} = \sigma_Y = 1$). Thus, the covariance between the true scores is equivalent to the correlation of these true scores $\sigma_{12} = r_{12}$ and $R^2 = var(\hat{Y})$. The squared multiple correlation is defined as

$$R^2 = b_1^2 + b_2^2 + 2b_1 b_2 r_{12}. \tag{21}$$

By knowing 21 and the four conditions of the ratio of the regression coefficients $(b_1 : b_2) \in (2 : 1, 1 : 1, 1 : 2, 1 : 0)$, we can find the regression coefficients b_1 and b_2 in two steps. The first step is to find the expression for b_1, given certain levels of R^2, r_{12}, ratio of b_1 and b_2. For a summary of these expressions for b_1, see Table 4. The second step is to compute calculate b_2 by knowing b_1. Note that in the fourth condition of the ratio of the regression coefficients (1:0), which means that $b_1 = R$, and therefore $b_2 = 0$.

Table 4 Summary of analytical expressions for b_1 for different ratios between b_1 and b_2

Ratio	Expression
$b_1 = b_2$	$b_1 = \sqrt{\frac{R^2}{2+2r_{12}}}$
$b_1 = 2b_2$	$b_1 = \sqrt{\frac{R^2}{\frac{5}{4}+r_{12}}}$
$2b_1 = b_2$	$b_1 = \sqrt{\frac{R^2}{5+4r_{12}}}$

Appendix C

In this section, we list R-code to apply the robust ANOVAs that were highlighted in this paper ($\eta^2 > 0.01$). These robust ANOVAs allow us to examine whether our results were affected by severe skewness of the residuals. Robust ANOVAs were performed in R using the WRS2 package (Mair and Wilcox 2020) and we used $\alpha = 0.05$ for significance testing. The function |bwtrim()| fits a mixed-ANOVA model with one between- and one within-subjects effects and |t2way()| fits a two-way ANOVA model. For the within-subjects effects in Part I and Part II of the simulation study, we fitted robust mixed-ANOVA models to test the differences in IV. For the between-subjects effects in Part I (see Table 2), we fitted two-way between-subjects robust ANOVAs on the differences in the average of the IV estimated in- and out-of-sample from OLS.

Table 5 Summary of the robust ANOVAs that were performed to test several effects from mixed-ANOVA in Part I

Between/Within-Subjects	Effect	Robust ANOVA
within-subjects	in/out-of sample	`bwtrim(IV_MSE~N*sample, id=ndesign, data=mydata)`*
between-subjects	$(b_1 : b_2)$ x r_{12}	`t2way(ave_IV_MSE~ratio*r12, data=mydata)`*
between-subjects	$(b_1 : b_2)$ x R^2	`t2way(ave_IV_MSE~ratio*R2, data=mydata)`*
between-subjects	ρ_2 x $(b_1 : b_2)$	`t2way(ave_IV_MSE~rho2*ratio, data=mydata)`*
between-subjects	ρ_1 x r_{12}	`t2way(ave_IV_MSE~rho1*r12, data=mydata)`*
between-subjects	ρ_2 x R^2	`t2way(ave_IV_MSE~rho2*R2, data=mydata)`*
between-subjects	ρ_1 x R^2	`t2way(ave_IV_MSE~rho1*R2, data=mydata)`*
between-subjects	r_{12} x R^2	`t2way(ave_IV_MSE~r12*R2, data=mydata)`*

* = the model also tests the main effects.

Table 6 Summary of the robust ANOVAs that were performed to test several effects from the mixed-ANOVA in Part II

Between/Within-Subjects	Effect	Robust ANOVA
within-subjects	ρ_1 x Methods	`bwtrim(IV_MSE~rho1*Estimates, id=ndesign, data=mydata)`*
within-subjects	ρ_2 x Methods	`bwtrim(IV_MSE~rho2*Estimates, id=ndesign, data=mydata)`*
within-subjects	N x Methods	`bwtrim(IV_MSE~N*Estimates, id=ndesign, data=mydata)`*

* = the model also tests the main effects.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s41237-024-00224-7.

Acknowledgements We thank Susan Niessen for providing the data and thank two anonymous reviewers for their helpful comments.

Funding The authors have not disclosed any funding.

Data availability Data of the empirical example can be found in the referred article https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0153663.

Declarations

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

Breiman L (2001) Statistical modeling: the two cultures (with comments and a rejoinder by the author). Stat Sci 16(3):199–231. https://doi.org/10.1214/ss/1009213726

Browne MW (2000) Cross-validation methods. J Math Psychol 44(1):108–132. https://doi.org/10.1006/jmps.1999.1279

Carroll RJ, Ruppert D, Stefanski LA, Crainiceanu CM (2006) Measurement error in nonlinear models: a modern perspective. CRC Press, Boca Raton

Chapman BP, Weiss A, Duberstein PR (2016) Statistical learning theory for high dimensional prediction: application to criterion-keyed scale development. Psychol Methods 21(4):603. https://doi.org/10.1037/met0000088

Cohen J, Cohen P, West SG, Aiken LS (2013) Applied multiple regression/correlation analysis for the behavioral sciences. Routledge, London

Cook JR, Stefanski LA (1994) Simulation-extrapolation estimation in parametric measurement error models. J Am Stat Assoc 89(428):1314–1328. https://doi.org/10.1080/01621459.1994.10476871

Darlington RB (1968) Multiple regression in psychological research and practice. Psychol Bull 69(3):161–182. https://doi.org/10.1037/h0025471

Darlington RB (1978) Reduced-variance regression. Psychol Bull 85(6):1238–1255. https://doi.org/10.1037/0033-2909.85.6.1238

De Rooij M, Weeda W (2020) Cross-validation: a method every psychologist should know. Adv Methods Pract Psychol Sci 3(2):248–263. https://doi.org/10.1177/2515245919898466

Evers A, Lucassen W, Meijer R, Sijtsma K (2010a) Cotan beoordelingssysteem voor de kwaliteit van tests. [COTAN Assessment system for the quality of tests]. Amsterdam, Netherlands: Nederlands Instituut van Psychologen

Evers A, Sijtsma K, Lucassen W, Meijer RR (2010b) The Dutch review process for evaluating the quality of psychological tests: history, procedure, and results. Int J Test 10(4):295–317. https://doi.org/10.1080/15305058.2010.518325

Grove WM, Zald DH, Lebow BS, Snitz BE, Nelson C (2000) Clinical versus mechanical prediction: a meta-analysis. Psychol Assess 12(1):19. https://doi.org/10.1037/1040-3590.12.1.19

Hair J Jr, Hair JF Jr, Hult GTM, Ringle CM, Sarstedt M (2021) A primer on partial least squares structural equation modeling (PLS-SEM). Sage Publications, Thousand Oaks

Harrell FE Jr (2015) Regression modeling strategies, 2nd edn. Springer, New York

Hastie T, Tibshirani R, Friedman J (2009) The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, New York

Hoerl AE, Kennard RW (1970) Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12(1):55–67. https://doi.org/10.2307/1271436

Hunsley J, Meyer GJ (2003) The incremental validity of psychological testing and assessment: conceptual, methodological, and statistical issues. Psychol Assess 15(4):446–455. https://doi.org/10.1037/1040-3590.15.4.446

IBM Corp. (2020) IBM SPSS statistics for Windows Version 27. IBM Corp., Armonk

Kim HY (2013) Statistical notes for clinical researchers: assessing normal distribution (2) using skewness and kurtosis. Restor. Dent. Endod. 38(1):52–54

Kline R (2015) Principles and practice of structural equation modeling, fourth edition. Methodology in the social sciences. Guilford Publications. https://books.google.nl/books?id=Q61ECgAAQBAJ

Kraemer N, Schaefer J, Boulesteix AL (2009) Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. BioMed Cent Bioinform. https://doi.org/10.1186/1471-2105-10-384

Lederer W, Seibold H, Küchenhoff H (2017) SIMEX: SIMEX- and MCSIMEX-algorithm for measurement error models. R Package Version 1.7

Lord M, Novick MR (1968) Statistical theories of mental test scores. Addison-Wesley, Oxford

Mair P, Wilcox R (2020) Robust statistical methods in R using the WRS2 package. Behav Res Methods 52:464–488

McNeish DM (2015) Using lasso for predictor selection and to assuage overfitting: a method long overlooked in behavioral sciences. Multivar Behav Res 50(5):471–484. https://doi.org/10.1080/00273171.2015.1036965

Meehl PE (1954) Clinical versus statistical prediction: a theoretical analysis and a review of the evidence. University of Minnesota Press, Minneapolis. https://doi.org/10.1037/11281-000

Mosier CI (1951) The need and means of cross validation. i. problems and designs of cross-validation. Educ Psychol Meas 11(1):5–11. https://doi.org/10.1177/001316445101100101

Niessen ASM, Meijer RR, Tendeiro JN (2016) Predicting performance in higher education using proximal predictors. PLoS ONE 11(4):e0153663. https://doi.org/10.1371/journal.pone.0153663

Niessen ASM, Meijer RR, Tendeiro JN (2018) Admission testing for higher education: a multi-cohort study on the validity of high-fidelity curriculum-sampling tests. PLoS ONE 13(6):e0198746. https://doi.org/10.1371/journal.pone.0198746

R Core Team (2022) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna

Sackett PR, Dahlke JA, Shewach OR, Kuncel NR (2017) Effects of predictor weighting methods on incremental validity. J Appl Psychol 102(10):1421. https://doi.org/10.1037/apl0000235

Schmidt FL, Hunter JE (1998) The validity and utility of selection methods in personnel psychology: practical and theoretical implications of 85 years of research findings. Psychol Bull 124(2):262. https://doi.org/10.1037/0033-2909.124.2.262

Sechrest L (1963) Incremental validity: a recommendation. Educ Psychol Meas 23(1):153–158. https://doi.org/10.1177/001316446302300113

Shmueli G (2010) To explain or to predict? Stat Sci 25(3):289–310

Spearman C (1904) The proof and measurement of association between two things. Am J Psychol 15(1):72–101. https://doi.org/10.2307/1422689

Stone M (1974) Cross-validatory choice and assessment of statistical predictions. J Roy Stat Soc Ser B (Methodol) 111–147. https://doi.org/10.1111/j.2517-6161.1974.tb00994.x

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J Roy Stat Soc Ser B (Methodol) 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Van Houwelingen J, Le Cessie S (1990) Predictive value of statistical models. Stat Med 9(11):1303–1325. https://doi.org/10.1002/sim.4780091109

Van Loon W, Fokkema M, Szabo B, De Rooij M (2020) Stacked penalized logistic regression for selecting views in multi-view learning. Inf Fusion 61:113–123

van Wieringen WN (2021) Lecture notes on ridge regression. arXiv preprint arXiv:1509.09169

Varma S, Simon R (2006) Bias in error estimation when using cross-validation for model selection. BioMed Cent Bioinform 7(1):91. https://doi.org/10.1186/1471-2105-7-91

Wainer H (1976) Estimating coefficients in linear models: it don't make no nevermind. Psychol Bull 83(2):213. https://doi.org/10.1037/0033-2909.83.2.213

Westfall J, Yarkoni T (2016) Statistically controlling for confounding constructs is harder than you think. PLoS ONE 11(3):e0152719. https://doi.org/10.1371/journal.pone.0152719

Yarkoni T, Westfall J (2017) Choosing prediction over explanation in psychology: lessons from machine learning. Perspect Psychol Sci 12(6):1–23. https://doi.org/10.1177/1745691617693393

Zou H, Hastie T (2005) Regularization and variable selection via the elastic net. J Roy Stat Soc Ser B (Stat Methodol) 67(2):301–320