Received: 13 November 2019 | Revised: 14 May 2021 | Accepted: 22 June 2021

DOI: 10.1111/1748-8583.12400

REVIEW/PROVOCATION

Human Resource Management Journal WILEY

Prediction in HRM research–A gap between rhetoric and reality

Marko Sarstedt[1,2] | Nicholas P. Danks[3]

[1]Otto-von-Guericke-University, Magdeburg, Germany

[2]Babeș-Bolyai University, Cluj-Napoca, Romania

[3]Trinity Business School, Trinity College, Dublin, Ireland

Correspondence

Marko Sarstedt, Otto-von-Guericke University Magdeburg, Faculty of Economics and Management, Universitaetsplatz 2, 39106 Magdeburg, Germany.
Email: marko.sarstedt@ovgu.de

Funding information

WOA Institution: OTTO VON GUERICKE UNIVERSITAET MAGDEBURG Blended DEAL: Projekt DEAL.

Abstract

There are broadly two dimensions on which researchers can evaluate their statistical models: explanatory power and predictive power. Using data on job satisfaction in ageing workforces, we empirically highlight the importance of distinguishing between these two dimensions clearly by showing that a model with a certain degree of explanatory power can produce vastly different levels of predictive power and vice versa—in the same and different contexts. In a further step, we review all the papers published in three top-tier human resource management journals between 2014 and 2018 to show that researchers generally confuse explanation and prediction. Specifically, while almost all authors rely solely on explanatory power assessments (i.e., assessing whether the coefficients are significant and in the hypothesised direction), they also derive practical recommendations, which inherently result from a predictive scenario. Based on our results, we provide HRM researchers recommendations on how to improve the rigour of their explanatory studies.

Abbreviations: AIC, Akaike Information Criterion; ARIMA, Autoregressive integrated moving average; AUC, Area under the curve; CFI, Comparative fit index; CV, Cross-validation; EP, Explanatory and predictive; ETS, Error trend seasonal; HRM, Human resource management; ISSP, International Social Survey Program; KNN, K-nearest neighbours; LOOCV, Leave-one-out cross-validation; MAE, Mean absolute error; RMSEA, Root mean square error of approximation; RMSE, Root mean square error; SMAPE, Symmetric mean absolute percentage error; SRMR, Standard root mean square residual; USA, United States of America

KEYWORDS
explanation, explanatory power, generalisability, prediction, predictive power, relevance

Practitioner notes

What is currently known?

- Explanation and prediction are two distinct statistical concepts.
- Establishing a model's predictive power is central to deriving practical recommendations.
- Research has brought forward numerous procedures for evaluating a model's explanatory and predictive power.

What this paper adds?

- An empirical example demonstrates the interplay between explanatory and predictive power.
- A review of prior human resource management (HRM) research shows that few researchers consider the predictive power of their models.
- HRM researchers generally confuse the concepts of explanation and prediction.
- The paper offers concrete recommendations on how to routinely implement predictive model evaluations in statistical analyses.

The implications for practitioners

- Results of research papers should be interpreted in the context of the analyses (explanation vs. prediction).
- Managerial recommendations derived from purely explanatory analyses should be challenged in terms of their predictive relevance.
- Predictive analyses should supplement explanatory modelling in making research practical and relevant to practice.

1 | INTRODUCTION

Socio-technical systems within the purview of human resource management (HRM) research are inherently complex due to their intricate underlying causal interactions and processes. Models are formal quantitative representations of theories and hypotheses that are devised to offer partial explanations of these complex systems (Lauenroth, 2003). HRM researchers build models to approximate these processes and apply inferential statistical methods to data in order to test the relationships between the assumed causes and effects. Specifically, the researchers' focus is on assessing whether the model fits the data and whether the model coefficients are significant and in the hypothesised direction. This approach, which is widely known as explanatory modelling, differs from predictive modelling. Predictive modelling applies estimated model parameters, generated from a sample at hand, to generate predictions for other observations not used in the model estimation. These observations can be kept separate from the main sample or they can be collected at a future point in time or even in another context (Shmueli et al., 2016).

While there is no scientific basis for choosing one form over the other (Hofman et al., 2017), the statistics literature has long established that the distinction between explanation and prediction has important consequences for all steps along the research process (e.g., Dowe et al., 2007; Forster, 2002; Forster & Sober, 1994; Shmueli, 2010; Shmueli & Koppius, 2011). In explanatory modelling, for example, the variable choice is based on the theorised model structure and on the operationalisation of the concepts under investigation. Conversely, in predictive modelling, the focus is on the association between dependent and independent variables without inferring an underlying causal structure—as is the case in many machine learning applications (Franklin, 2005).

Further differences arise when interpreting the findings derived from models that are validated by means of an explanatory rather than a predictive lens. With their focus on theory and inference, models with high explanatory power allow for translating statistical results regarding hypothesised relationships into scientific conclusions. Conversely, models with high predictive power allow researchers to make claims regarding their generalisability and practical performance. While carefully designed explanatory models can potentially achieve high levels of predictive power, it cannot be assumed to be true and therefore needs to be explicitly tested (Shmueli, 2010). A model with high explanatory power can perform very poorly in terms of prediction, and a model with high predictive power can perform very poorly in terms of explanation (Konishi & Kitagawa, 2007). Hence, the conceptual differences between explanation and prediction have important implications that go beyond a more careful choice of language. Specifically, these differences dictate the extent to which inferential conclusions can be drawn and our ability to generate practical prescriptions from our analyses.

In this paper, we empirically demonstrate the interplay between explanatory and predictive (EP) power by replicating Drabe et al.'s (2015) model on job satisfaction in ageing workforces. Specifically, we show that a model with a certain degree of explanatory power can produce vastly different levels of predictive power and vice versa. In addition, our results highlight that a model developed in one context, for example, time of data collection or country, does not necessarily generalise well to other contexts in terms of predictive power. In a further step, we show that HRM researchers generally confuse the explanation and prediction concepts when evaluating and deriving implications from their models. Specifically, our review of all the quantitative papers published in three top-tier HRM journals demonstrates that the majority of authors follows an explanatory perspective in their analysis. Despite this focus on association-based modelling, the authors' discussion of their results and the managerial implications suggest that their findings hold in a predictive scenario and generalise beyond the data and specific context at hand. In consideration of these results, we conclude that HRM researchers should supplement their exclusive focus on explanatory power with an evaluation of the predictive power and provide guidelines for implementing predictive model evaluations in HRM research.

2 | MODEL EVALUATION: EXPLANATORY AND PREDICTIVE POWER

There are broadly two dimensions according to which researchers can evaluate their models: explanatory power and predictive power. Explanatory power refers to 'the strength of association indicated by a statistical model' (Shmueli & Koppius, 2011, p. 561). Researchers evaluate their models' explanatory power based on *F*-type metrics and the R^2 (Cohen, 1988), followed by an assessment of the model coefficients in terms of their significance, direction and size. The model coefficients are typically of primary concern as they represent the hypotheses of interest. These coefficients and metrics have a common ground in that they are estimated and evaluated using a single common dataset—they are therefore usually referred to as in-sample metrics. In a number of methodological applications, an assessment of model fit, which describes the model's ability to reasonably approximate the underlying data-generating process, precedes the explanatory power's assessment. Most notably, structural equation modelling strongly emphasises the model fit concept established via the χ^2 statistic or alternative fit indices, such as the Comparative Fit Index (CFI), root mean square error of approximation (RMSEA) and standard root mean square residual (SRMR) (Bagozzi & Yi, 2012). These in-sample metrics quantify the divergence between the empirical

covariance matrix and the model-implied covariance matrix as an indication of how well the hypothesised model fits the data. A number of statisticians would consider these metrics as indicative of a model's descriptive power, that is, as 'summarizing or representing the data structure in a compact manner' (Shmueli, 2010, p. 3). However, model fit tests have become standard for explanatory model validation (e.g., Bagozzi & Yi, 2012), and we therefore consider the use of model fit metrics as an element of explanatory theory testing.

A common misconception in the social sciences is that of viewing in-sample metrics, such as the R^2 or the size of the estimated parameters in a regression, as indicative of predictive power. Shmueli and Koppius (2011, p. 562) note that 'a model with a very high R^2 indicates a strong relationship within the data used to build that model, but the same model might have very low predictive accuracy in practice'. Analogously, a well-fitting model—in a χ^2 sense—can perform very poorly when it comes to prediction (Rigdon, 2012). Such low predictive power can result from an overly complex model that overfits the data by tapping spurious patterns (i.e., random error) in the specific sample (Makridakis et al., 2009). A model which overfits can cause misleading confidence in model quality or generalisability—thus, model fit statistics should be accompanied by predictive metrics unless the research focus is purely on understanding the relationships between the variables of interest (Shmueli, 2010).

Assessing a model's predictive power generally requires separating the full sample of data into training and validation subsets. The training set, generally 70%–80% of the original data, is used to estimate the model parameters, for example, the beta weights in a regression model. The validation set, the data not used in the training set, is then used to evaluate the estimated model's predictive power (Hastie et al., 2013). The observations in the validation set are usually selected randomly, except in a specific context, for example in time series, where the observations are chosen as the series' last periods (Bergmeir & Benitez, 2012). Alternatively, researchers can assess their model's predictive power by using newly collected data. This approach has the advantage that it avoids estimating and evaluating the model on data that may have sample-specific patterns, which might bias the prediction error estimates.

Combining the model estimates generated from the training set and the validation set's observations of the independent variables, allows producing observation-level predictions for the validation sample's dependent variables. Researchers can then assess their models' predictive power by using out-of-sample metrics that quantify the divergence between the actual and the predicted values in the holdout data.

When the outcome variable of interest is continuous, predictions can be evaluated by using metrics such as the mean absolute error (MAE), root mean square error (RMSE) and symmetric mean absolute percentage error (Hastie et al., 2013). When the outcome variable is non-continuous, such as multi-class or binary, researchers frequently use misclassification or hit rates, confusion matrices and area under the curve to quantify the out-of-sample error rate (Japkowicz & Shah, 2011).

The most commonly used metric in an out-of-sample context is the RMSE, which expresses the standard deviation of the prediction error in the same scale as the outcome variable. If the error is normally distributed, researchers can apply the 65-95-99 rule according to which 65% of all predictions will fall within the interval of ± 1 RMSE of the true value (Chai & Draxler, 2014); intervals of ± 2 and ± 3 produce the corresponding 95% and 99% intervals, respectively. For example, if a respondent's true value for job satisfaction was 5, measured on a scale from 1 to 7; the model with an RMSE of 0.7 would predict a job satisfaction between 4.3 and 5.7 in 65% of the cases.

The MAE also features prominently in research because of its intuitive interpretation—it quantifies the average error made by a model in the scale of the outcome variable (Hyndman, 2006). For example, if a model predicts an annual salary with a MAE of 5000 and this model generates a prediction for an individual of $50,000, we can quantify the error related to this prediction as being $\pm$$5000 on average.

Single point estimates of prediction error can be subject to large bias and variance (Hastie et al., 2013). In order to avoid such issues, researchers can draw on cross-validation (CV) methods, which reuse part of the dataset to estimate the prediction error. A popular method is *k*-fold CV, which splits the dataset into *k* equally sized subsets and then iteratively combines $k-1$ subsets in training samples to predict each remaining subset. The dataset is generally split randomly, but there are variants, such as stratified CV, in which researchers can enforce a certain data distribution in each subset. A subset of *k*-fold CV, where *k* is set equal to the number of observations, is called

leave-one-out CV (LOOCV) (Burman, 1989). LOOCV is particularly useful in the case of small sample sizes, as it generates predictions using all the data, thereby averaging out the sampling error, which could be considerable when using only one very small validation sample. For a thorough discussion of CV techniques and their relative strengths and merits, see Arlot and Celisse (2010).

Another popular metric of a model's predictive power is the Akaike Information Criterion (AIC) (Akaike, 1973). Based on the concept of the Kullback–Leibler distance, the AIC quantifies the relative amount of information lost when comparing a model to the unknown true model that generates the observed data and thereby defines the correlation patterns between the variables of interest (Burnham & Anderson, 2002). Research has produced numerous variants and extensions of the AIC, for example, the AIC_3, AICu and AICc (McQuarrie & Tsai, 1998; Sugiura, 1978), which follow the same conceptual paradigm. Although these metrics' estimation draws on the entire sample data (i.e., they are in-sample metrics), they approximate a model's predictive power well. However, their use is confined to model comparisons; that is, the AIC and related metrics do not allow a stand-alone assessment of a model's predictive power. Consequently, choosing a model from a set of competing models based on, for example the AIC, does not guarantee that the selected model has good absolute predictive power.

Importantly, all metrics for assessing a model's predictive power—whether used for a stand-alone assessment or in the context of model comparisons—can be computed for any statistical model, regardless of whether the model was derived with an explanatory or predictive focus.

3 | IMPROVING HRM RESEARCH THROUGH PREDICTIVE METHODS

Gregor (2006) describes three components of theory: (1) *Causal explanations*, which are statements derived from causal reasoning which describe the relationships among phenomena; (2) *Testable propositions* (hypotheses), which are statements of relationships between constructs in such a form that they can be tested empirically and (3) *Prescriptive statements*, which are statements in the theory specifying how outcomes can be achieved in practice. Gregor (2006) further suggests that the term explanatory statistical modelling be used for the purpose of describing the testing of hypotheses that specify *how*, *why* and *when* certain empirical phenomena occur. In contrast, predictive modelling serves the purpose of generating (prescriptive) statements about *what will be*, but not why. That is, explanatory modelling is concerned with why things happen, while predictive modelling is concerned with describing what will happen. Correspondingly, researchers engaging in explanatory modelling are primarily concerned with the significance, direction and size of the coefficients, which represent the relationships hypothesised in the theoretical model (Shmueli & Koppius, 2011). On the contrary, predictive modelling is solely concerned with the prediction of one or more outcome variable(s) from a set of explanatory variables, without explaining the underlying causal connections between the variables involved (Shmueli, 2010). In its purest sense, predictive modelling implies that the mechanisms through which the predictions are generated are thought unnecessary, yielding models that include explanatory variables that are not supported by theory or sometimes even logic. The focus is on the model's ability to predict the outcome variables and not on assessing the model coefficients in terms of their significance, direction, and size. From this perspective, predictive modelling is by no means a substitute nor required when the focus is purely on understanding the model relationships.

However, these two forms of modelling are not mutually exclusive but can be applied in tandem in one empirical analysis by means of a balanced EP approach (Gregor, 2006; Sharma et al., 2021). When a researcher follows an EP approach, the goal of modelling is to select a model that not only offers theoretical explanations and testable propositions but also allows for deriving prescriptive statements. While the emphasis is placed on explanation as the primary goal of the empirical analyses, predictive analyses of such explanatory models need to be conducted in order to assess whether prescriptive statements can be readily derived from the results. Such a model would describe not only 'what is, how and why', but also 'what will be' (Gregor, 2006, p. 626), allowing for accurate and robust generalisable prescriptions to be made.

Supplementing an explanatory perspective with a predictive focus is of particular importance to HRM researchers given the very practical nature of the research and the direct impact on managerial decision making. Practical prescriptions—as characteristically documented in 'Managerial Implications' sections—inherently follow a predictive perspective (Hair & Sarstedt, 2021). For example, researchers frequently make conditional statements that foreshadow a specific result if a specific activity is implemented (i.e., prescriptive statements): 'In order to foster work engagement, organisations should offer interventions aimed at increasing personal resources'. Due to this mismatch between the explanation and prediction perspectives, researchers wanting to make practical, generalisable recommendations are required to conduct an additional assessment of their explanatory models' predictive power—even when the recommendations have been derived from an explanatory perspective (Shmueli & Koppius, 2011). Without such an additional assessment, it is questionable whether the model estimates produce similar results in terms of the outcome variables across time, samples and contexts.

Putting the EP approach into practice requires balancing model parsimony and explanatory power. While complex models might perform very well in terms of explanatory power, they run the risk of becoming overfit. In the case of overfit, a model with very high explanatory power might have substantially weaker predictive power because it fits the data sample at hand too closely (Hastie et al., 2013)—a characteristic that we will empirically demonstrate in the next section. Hence, researchers need to carefully evaluate their models' explanatory power and predictive performance (Aho et al., 2014; Sharma et al., 2019). In addition to evaluating overall model performance, researchers need to carefully evaluate the differential contribution of independent variables to their model's EP power. Non-zero, non-significant effects might yield improvements in out-of-sample predictive power and thus provide evidence for their inclusion in the model. Similarly, the model might contain independent variables that are statistically significant, but do not contribute to the predictive power of the model. Evaluating the impact of individual independent variables on predictive power by first evaluating predictive power with and then without the variable can provide evidence of practical usefulness beyond significance alone.

In summary, researchers should routinely seek to validate their models' predictive performance on further samples. If possible, researchers should collect additional samples that vary in either context (such as nature of company under analyses, or industry), time (such as different financial periods, or a second dataset several months later) and sample (such as sampling a second time from the same context and time). Such additional samples allow one to establish the generalisability of the model, thereby providing the basis for answering questions such as: 'What will the predicted staff turnover be in the next forecast period, and what is the uncertainty associated with that prediction?' Alternatively, researchers may randomly split their dataset into training and holdout samples, provided that the set is large enough to warrant sufficient levels of statistical power.

Finally, it is important to note that predictive analysis is no post-hoc panacea for poorly planned explanatory models or models that lack causal support. Predictive analysis will not ameliorate the need for designing studies and empirical analyses that provide accurate and sufficient scientific evidence to answer the specific research questions of interest to the researchers. A wrong model remains wrong. However, given that all models are likely to be wrong but may be useful approximations (Box, 1976, Box et al., 1978), it is all the more important to gauge the practical utility of such models and communicate this to the readers of research.

4 | PREDICTIVE POWER VS EXPLANATORY POWER: AN EMPIRICAL DEMONSTRATION

4.1 | Design and data

To demonstrate the interplay between EP power, we draw on Drabe et al.'s (2015) model on job satisfaction in ageing workforces. Using data from the 2005 International Social Survey Program (ISSP), these authors analysed

the impact of several situational antecedents on job satisfaction in the USA, Germany and Japan. We replicate their model using a more recent dataset from the ISSP gathered in 2015 (ISSP Research Group, 2017).[1]

To highlight the importance of reporting predictive power, we engage in a thought experiment that seeks to answer two questions:[2] (1) 'If two researchers were conducting their studies independently and collected data, how would their respective datasets perform in predicting each other's data?' and (2) 'How does the predictive power of these models compare to their explanatory power?' In a second step, we provide a practical demonstration of how predictive power can be estimated and evaluated across different contexts. To do so, we use Drabe et al.'s (2015, Table 4) model to predict the 2015 ISSP data and analyse the power of models developed in one country in predicting a sample from another country.

4.2 | A thought experiment

Drawing on the 2015 ISSP data collected in the USA and the model shown in Table A1, we demonstrate how predictive power and explanatory power can vary in one and the same model. We do so by randomly drawing two subsamples—each consisting of 250 observations—from the original dataset ($n = 915$) with 1000 replications.[3] One subsample serves as training sample, the other as holdout sample, both being collected within the same context and from the same population. This procedure emulates the collecting of data by two independent researchers in the same context and allows us to compare the predictive power of the model on an unseen, but contextually similar dataset. We fit the regression model to the training dataset and generate predictions for the holdout set. We first calculate the R^2 of the training samples and contrast them with the out-of-sample RMSE on the predictions generated using the holdout samples. In the next step, we calculate the RMSE using the training samples (i.e., the in-sample RMSE) as an alternative measure of the model's explanatory power and contrast it with the out-of-sample RMSE. Figure 1a plots the relationship between the (in-sample) explanatory power on the grounds of R^2 and the out-of-sample RMSE. Figure 1b plots the relationship between explanatory power using the in-sample RMSE and the out-of-sample RMSE. To perform our analyses, we use the caret package (Kuhn, 2020) for the R Statistical Environment (R Core Team, 2021). The code for the analyses of both the empirical example and thought experiment are included in Appendix B.[4]

The results presented in Figure 1 show that, for a given in-sample metric (R^2 or in-sample RMSE) value, the out-of-sample RMSE can take a broad range of different values. This finding indicates that there is no fixed relationship between explanatory power of a model fitted to a training dataset and predictive power of that model on a further holdout dataset. That is, two models with a very similar in-sample performance (R^2 or in-sample RMSE) can have very different levels of predictive power.

The relationship between R^2 and out-of-sample RMSE has a moderate positive correlation (Figure 1a; $\rho = 0.20$), indicating that a higher R^2 is associated with a higher out-of-sample RMSE. Similarly, the relationship between in-sample RMSE and out-of-sample RMSE has a moderate negative correlation (Figure 1b; $\rho = -0.33$), indicating that a lower in-sample RMSE (i.e., higher explanatory power) is associated with a higher out-of-sample RMSE. These results highlight the risk of overfit when a model has strong explanatory power at the cost of out-of-sample predictive power.

To further explore the degree of variation in predictive power, we focus on those replications, which produced a similar R^2 value to the original dataset ($R^2 = 0.444$).[5] Specifically, we select all training datasets from the 1000 replication runs with an R^2 value between 0.434 and 0.454 ($R^2 = 0.444 \pm 0.01;\ n^* = 152$) and analyse the predictive performance on the corresponding holdout sets. Figure 2 shows the distribution of RMSE for this model on these 152 subsamples.

The results show that the out-of-sample RMSE estimated on a holdout set for the samples ranges from 0.81 to 1.07—a 32.1% increase in uncertainty associated with the predictions generated from the model. Thus, models with seemingly very similar fit in terms of R^2 could be overly fitted to the idiosyncrasies (error or otherwise) of the data, and thus perform poorly when out-of-sample predictions are generated.

FIGURE 1 The relationship between (a) (in-sample) R^2 and out-of-sample root mean square error (RMSE); and (b) in-sample RMSE and out-of-sample RMSE

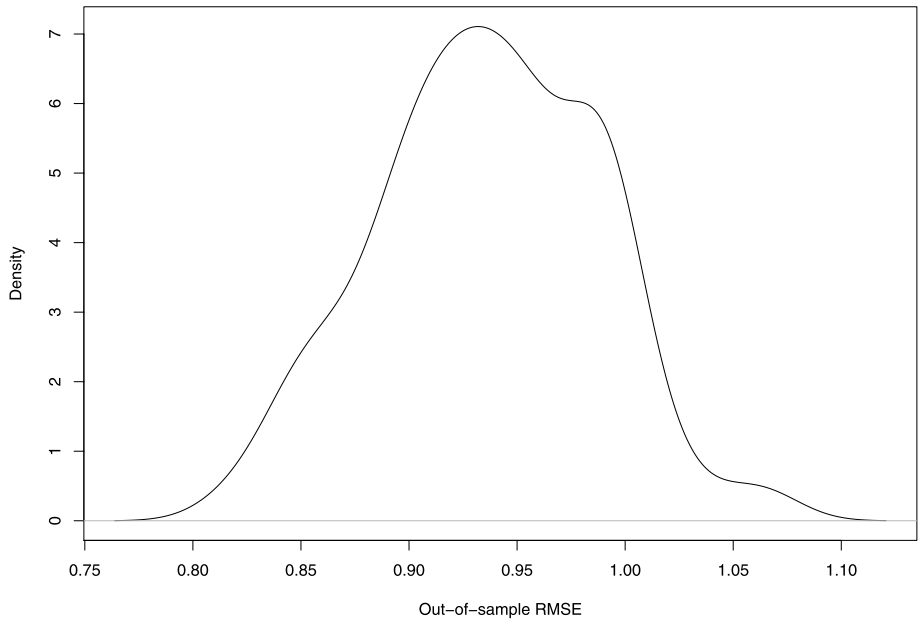

FIGURE 2 Density plot of predictive power for subsamples with $0.434 < R^2 < 0.454$. RMSE, root mean square error

Jointly, these results underline the importance of evaluating a model's predictive power to determine if it generalises well to other datasets and to ascertain that the model estimation is not subject to overfit. Such an analysis requires that benchmarks of predictive performance be generated for similar models and contexts and reported for comparative purposes. In the following section, we will illustrate such an analysis for the Drabe et al. (2015) model and ISSP data.

4.3 | Predictive power across countries and times

To illustrate the benefits of assessing predictive power across contexts, we first use the model estimates from one country (e.g., USA) to predict the observations from another country (e.g., Germany and Japan). Each model is being estimated and assessed using the 2015 ISSP data.

The results in Table 1 (upper part) show that while the Japanese model has the highest R^2 (0.519), it lags behind the German model in terms of predicting the USA data—as indicated by higher RMSE and MAE values. At the same time, both the German and USA models do not predict the Japanese data very well (e.g., RMSE = 0.963 and 1.012 for Germany and USA, respectively). These results suggest that there is some idiosyncrasy in the Japanese data that fits its own model, but which cannot be predicted well by other models. This leads us to surmise this model might be overfit or that there is some substantial predictor of job satisfaction in the Japanese data that is not included in the model. Further research is necessary to find the reasons for the disparity in predictive performance.

Similarly, the German model has the lowest explanatory power ($R^2 = 0.384$) and relatively low predictive power for the Japanese and USA data with RMSE values of 0.963 and 0.960 respectively. Yet, the German data is fairly well predicted by both the Japanese and USA models with RMSE values of 0.816 and 0.851, respectively. Further research is necessary to determine why the model fits the German data relatively poorly, and why the German data is well predicted by the Japanese and USA models.

TABLE 1 Predictive power of country models

		Country data to be predicted		
		Germany	**Japan**	**USA**
Country data used for estimation	Germany	**0.384**	0.963	0.960
			0.753	*0.744*
	Japan	0.816	**0.519**	0.989
		0.624		*0.757*
	USA	0.851	1.012	**0.444**
		0.658	*0.769*	
Drabe et al. (2015)		0.937	0.809	0.977
		0.707	*0.614*	*0.740*

Note: Bolded values on the diagonal represent each model's explanatory power (R^2); in each cell the top value is the RMSE and the value below it (italicised) is the MAE.

Abbreviations: MAE, mean absolute error; RMSE, root mean square error.

In the second step, we apply Drabe et al.'s (2015) estimates generated using the 2005 data to predict the 2015 ISSP data. The lower part of Table 1 shows the results. We find that the model estimated on the 2005 ISSP data has a much lower power in predicting the 2015 data for Germany and USA compared to Japan. Specifically, the analysis produces RMSE values of 0.937 and 0.977 for Germany and USA versus 0.809 for Japan. These results suggest a significant change in the antecedents of job satisfaction between 2005 and 2015 in Germany and USA that the predictive models fail to capture.

We note that our investigation of the predictive power of models across country contexts might not be specifically relevant to the original research questions explored in the Drabe et al. (2015) article. However, we conduct this investigation to illustrate that the practice of using an established model as a basis for follow-up research or deriving managerial recommendations does not guarantee high predictive power, even when the same model achieved sufficient levels of predictive power in the original context (i.e., country). Researchers wishing to employ predictive analysis would need to tailor their objectives more closely to the research project and topic under study and motivate the contexts explicitly.

5 | EXPLANATION AND PREDICTION IN HRM RESEARCH

After having empirically shown the interplay between a model's EP performance and demonstrated the importance of predictive power assessments, we investigate if HRM researchers' approach to model evaluation corresponds to their framing of managerial implications and if and how predictive analytics have been used. To this end, we evaluated all articles published in the five-year period between 2014 and 2018 in three top-tier journals in the field (*Human Resource Management*, *Human Resource Management Journal* and *Personnel Psychology*).

The full article list ($n = 603$) was originally refined by excluding all conceptual and qualitative papers as well as editorials. This initial screening resulted in 432 articles, each of which was then manually examined to ascertain whether the authors assessed their models' explanatory power, predictive power or both.[6] To be precise, if a study considered in-sample metrics or confined its analysis to the assessment of model parameters in terms of their significance, direction and size—potentially preceded by a model fit assessment—the model evaluation was deemed explanatory. If a study used a validation dataset or CV to generate an estimate of a model's predictive power or engaged in predictive model comparisons, the evaluation was deemed predictive. In the final step, we investigated

TABLE 2 Results of the literature review

	Human Resource Management (n_{total} = 311)	*Human Resource Management Journal* (n_{total} = 177)	*Personnel Psychology* (n_{total} = 115)
Number of quantitative papers	224	106	102
Model evaluation			
Only descriptive power[a]	7	3	4
Only explanatory power	217	102	97
Only predictive power	0	0	0
Explanatory and predictive power	0	1[b]	1
Predictive claims in managerial implications section	211	103	98

[a]Descriptive power refers to all analyses aimed at compactly summarising or representing the data structure by using measures of location, association and dependence.

[b]The authors use the AIC for predictive model comparisons, which they label as explanatory.

Abbreviation: AIC, Akaike Information Criterion.

whether the authors used predictive statements when they described their studies' managerial implications. As stated, predictive statements include conditional sentences (if-then) or formulations that foreshadow a specific result if an activity is implemented.

Table 2 shows the results of this review, Table 3 lists a number of illustrative quotes from articles published in the three journals in which authors overstated prescriptive statements in their results discussion and Table 4 lists a number of illustrative quotes from articles published in the three journals where strong prescriptions are justified by the research design. Two major findings emerged from this literature review.

First, practically all of the 432 quantitative studies focused on association-based explanatory power assessments (Table 2); only one study explicitly evaluated the model's predictive power. Specifically, Speer (2018) presents a combination of text mining-based sentiment analysis—configured using a validation sample—and regression analysis to predict the following year's employee performance ratings. In another study, Lauring and Jonasson (2018, p. 400) engage in model comparisons based on the AIC and other metrics to select the model with the highest 'goodness-of-fit'. Given that the AIC tends to select the model with the highest predictive power (Aho et al., 2014), the authors factually engage in predictive model comparisons without identifying their model evaluation's focus as such. Our review shows that this confusion of evaluation foci is common. In only one study, the author clearly differentiates between explanatory and predictive power by stating in a footnote: 'The sign and statistical significance of each regression coefficient are of primary interest here, rather than the magnitude, because our intention is to determine whether a positive relationship exists, as opposed to using the models for prediction' (Bello-Pintado, 2015, p. 326).

From this literature review, it is clear that, barring a few examples, HRM researchers are generally not employing predictive analyses in their research. This is particularly worrying given that predictive and explanatory goals are distinct, though both are essential to scientific research (Dubin, 1969).

Second, practically all the studies (412 out of 432 studies; 95.37%) include prescriptive statements in their managerial implications sections, despite their focus on explanatory power in the model evaluation (Table 2). Two examples illustrate this typical practice—nothing about these examples is unusual. Wagstaff et al. (2015) analyse the relationships between social support seeking, core self-evaluations and withdrawal behaviours by using a series of hierarchical multiple regression analyses. The authors report the *F*-statistics for each of their models to support their

TABLE 3 Illustrative quotes from the literature review where articles have *overstated* practical prescriptions

Reference	Quote
Human Resource Management	
Semeijn et al. (2014, p. 789)	'We would like to recommend that in order to fully benefit from multisource feedback within organisations, more intensive communication and collaborations on the subject are to be advised, as they will positively influence the quality of the ratings'.
Wagstaff et al. (2015, p. 683)	'Managers will benefit from hiring individuals with high core self-evaluations because beyond increasing individuals' sales volume, task performance, rated performance, service quality orientation and service climate (…), these individuals are less likely to withdraw from the organisation when social support seeking is a major response to perceived discrimination'.
Hui et al. (2015, p. 451)	'We also propose that organisations should offer participative decision-making opportunities to their employees, who need to feel that they are part of an in-group or have special status, which in turn will prompt them to reciprocate with more citizenship-like behaviours'.
Glass and Cook (2018, p. 834)	'In order to maximise business and equity policies and practices, firms should be highly attentive to board composition irrespective of CEO gender. When men lead, firms should seek to integrate boards with women directors. Doing so will improve firms' governance, community engagement, and diversity'.
Sanders et al. (2018, p. 1464)	'First, our findings demonstrate while the implementation of performance-based rewards can encourage innovative behaviors, managers will achieve these effects only when these practices are implemented and communicated in a way that is understood by employees (…), and provide clarity and consistency. (…) For this reason, management should consider clarity and consistency when communicating their HR policies to employees'.
Human Resource Management Journal	
Van de Voorde and Beijer (2015, p. 75)	'In particular, the study suggests that line managers can positively stimulate employee commitment in their work unit by increasing the coverage of employees by HPWS in their work unit. By communicating that the intended goal of the HPWS practices reflects care and support for employees, line managers can further enhance employee commitment and reduce feelings of job strain'.
Ng and Sears (2017, p. 144)	'The use of these policies can be expected to enhance a firm's reputation and attract a greater number of women to the firm, thus providing a pipeline of female talent for future promotions'.
Edwards and Kudret (2017, p. 186)	'Where the organisation consistently acts in a fair and socially responsible way towards employees as well as other social and non-social stakeholders across the board, this will elicit increased levels of employee commitment, pride and performance'.
Jung et al. (2018, p. 421)	'In order to minimise negative impacts, our results suggest that HR managers need to have policies in place that convert the employment arrangements of TAWs who meet certain criteria into full-time, permanent work'.
Lo Presti et al. (2018, p. 438)	'In order to successfully manage their career and stay on top of trends, freelancers should engage in career learning and continually develop their employability and positive career attitudes'.
Personnel Psychology	
Grijalva et al. (2015, p. 30)	'Our findings suggest that assuming lower narcissism scores are better is not always accurate. Instead, narcissism levels near the population mean will be

TABLE 3 (Continued)

Reference	Quote
	associated with the most positive leadership outcomes. Thus, individuals with average levels of narcissism should be preferred over those with either very low or very high levels'.
Pieper (2015, p. 849)	'For example, when positions open up, organisations should proactively seek referrals from their high performing employees, who are likely to refer high-quality candidates, who in turn will produce high quality, cost-effective work'.
Kauppila (2016, p. 385–386)	'Rather, this finding implies that the more extensively top managers decentralise responsibilities, the more they need to work to assure that employees in leadership positions have what is needed to establish high-quality relationships with all of their followers. Accordingly, top managers should design decentralised settings in such a way that group leaders do not have too many followers relative to their time and other resources'.
Deng et al. (2017, p. 490)	'To the extent that organisational display rules emphasise the suppression of negative emotions, employees are more likely to utilise surface acting; if display rules emphasise positive expressions, however, employees 'focus more on trying to experience a positive emotional state' (…). Hence, service organisations should place more emphasis on positive rather than negative emotion norms so as to promote their employees' productive emotion regulation and, in doing so, alleviate employees' ego depletion and reduce coworker harming'.
Seibert et al. (2017 p. 387)	'The other informal developmental experience examined here, developmental supervision, positively related to the manager's network of supportive relationships. Thus, organisations should also train their supervisors to engage in role modelling (e.g., setting a positive example with their own leader behaviour), and in coaching behaviour (e.g., challenging thinking and assumptions, driving results and creating accountability for goals)'.

models' significance and R^2 values, followed by the beta weights' interpretation. In their discussion of the practical implications, the authors conclude that 'managers *will* benefit from hiring individuals with high core self-evaluations because beyond increasing individuals' sales volume, task performance, rated performance, service quality orientation, and service climate (…), these individuals are less likely to withdraw from the organisation when social support seeking is a major response to perceived discrimination' (Wagstaff et al. 2015, p. 683; emphasis added). Specifically, although the analysis draws on in-sample metrics, the authors conclude that their results apply in a predictive scenario, which is, however, not necessarily the case—as shown in our empirical illustration. Similarly, in their analysis of how different facets of corporate social responsibility impact commitment, organisational pride and employee performance by using structural equation modelling, Edwards and Kudret (2017, p. 181) note that the results of various fit indices, such as CFI and SRMR, suggest that the 'model fitted the data well' and continue with the analysis of the path coefficients. The authors then conclude that 'where the organisation consistently acts in a fair and socially responsible way towards employees as well as other social and non-social stakeholders across the board, this *will* elicit increased levels of employee commitment, pride and performance' (Edwards & Kudret, 2017, p. 186; emphasis added). Such a conclusion, however, should be backed up by a predictive power assessment.

In summary, these studies provide valuable insights into the underlying research topics and adequately test the assumed relationships from an explanatory perspective. However, the next step, generating practical prescriptions, requires additional predictive power assessments, ideally using data from a different context.

The remaining 20 of the 432 quantitative studies (4.63%) that avoid making predictive claims in the managerial implications sections tend to summarise the results in the context of the study setting (e.g., Swart et al., 2014; Teague & Roche, 2014). Specifically, the authors do not foreshadow outcomes if certain activities are implemented,

TABLE 4 Illustrative quotes from the literature review where articles have *correctly* stated practical prescriptions

Reference	Quote
Human Resource Management	
Baum and Kabst (2014, p. 368)	'First, our results indicate that high-information-recruitment practices such as websites may be utilised to develop an employer brand'.
Firfiray and Mayo (2017, p. 643)	'HR departments should be attentive to the inclusion of employment inducements such as WLBs in their recruitment materials as they can positively influence applicant perceptions of recruiting organizations'.
Deery et al. (2017, p. 1047)	'Our study revealed in particular the difficulty of performing multiple roles at a high level. Combining high levels of conscientiousness with high levels of task performance was associated with clearly identifiable negative outcomes'.
Human Resource Management Journal	
Swart et al. (2014, p. 283)	'We have shown that client commitment, especially client continuance commitment, has a negative relationship with knowledge sharing, whereas the influence of commitment to the other foci is positive. Indeed, it seems that continuance commitment to the client is quite unlike commitment to the other foci'.
Teague & Roche (2014, p. 189)	'Second, our findings suggest that it is necessary to recognise theoretically that a variety of adjustment routes are available to firms facing deep recession, involving different bundles of HR practices. Firms commonly bundle HR practices in retrenchment programmes but not, it seems, often in the ways associated with employment stabilisation, responsible restructuring or pure restructuring'.
Personnel Psychology	
Chao et al. (2017, p. 280)	'Guided by the contact hypothesis, our study demonstrated that the quality of international contact experience matters for CQ development. In particular, adjustment in the social domain was shown to play a critical role in fostering motivational and behavioural CQs. This suggests that the provision of support that enhances social adjustment could enable sojourners to garner more CQ benefit from their international experience'.
Slaughter et al. (2014, p. 877)	'In terms of practical implications, one could cautiously interpret our findings as suggesting that researchers should perhaps measure AH and use it in combination with ISJT scores when selecting employees'.

which corresponds to the studies' explanatory focus of their model evaluations. For example, Deery et al. (2017, p. 1047) perform a confirmatory factor analysis and regression analysis of survey data and do well to limit the prescription or generalisability of the claims that can be generated from their results (Table 4):

> "Our study revealed in particular the difficulty of performing multiple roles at a high level. Combining high levels of conscientiousness with high levels of task performance was associated with clearly identifiable negative outcomes".

6 | DISCUSSION

Predictive model evaluation practices and their implications for deriving meaningful managerial recommendations have been largely overlooked in HRM research. Assessing a model's predictive power is at the heart of a scientific enterprise. Researchers evaluate and compare theories based on their ability to make falsifiable predictions about

TABLE 5 Software packages and reference materials for conducting predictive analysis in *R*

Purpose	Software	Details	Reference
Cross-sectional data			
Cross-validation and generation of predictive metrics	caret package (Kuhn, 2020)	The caret package provides a wide range of tools for training predictive models on cross-sectional data. It incorporates automated slicing of data into training and holdout datasets, and multiple cross-validation techniques such as *k*-fold and LOOCV. caret works on a wide range of models such as multiple regression, logit regression, classification tasks such as KNN, and regression trees.	Kuhn (2008)
Time-series data			
Cross-validation and generation of predictive metrics	forecast package (Hyndman et al., 2021)	The forecast package provides a wide range of tools for training predictive models on time-series data. It incorporates automated slicing of data into training and holdout datasets and LOOCV. Forecast works on a wide range of time-series models such as multiple regression, ARIMA, and ETS.	Hyndman and Khandakar (2008)
Survey and observational data			
Cross-validation and generation of prediction metrics for structural equation models	seminr package (Ray et al., 2021)	The seminr package provides a wide range of tools for structural equation modelling, including cross-validation and different prediction metrics.	Hair et al., (2022)

Abbreviations: ARIMA, autoregressive integrated moving average; ETS, error trend seasonal; KNN, K-nearest neighbours; LOOCV, leave-one-out cross-validation.

new observations. 'Historically, this process of prediction-driven explanation has proven uncontroversial in the physical sciences, especially in cases where theories make relatively unambiguous predictions and data are plentiful' (Hofman et al., 2017, p. 486). Despite prediction's central role in science, HRM researchers generally deemphasise its importance in their model evaluations relative to explanatory power. That is, they focus on the form of the input-output relationship instead of testing whether their models accurately predict new output data *given* the input. While both explanation and prediction have a raison d'être (Hofman et al., 2017), a singular focus on explanation is not enough when researchers seek to derive managerial implications, which are inherently predictive by nature. This is because models with high explanatory power do not inherently possess predictive power—as empirically illustrated in our replication of Drabe et al.'s (2015) model on job satisfaction in ageing workforces. Specifically, we show that a model with a certain degree of explanatory power can produce vastly different levels of predictive power and vice versa. In addition, our results highlight that a model developed in one context, such as time and place of data collection, does not necessarily generalise well to other contexts in terms of predictive power.

There is much to gain for HRM research by putting more focus on predictive power assessments. Long ago, Kaplan (1964, p. 350) noted that 'it remains true that if we can predict successfully on the basis of a certain explanation, we have a good reason, and perhaps the best sort of reason, for accepting the explanation'. That is, the actual results of

decision-making are generally a critical test of explanations' relevance. Furthermore, a stronger prediction focus widens the field's potential to foster an understanding of behavioural phenomena. For example, a change in focus can help researchers develop new theories or improve existing models. It can also guide the comparison of competing models derived from different theories by selecting a model that generalises well to other samples, while avoiding models that overfit the data by tapping spurious sample-specific patterns (Shmueli & Koppius, 2011). Finally, the routine reporting of predictive performance metrics would facilitate follow-up meta-analyses that quantify the expected predictive performance of seminal models used in HRM across different contexts.

In consideration of the above, HRM researchers should place greater emphasis on validating their models' predictive power. This requires researchers to engage in an out-of-sample prediction assessment by using CV or replication samples. Such analyses have become the norm in many fields, such as sensory product research, and research offers clear guidance on how to implement them (e.g., Rodriguez et al., 2010). Specifically, studies that draw on large-scale empirical datasets, for example, from online social networking sites, can readily implement predictive validation techniques. Consumer researchers should also plan for predictive assessments in their research designs, for example, by including holdout tasks in choice experiments. In Table 5, we offer a list of popular software packages for the R Statistical Environment (R Core Team, 2021) for applying predictive techniques in empirical HRM papers along with references, which researchers may consult in order to investigate the domain of predictive analytics in more depth.

Finally, researchers should place less emphasis on methods that follow an explanation-only paradigm and should rather consider methods that bridge the apparent dichotomy between explanation and prediction. For example, partial least squares, which has recently gained prominence in a variety of fields including HRM (Ringle et al., 2020), was designed as a causal-predictive approach to structural equation modelling that emphasises prediction in the estimation of a model having a structure that is anchored in causal explanations (Hair et al., 2019). Similarly, researchers can employ the more recently developed generalised structured component analysis (Hwang & Takane, 2004) approach, which has attracted considerable research attention from users and methodologists alike (Hwang et al., 2020). Both approaches to structural equation modelling allow researchers to readily assess their models' predictive power by using CV-type methods (Cho et al., 2019; Shmueli et al., 2016). In addition, recent research has underlined their efficacy for prediction-oriented model selection. For example, Sharma et al. (2021) have recently shown that model selection criteria such as the Bayesian Information Criterion (Schwarz, 1978) and Geweke and Meese's (1981) criterion can reliably substitute out-of-sample criteria that require a holdout sample. Researchers can also draw on model selection criteria to generate averaged predictions across multiple models. These model-averaged predictions tend to perform well in terms of EP power, thereby improving confidence in any prescriptions derived from the models (Danks et al., 2020).

When applying these methods, however, researchers need to acknowledge that there is an inherent tension between models that perform well in terms of explanation versus those with a high predictive power (Hair & Sarstedt, 2021). Researchers should therefore first establish their research's main goal—usually explanation—and evaluate the model's performance not only in terms of this goal (i.e., explanatory power) but also in terms of the other goal (i.e., predictive power), thus applying an EP lens. This analysis might not produce the best predictive model, but researchers could define a minimal threshold regarding how well the explanatory model should predict. Researchers could also compare models that are equally strong in terms of explanation, although one model will have a higher predictive power. Implementing such procedures is a fundamental step towards increasing HRM research's rigour and relevance.

Future researchers should look into the role of Bayesian statistical approaches for integrating causal thinking with predictive modelling in HRM research. Drawing on techniques such as Bayesian predictive model selection could offer more confidence when drawing causal conclusions and balancing in-sample and out-of-sample fit (Piironen & Vehtari, 2017; Vehtari & Ojanen, 2012). Finally, while our elaborations focused on the EP approach (Gregor, 2006) and the role of predictive power assessments, future research should critically discuss the conditions for causal inference in HRM studies (e.g., Holland, 1986, Pearl, 2009).

ACKNOWLEDGEMENT

Open Access funding enabled and organized by Projekt DEAL.

[Correction added on 26th July: Acknowledgement section is updated in this version.]

CONFLICT OF INTEREST

The authors declare no conflict of interest.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from GESIS– Leibniz-Institute for the Social Sciences at https://www.gesis.org/en/issp/modules/issp-modules-by-topic/work-orientations/2015, doi:10.4232/1.12848. The analysis code can be downloaded from GitHub at https://github.com/NicholasDanks/Prediction-in-HRM-research.

ORCID

Marko Sarstedt https://orcid.org/0000-0002-5424-4268

Nicholas P. Danks https://orcid.org/0000-0001-6902-2708

ENDNOTE

[1] See Appendix A for an overview of country-specific model estimates.

[2] We would like to thank the anonymous reviewer for making this suggestion.

[3] The sample size yields a high level of statistical power for the estimation of the linear model (GPower, $f^2 = 0.1$, $\alpha = 0.05$, $1 - \beta = 0.95$, $n = 10$; Erdfelder et al., 1996). In Appendix C, we replicate the analysis using larger sample sizes of 350 and 450 in order to demonstrate that the findings are robust for higher sample sizes.

[4] The code can also be downloaded from GitHub at https://github.com/NicholasDanks/Prediction-in-HRM-research.

[5] For the sake of clarity and brevity, we focus our evaluation on the R^2 and out-of-sample RMSE metrics to conduct this evaluation. The results and conclusions when comparing in-sample RMSE and out-of-sample RMSE are highly similar.

[6] We also identified a number of cases where the authors confined their analyses to an assessment of descriptive power by using measures of location, association and dependence (see Table 2).

[7] The code is included in an Appendix, the code can also be downloaded from GitHub at: https://github.com/NicholasDanks/Prediction-in-HRM-research

REFERENCES

Aho, K., Derryberry, D., & Peterson, T. (2014). Model selection for ecologists: The worldviews of AIC and BIC. *Ecology*, *95*(3), 631–636. https://doi.org/10.1890/13-1452.1

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. N. Petrov, & F. Csáki (Eds.), *Selected papers of Hirotugu Akaike* (pp. 199–213). Springer. https://doi.org/10.1007/978-1-4612-1694-0_15

Arlot, S., & Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Statistics Surveys*, *4*, 40–79. https://doi.org/10.1214/09-SS054

Bagozzi, P. R., & Yi, Y. (2012). Specification, evaluation, and interpretation of structural equation models. *Journal of the Academy of Marketing Science*, *40*(1), 8–34. https://doi.org/10.1007/s11747-011-0278-x

Baum, M., & Kabst, R. (2014). The effectiveness of recruitment advertisements and recruitment websites: Indirect and interactive effects on applicant attraction. *Human Resource Management*, *53*(3), 353–378. https://doi.org/10.1002/hrm.21571

Bello-Pintado, A. (2015). Bundles of HRM practices and performance: Empirical evidence from a Latin American context. *Human Resource Management Journal*, *25*(3), 311–330. https://doi.org/10.1111/1748-8583.12067

Bergmeir, C., & Benítez, J. M. (2012). On the use of cross-validation for time series predictor evaluation. *Information Sciences*, *191*, 192–213. https://doi.org/10.1016/j.ins.2011.12.028

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, *71*(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Box, G. E., Hunter, W. H., & Hunter, S. (1978). *Statistics for experimenters (664)*. John Wiley and sons.

Burman, P. (1989). A comparative study of ordinary cross-validation, V-fold cross-validation and the repeated learning-testing methods. *Biometrika*, *76*(3), 503–514. https://doi.org/10.1093/biomet/76.3.503

Burnham, K. P., & Anderson, D. R. (2002). *Model selection and multimodel inference: A practical information-theoretic approach*. Springer.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)?–Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247–1250.

Chao, M. M., Takeuchi, R., & Farh, J.-L. (2017). Enhancing cultural intelligence: The roles of implicit culture beliefs and adjustment. *Personnel Psychology*, *70*(1), 257–292. https://doi.org/10.1111/peps.12142

Cho, G., Jung, K., & Hwang, H. (2019). Out-of-bag prediction error: A cross validation index for generalized structured component analysis. *Multivariate Behavioral Research*, *54*(4), 505–513. https://doi.org/10.1080/00273171.2018.1540340

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Lawrence Erlbaum Associates.

Danks, N., Sharma, P. N., & Sarstedt, M. (2020). Model selection uncertainty and multimodel averaging in partial least squares structural equation modelling (PLS-SEM). *Journal of Business Research*, *113*, 13–24. https://doi.org/10.1016/j.jbusres.2020.03.019

Deery, S., Rayton, B., Walsh, J., & Kinnie, N. (2017). The costs of exhibiting organizational citizenship behavior. *Human Resource Management*, *56*(6), 1039–1049. https://doi.org/10.1002/hrm.21815

Deng, H., Walter, F., Lam, C. K., & Zhao, H. H. (2017). Spillover effects of emotional labor in customer service encounters toward coworker harming: A resource depletion perspective. *Personnel Psychology*, *70*(2), 469–502. https://doi.org/10.1111/peps.12156

Dowe, D. L., Gardner, S., & Oppy, G. R. (2007). Bayes not bust! Why simplicity is no problem for Bayesians. *The British Journal for the Philosophy of Science*, *58*(4), 709–754. https://doi.org/10.1093/bjps/axm033

Drabe, D., Hauff, S., & Richter, N. F. (2015). Job satisfaction in aging workforces: An analysis of the USA, Japan and Germany. *International Journal of Human Resource Management*, *26*(6), 783–805. https://doi.org/10.1080/09585192.2014.939101

Dubin, R. (1969). *Theory building*. The Free Press.

Edwards, M. R., & Kudret, S. (2017). Multi-foci CSR perceptions, procedural justice and in-role employee performance: The mediating role of commitment and pride. *Human Resource Management Journal*, *27*(1), 169–188. https://doi.org/10.1111/1748-8583.12140

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, & Computers*, *28*(1), 1–11. https://doi.org/10.3758/BF03203630

Firfiray, S., & Mayo, M. (2017). The lure of work-life benefits: Perceived person-organization fit as a mechanism explaining job seeker attraction to organizations. *Human Resource Management*, *56*(4), 629–649. https://doi.org/10.1002/hrm.21790

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*(3), S124–S134. https://doi.org/10.1086/341840

Forster, M. R., & Sober, E. (1994). How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions. *The British Journal for the Philosophy of Science*, *45*(1), 1–35. https://doi.org/10.1093/bjps/45.1.1

Franklin, J. (2005). The elements of statistical learning: Data mining, inference and prediction. *The Mathematical Intelligencer*, *27*(2), 83–85. https://doi.org/10.1007/BF02985802

Geweke, J., & Meese, R. (1981). Estimating regression models of finite but unknown order. *International Economic Review*, *22*(1), 55–70. https://doi.org/10.1016/0304-4076(81)90091-9

Glass, C., & Cook, A. (2018). Do women leaders promote positive change? Analyzing the effect of gender on business practices and diversity initiatives. *Human Resource Management*, *57*(4), 823–837. https://doi.org/10.1002/hrm.21838

Gregor, S. (2006). The nature of theory in information systems. *MIS Quarterly*, *30*(3), 611–642. https://doi.org/10.2307/25148742

Grijalva, E., Harms, P. D., Newman, D. A., Gaddis, B. H., & Fraley, R. C. (2015). Narcissism and leadership: A meta-analytic review of linear and nonlinear relationships. *Personnel Psychology*, *68*(1), 1–47. https://doi.org/10.1111/peps.12072

Hair, J. F., Hult, T. M., Ringle, C. M., Sarstedt, M., Danks, N. P., & Ray, S. (2022). *Partial least squares structural equation modeling (PLS-SEM) using R–A workbook*. Springer Nature.

Hair, J. F., & Sarstedt, M. (2021). Explanation plus prediction—The logical focus of project management research. *Project Management Journal*, *52*(4), 319–322. https://doi.org/10.1177/8756972821999945

Hair, J. F., Sarstedt, M., & Ringle, C. M. (2019). Rethinking some of the rethinking of partial least squares. *European Journal of Marketing*, *53*(4), 566–584. https://doi.org/10.1108/EJM-10-2018-0665

Hastie, T., Tibshirani, R., & Friedman, J. H. (2013). *The elements of statistical learning: Data mining, inference, and prediction* (2nd ed.2009. Corr. 10th printing 2013 edition). Springer.

Hofman, J. A. M., Sharma, A., & Watts, D. J. (2017). Prediction and explanation in social systems. *Science*, *355*(6324), 486–488. https://doi.org/10.1126/science.aal3856

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, *81*(396), 945–960.

Hui, C., Lee, C., & Wang, H. (2015). Organizational inducements and employee citizenship behavior: The mediating role of perceived insider status and the moderating role of collectivism. *Human Resource Management*, *54*(3), 439–456. https://doi.org/10.1002/hrm.21620

Hwang, H., Sarstedt, M., Cheah, J.-H., & Ringle, C. M. (2020). A concept analysis of methodological research on composite-based structural equation modeling: Bridging PLSPM and GSCA. *Behaviormetrika*, *47*(1), 219–241. https://doi.org/10.1007/s41237-019-00085-5

Hwang, H., & Takane, Y. (2004). Generalized structured component analysis. *Psychometrika*, *69*(1), 81–99. https://doi.org/10.1007/BF02295841

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., Yasmeen, F. (2021). forecast: Forecasting functions for time series and linear models. R package version 8.14. https://pkg.robjhyndman.com/forecast/

Hyndman, R. J. (2006). Another look at forecast-accuracy metrics for intermittent demand. *Foresight: The International Journal of Applied Forecasting*, *4*(4), 43–46.

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: The forecast package for R. *Journal of Statistical Software*, *27*(3), 1–22. https://doi.org/10.18637/jss.v000.i00

ISSP Research Group. (2017). ZA6770: International social survey programme: Work orientations IV–ISSP 2015. https://doi.org/10.4232/1.12848

Japkowicz, N., & Shah, M. (2011). *Evaluating learning algorithms: A classification perspective*. Cambridge University Press. https://doi.org/10.1017/CBO9780511921803

Jung, H.-J., Noh, S.-C., & Kim, I. (2018). Relative deprivation of temporary agency workers in the public sector: The role of public service motivation and the possibility of standard employment. *Human Resource Management Journal*, *28*(3), 410–426. https://doi.org/10.1111/1748-8583.12186

Kaplan, A. (1964). *The conduct of inquiry: Methodology for behavioral science*. Chandler Publishing.

Kauppila, S.-P. (2016). When and how does LMX differentiation influence followers' work outcomes? The interactive roles of one's own LMX status and organizational content. *Personnel Psychology*, *69*(2), 357–393. https://doi.org/10.1111/peps.12110

Konishi, S., & Kitagawa, G. (2007). *Information criteria and statistical modeling*. Springer. https://doi.org/10.1007/978-0-387-71887-3

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

Kuhn, M. (2020). caret: Classification and regression training. R package version 6.0-86. https://CRAN.R-project.org/package=caret

Lauenroth, W. K. (2003). *Models in ecosystem science*. Princeton University Press.

Lauring, J., & Jonasson, C. (2018). Can leadership compensate for deficient inclusiveness in global virtual teams? *Human Resource Management Journal*, *28*(3), 392–409. https://doi.org/10.1111/1748-8583.12184

Lo Presti, A., Pluviano, S., & Briscoe, J. P. (2018). Are freelancers a breed apart? The role of protean and boundaryless career attitudes in employability and career success. *Human Resource Management Journal*, *28*(3), 427–442. https://doi.org/10.1111/1748-8583.12188

Makridakis, S., Hogarth, R. M., & Gaba, A. (2009). Forecasting and uncertainty in the economic and business world. *International Journal of Forecasting*, *25*(4), 794–812. https://doi.org/10.1016/j.ijforecast.2009.05.012

McQuarrie, A. D., & Tsai, C. L. (1998). *Regression and time series model selection* (Vol. 43). World Scientific. https://doi.org/10.1142/3573

Ng, E. S., & Sears, G. J. (2017). The glass ceiling in context: The influence of CEO gender, recruitment practices and firm internationalisation on the representation of women in management. *Human Resource Management Journal*, *27*(1), 133–151. https://doi.org/10.1111/1748-8583.12135

Pearl, J. (2009). *Causality: Models, reasoning, and inference*. Cambridge University Press. https://doi.org/10.1017/CBO9780511803161

Pieper, J. R. (2015). Uncovering the nuances of referral hiring: How referrer characteristics affect referral hires' performance and likelihood of voluntary turnover. *Personnel Psychology*, *68*(4), 811–858. https://doi.org/10.1111/peps.12097

Piironen, J., & Vehtari, A. (2017). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Ray, S., Danks, N. P., & Valdez, A. C. (2021). seminr: Domain-specific language for building and estimating structural equation models. R package version 2.0.0. https://cran.r-project.org/web/packages/seminr/index.html

R Core Team. (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing. https://www.R-project.org/ Accessed 3 March 2021.

Rigdon, E. E. (2012). Rethinking partial least squares path modeling: In praise of simple methods. *Long Range Planning*, *45*(5/6), 341–358. https://doi.org/10.1016/j.lrp.2012.09.010

Ringle, C. M., Sarstedt, M., Mitchell, R., & Gudergan, S. P. (2020). Partial least squares structural equation modeling in HRM research. *International Journal of Human Resource Management*, *31*(12), 1617–1643. https://doi.org/10.1080/09585192.2017.1416655

Rodriguez, J. D., Perez, A., & Lozano, J. A. (2010). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *32*(3), 569–575. https://doi.org/10.1109/TPAMI.2009.187

Sanders, K., Jorgensen, F., Shipton, H., van Rossenberg, Y., Cunha, R., Li, X., Rodrigues, R., Wong, S. I., & Dysvik, A. (2018). Performance-based rewards and innovative behaviors. *Human Resource Management*, *57*(6), 1455–1468. https://doi.org/10.1002/hrm.21918

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*(2), 461–464. https://doi.org/10.1214/aos/1176344136

Seibert, S. E., Sargent, L. D., Kraimer, M. L., & Kiazad, K. (2017). Linking developmental experiences to leader effectiveness and promotability: The mediating role of leadership self-efficacy and mentor network. *Personnel Psychology*, *70*(2), 357–397. https://doi.org/10.1111/peps.12145

Semeijn, J. H., van der Heijden, B. I. J. M., & van der Lee, A. (2014). Multisource ratings of managerial competencies and their predictive value for managerial and organizational effectiveness. *Human Resource Management*, *53*(5), 773–794. https://doi.org/10.1002/hrm.21592

Sharma, P., Sarstedt, M., Shmueli, G., Kim, K. H., & Thiele, K. O. (2019). PLS-based model selection: The role of alternative explanations in information systems research. *Journal of the Association for Information Systems*, *20*(4), 346–397.

Sharma, P. N., Shmueli, G., Sarstedt, M., Danks, N., & Ray, S. (2021). Prediction-oriented model selection in partial least squares path modelling. *Decision Sciences*, *52*(3), 567–607. https://doi.org/10.1111/deci.12329

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310. https://doi.org/10.1214/10-STS330

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572. https://doi.org/10.2307/23042796

Shmueli, G., Ray, S., Velasquez Estrada, J. M., & Chatla, S. B. (2016). The elephant in the room: Evaluating the predictive performance of PLS models. *Journal of Business Research*, *69*(10), 4552–4564. https://doi.org/10.1016/j.jbusres.2016.03.049

Slaughter, J. E., Christian, M. S., Podsakoff, N. P., Sinar, E. F., & Lievens, F. (2014). On the limitations of using situational judgment tests to measure interpersonal skills: The moderating influence of employee anger. *Personnel Psychology*, *67*(4), 847–885. https://doi.org/10.1111/peps.12056

Speer, A. B. (2018). Quantifying with words: An investigation of the validity of narrative-derived performance scores. *Personnel Psychology*, *71*(3), 299–333. https://doi.org/10.1111/peps.12263

Sugiura, N. (1978). Further analysis of the data by Akaike's information criterion and the finite corrections. *Communications in Statistics– Theory and Methods*, *7*(1), 13–26. https://doi.org/10.1080/03610927808827599

Swart, J., Kinnie, N., van Rossenberg, Y., & Yalabik, Z. Y. (2014). Why should I share my knowledge? A multiple foci of commitment perspective. *Human Resource Management Journal*, *24*(3), 269–289. https://doi.org/10.1111/1748-8583.12037

Teague, P., & Roche, W. K. (2014). Recession bundles: HR practices in the Irish economic crisis. *Human Resource Management Journal*, *24*(2), 176–192. https://doi.org/10.1111/1748-8583.12019

Van de Voorde, K., & Beijer, S. (2015). The role of employee HR attributions in the relationship between high-performance work systems and employee outcomes. *Human Resource Management Journal*, *25*(1), 62–78. https://doi.org/10.1111/1748-8583.12062

Vehtari, A., & Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, *6*, 142–228. https://doi.org/10.1214/12-ss102

Wagstaff, M., del Carmen Triana, M., Kim, S., & Al-Riyami, S. (2015). Social support seeking, core self-evaluations, and withdrawal behaviors. *Human Resource Management*, *54*(4), 673–687. https://doi.org/10.1002/hrm.21634

How to cite this article: Sarstedt, M., & Danks, N. P. (2022). Prediction in HRM research–A gap between rhetoric and reality. *Human Resource Management Journal*, *32*(2), 485–513. https://doi.org/10.1111/1748-8583.12400

Appendix A

TABLE A1 New estimation of country specific models and predictive metrics

Coefficients	All	USA	Japan	Germany
Gender (reference: female)	−0.002	0.027	−0.010	−0.019
Age (reference: ≤35)				
36–49	0.029	0.138*	−0.027	−0.019
≥50	0.122***	0.215***	0.173*	0.011
Education	−0.053***	−0.117***	0.012	−0.069*
Income	0.094***	0.116***	0.119***	0.098**
Advancement opportunities	0.041*	0.065*	−0.033	0.013
Job security	0.093***	0.112***	0.092**	0.075*
Interesting job	0.385***	0.406***	0.351***	0.362***
Independent work	0.037*	−0.008	0.038	0.039
Good relationships with management	0.259***	0.249***	0.296***	0.257***
Good relationships with colleagues	0.102***	0.059	0.199***	0.094**
R^2	0.497	0.444	0.519	0.384
Adj. R^2	0.495	0.437	0.511	0.377
n	2449	915	635	899
CV R^2	0.491	0.435	0.500	0.374
CV RMSE	0.891	0.912	0.949	0.786
CV MAE	0.681	0.687	0.723	0.600

Note: Standardised regression coefficients; CV is 10-fold cross validation.

Abbreviations: CV, cross-validation; MAE, mean absolute error; RMSE, root mean square error.

***$p \leq 0.001$; **$p \leq 0.01$; *$p \leq 0.05$.

Appendix B

Code for replicating the thought experiment and empirical study[7]

```
library(caret)
library(MASS)
data <- read.csv(file = "2015 ISSP data.csv")

# Reproducing Drabe et al, USA
usa_data <- na.omit(data[data$country ==36, ])
jp_data <- na.omit(data[data$country ==19, ])
de_data <- na.omit(data[data$country ==14, ])

# Total data
all_data <- rbind(usa_data, jp_data, de_data)

# Estimate USA model
lm_usa <- lm(scale(job_sat) ~ factor(sex) + factor(Age_cat) + scale(educyrs) +
scale(income) + scale(advancement) + scale(security) + scale(interesting) +
scale(independent) + scale(rel_mgmt) + scale(rel_clgs), data = usa_data )
summary(lm_usa)

# Estimate Japan model
lm_jp <- lm(scale(job_sat) ~ factor(sex) + factor(Age_cat) + scale(educyrs) +
scale(income) + scale(advancement) + scale(security) + scale(interesting) +
scale(independent) + scale(rel_mgmt) + scale(rel_clgs), data = jp_data)
summary(lm_jp)

# Estimate German model
lm_de <- lm(scale(job_sat) ~ factor(sex) + factor(Age_cat) + scale(educyrs) +
scale(income) + scale(advancement) + scale(security) + scale(interesting) +
scale(independent) + scale(rel_mgmt) + scale(rel_clgs), data = de_data)
summary(lm_de)

# Estimate All model
lm_all <- lm(scale(job_sat) ~ factor(sex) + factor(Age_cat) + scale(educyrs) +
scale(income) + scale(advancement) + scale(security) + scale(interesting) +
scale(independent) + scale(rel_mgmt) + scale(rel_clgs), data = all_data)
summary(lm_all)

# CV predictive metrics for each country ----
# Obtain CV predictive metrics for the model
data_ctrl <- trainControl(method = "cv", number = 10)
usa_model_caret <- train(job_sat ~ factor(sex) + factor(Age_cat) + educyrs +
income + advancement + security + interesting + independent + rel_mgmt +
rel_clgs,
                         data = usa_data,
                         trControl = data_ctrl,    # folds
                         method = "lm",            # specifying regression model
                         na.action = na.pass)
```

```
# Inspect the USA model
summary(usa_model_caret)

# Output the cross-validated predictive metrics
usa_model_caret$results

# show the results from each cross-validation sub-sample
usa_model_caret$resample

# Obtain CV predictive metrics for the german model
data_ctrl_de <- trainControl(method = "cv", number = 10)
model_caret_de <- train(job_sat ~ factor(sex) + factor(Age_cat) + educyrs +
income + advancement + security + interesting + independent + rel_mgmt +
rel_clgs,
                        data = de_data,
                        trControl = data_ctrl,  # folds
                        method = "lm",          # specifying regression model
                        na.action = na.pass)

# Inspect the German model
summary(model_caret_de)

# Output the cross-validated predictive metrics
model_caret_de$results

# show the results from each cross-validation sub-sample
model_caret_de$resample

# Obtain CV predictive metrics for the japanese model
data_ctrl_jp <- trainControl(method = "cv", number = 10)
model_caret_jp <- train(job_sat ~ factor(sex) + factor(Age_cat) + educyrs +
income + advancement + security + interesting + independent + rel_mgmt +
rel_clgs,
                        data = jp_data,
                        trControl = data_ctrl_jp, # folds
                        method = "lm",            # specifying regression model
                        na.action = na.pass)

# Inspect the Japanese model
summary(model_caret_jp)

# Output the cross-validated predictive metrics
model_caret_jp$results

# show the results from each cross-validation sub-sample
model_caret_jp$resample


# Obtain CV predictive metrics for the all model
data_ctrl_all <- trainControl(method = "cv", number = 10)
model_caret_all <- train(job_sat ~ factor(sex) + factor(Age_cat) + educyrs +
income + advancement + security + interesting + independent + rel_mgmt +
rel_clgs,
                         data = all_data,
                         trControl = data_ctrl_all, # folds
                         method = "lm",            # specifying regression model
                         na.action = na.pass)

# Inspect the All model
```

```
summary(model_caret_all)

# Output the cross-validated predictive metrics
model_caret_all$results

# show the results from each cross-validation sub-sample
model_caret_all$resample

# Countries predicting each other ----

# USA Predicting Japan
prediction <- predict(lm_usa, newdata = jp_data)
unscaled_predictions <- (prediction * attr(lm_usa$fitted.values,
"scaled:scale"))+attr(lm_usa$fitted.values, "scaled:center")
USA_japan_RMSE <- sqrt(mean((jp_data[, "job_sat"] - unscaled_predictions)^2))
USA_japan_MAE <- mean(abs(jp_data[, "job_sat"] - unscaled_predictions))

# USA Predicted Germany
prediction <- predict(lm_usa, newdata = de_data)
unscaled_predictions <- (prediction * attr(lm_usa$fitted.values,
"scaled:scale"))+attr(lm_usa$fitted.values, "scaled:center")
USA_germany_RMSE <- sqrt(mean((de_data[, "job_sat"] - unscaled_predictions)^2))
USA_germany_MAE<- mean(abs(de_data[, "job_sat"] - unscaled_predictions))

# Germany Predicting Japan
prediction <- predict(lm_de, newdata = jp_data)
unscaled_predictions <- (prediction * attr(lm_de$fitted.values,
"scaled:scale"))+attr(lm_de$fitted.values, "scaled:center")
DE_japan_RMSE <- sqrt(mean((jp_data[, "job_sat"] - unscaled_predictions)^2))
DE_japan_MAE <- mean(abs(jp_data[, "job_sat"] - unscaled_predictions))

# Germany Predicting USA
prediction <- predict(lm_de, newdata = usa_data)
unscaled_predictions <- (prediction * attr(lm_de$fitted.values,
"scaled:scale"))+attr(lm_de$fitted.values, "scaled:center")
DE_USA_RMSE <- sqrt(mean((usa_data[, "job_sat"] - unscaled_predictions)^2))
DE_USA_MAE<- mean(abs(usa_data[, "job_sat"] - unscaled_predictions))

# Japan Predicting Germany
prediction <- predict(lm_jp, newdata = de_data)
unscaled_predictions <- (prediction * attr(lm_jp$fitted.values,
"scaled:scale"))+attr(lm_jp$fitted.values, "scaled:center")
JP_DE_RMSE <- sqrt(mean((de_data[, "job_sat"] - unscaled_predictions)^2))
JP_DE_MAE <- mean(abs(de_data[, "job_sat"] - unscaled_predictions))

# Japan Predicting USA
prediction <- predict(lm_jp, newdata = usa_data)
unscaled_predictions <- (prediction * attr(lm_jp$fitted.values,
"scaled:scale"))+attr(lm_jp$fitted.values, "scaled:center")
JP_USA_RMSE <- sqrt(mean((usa_data[, "job_sat"] - unscaled_predictions)^2))
JP_USA_MAE<- mean(abs(usa_data[, "job_sat"] - unscaled_predictions))


# What would happen if our researcher only had a sample of 250?
results <- matrix(0, nrow = 1000, ncol = 3)
for (i in 1:1000) {
  set.seed(123+i)
  index <- sample(rownames(usa_data), size = 500, replace = FALSE,)
  train <- index[1:((length(index)*0.5))]
```

```
  test <- index[((length(index)*0.5)+1):length(index)]
  train_lm <- lm(as.formula(lm_usa$call), data = usa_data[train, ])
  unscaled_fitted_values <- (train_lm$fitted.values *
attr(train_lm$fitted.values, "scaled:scale"))+attr(train_lm$fitted.values,
"scaled:center")
  train_rmse <- sqrt(mean((usa_data[train, "job_sat"] -
unscaled_fitted_values)^2))
  prediction <- predict(train_lm, newdata = usa_data[test,])
  unscaled_predictions <- (prediction * attr(train_lm$fitted.values,
"scaled:scale"))+attr(train_lm$fitted.values, "scaled:center")
  test_rmse <- sqrt(mean((usa_data[test, "job_sat"] - unscaled_predictions)^2))
  results[i,1] <- train_rmse
  results[i,2] <- test_rmse
  results[i,3] <- summary(train_lm)$r.squared
}
colnames(results) <- c("IS-RMSE", "OOS-RMSE", "R2")

# Calculate the range of RMSE for models with 0.343 <= R2 <= 0.454
range(results[results[,"R2"] >= 0.434 & results[,"R2"] <= 0.454,"OOS-RMSE"])

# Calculate the number of models in this range
sum(results[,"R2"] >= 0.434 & results[,"R2"] <= 0.454)

par(mfrow=c(1,2))
# view the relationship between OOS-MSE and R2
plot(results[,"R2"], results[,"OOS-RMSE"], main = "1.a", xlab = "R-Squared",
ylab = "Out-of-sample RMSE")
abline(a = lm(`OOS-RMSE` ~ R2, data = as.data.frame(results))$coefficients[1], b
= lm(`OOS-RMSE` ~ R2, data = as.data.frame(results))$coefficients[2])
abline(v = c(0.434, 0.454), lty = 2)

# view the relationship between OOS-MSE and IS-MSE
plot(results[,"IS-RMSE"], results[,"OOS-RMSE"], main = "1.b", xlab = "In-sample
RMSE", ylab = "Out-of-sample RMSE")
abline(a = lm(`OOS-RMSE` ~ `IS-RMSE`, data =
as.data.frame(results))$coefficients[1], b = lm(`OOS-RMSE` ~ `IS-RMSE`, data =
as.data.frame(results))$coefficients[2])

# distribution of the RMSE for R2 >0.44 < 0.45
plot(density(results[results[,"R2"] > 0.434 & results[,"R2"] < 0.454, "OOS-
RMSE"]), xlab = "RMSE", main = "Density Plot of RMSE")

## Use 2005 data to predict 2015 data ----

# The Drabe USA model

sex_proc <- usa_data$sex - 1
job_sat_proc <- (usa_data$job_sat - 5.27) / 1.20
Age_cat_proc1 <- rep(0, nrow(usa_data))
Age_cat_proc1[usa_data$Age_cat == 1] <- 1
Age_cat_proc2 <- rep(0, nrow(usa_data))
Age_cat_proc2[usa_data$Age_cat == 2] <- 1
Age_cat_proc3 <- rep(0, nrow(usa_data))
Age_cat_proc3[usa_data$Age_cat == 3] <- 1
educyrs_proc <- (usa_data$educyrs - 12.89)/5.23
income_proc <- (usa_data$income - 2.65) / 1.10
advancement_proc <- (usa_data$advancement - 2.65) / 1.14
security_proc <- (usa_data$security - 3.74) / 1.16
interesting_proc <- (usa_data$interesting - 3.94) / 1.02
```

```
rel_clgs_proc <- (usa_data$rel_clgs - 4.17) / 0.77
rel_mgmt_proc <- (usa_data$rel_mgmt - 3.88) / 0.93
independent_proc <- (usa_data$independent - 3.69) / 1.25

prediction <- -0.01*sex_proc + 0.04*Age_cat_proc2 + 0.085*income_proc +
0.059*advancement_proc + 0.067*security_proc + 0.309*interesting_proc +
0.04*independent_proc + 0.288*rel_mgmt_proc + 0.115*rel_clgs_proc
unscale_prediction <- (prediction * 1.20)+5.27
RMSE_2005_2015 <- sqrt(mean((usa_data$job_sat - unscale_prediction)^2))
MAE_2005_205 <- mean(abs(usa_data$job_sat - unscale_prediction))

# The Drabe Germany model

sex_proc <- de_data$sex - 1
job_sat_proc <- (de_data$job_sat - 5.27) / 1.20
Age_cat_proc1 <- rep(0, nrow(de_data))
Age_cat_proc1[de_data$Age_cat == 1] <- 1
Age_cat_proc2 <- rep(0, nrow(de_data))
Age_cat_proc2[de_data$Age_cat == 2] <- 1
Age_cat_proc3 <- rep(0, nrow(de_data))
Age_cat_proc3[de_data$Age_cat == 3] <- 1
educyrs_proc <- (de_data$educyrs - 12.89)/5.23
income_proc <- (de_data$income - 2.65) / 1.10
advancement_proc <- (de_data$advancement - 2.65) / 1.14
security_proc <- (de_data$security - 3.74) / 1.16
interesting_proc <- (de_data$interesting - 3.94) / 1.02
rel_clgs_proc <- (de_data$rel_clgs - 4.17) / 0.77
rel_mgmt_proc <- (de_data$rel_mgmt - 3.88) / 0.93
independent_proc <- (de_data$independent - 3.69) / 1.25

prediction <- -0.01*sex_proc + 0.04*Age_cat_proc2 + 0.085*income_proc +
0.059*advancement_proc + 0.067*security_proc + 0.309*interesting_proc +
0.04*independent_proc + 0.288*rel_mgmt_proc + 0.115*rel_clgs_proc
unscale_prediction <- (prediction * 1.20)+5.27
RMSE_2005_2015 <- sqrt(mean((de_data$job_sat - unscale_prediction)^2))
MAE_2005_205 <- mean(abs(de_data$job_sat - unscale_prediction))

# The Drabe Japan model

sex_proc <- jp_data$sex - 1
job_sat_proc <- (jp_data$job_sat - 5.27) / 1.20
Age_cat_proc1 <- rep(0, nrow(jp_data))
Age_cat_proc1[jp_data$Age_cat == 1] <- 1
Age_cat_proc2 <- rep(0, nrow(jp_data))
Age_cat_proc2[jp_data$Age_cat == 2] <- 1
Age_cat_proc3 <- rep(0, nrow(jp_data))
Age_cat_proc3[jp_data$Age_cat == 3] <- 1
educyrs_proc <- (jp_data$educyrs - 12.89)/5.23
income_proc <- (jp_data$income - 2.65) / 1.10
advancement_proc <- (jp_data$advancement - 2.65) / 1.14
security_proc <- (jp_data$security - 3.74) / 1.16
interesting_proc <- (jp_data$interesting - 3.94) / 1.02
rel_clgs_proc <- (jp_data$rel_clgs - 4.17) / 0.77
rel_mgmt_proc <- (jp_data$rel_mgmt - 3.88) / 0.93
independent_proc <- (jp_data$independent - 3.69) / 1.25

prediction <- -0.01*sex_proc + 0.04*Age_cat_proc2 + 0.085*income_proc +
0.059*advancement_proc + 0.067*security_proc + 0.309*interesting_proc +
0.04*independent_proc + 0.288*rel_mgmt_proc + 0.115*rel_clgs_proc
```

```
unscale_prediction <- (prediction * 1.20)+5.27
RMSE_2005_2015 <- sqrt(mean((jp_data$job_sat - unscale_prediction)^2))
MAE_2005_205 <- mean(abs(jp_data$job_sat - unscale_prediction))
```

Appendix C

To illustrate that the results described in the thought experiment and visualised in Figure 1 are robust for different sizes of training and holdout samples, we re-run the analyses using two additional sample sizes of $n = 350$ and $n = 450$. We visualise these results in Figures C1 and C2, respectively.

While the results remain largely consistent with the original analysis as sample size increases, we also find that the range of the out-of-sample RMSE given an in-sample metric (R^2 or in-sample RMSE) becomes narrower with increasing sample sizes in training and holdout samples—see, for example, Figure 1a versus Figure C2a. This result is to be expected because the number of possible permutations of two samples (i.e., training and holdout samples) of 450 observations drawn from a population of 915 is dramatically reduced from that of drawing two samples of 250 observations from the same population. That is, when the proportion of respondents drawn from the population increases, the variation in estimated metrics decreases. Nonetheless, overfit remains apparent in that as explanatory power increases, predictive power decreases on average. This finding is also supported when analysing all replications with an R^2 value between 0.434 and 0.454 ($R^2 = 0.444 \pm 0.01;\ n^* = 268$) in terms of their predictive performance on the holdout sets. These results show that the out-of-sample RMSE estimated on a holdout set for the samples ranges from 0.85 to 1.00—a 19.1% increase in uncertainty associated with the predictions generated from the model.

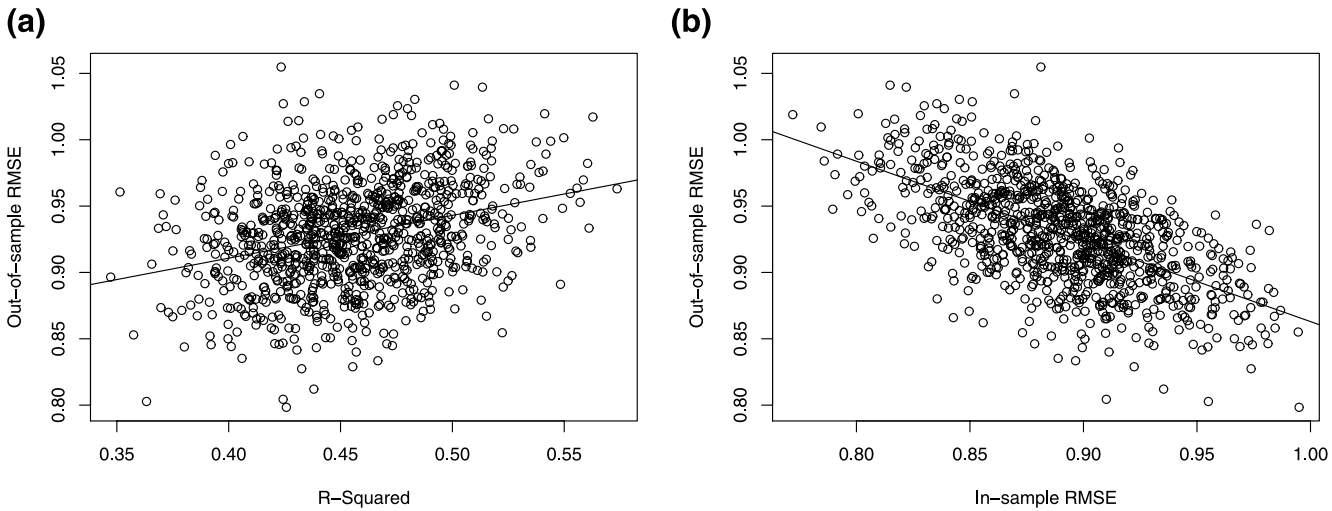

FIGURE C1 The relationship between (a) (in-sample) R^2 and out-of-sample root mean square error (RMSE); and (b) in-sample RMSE and out-of-sample RMSE for $n = 350$

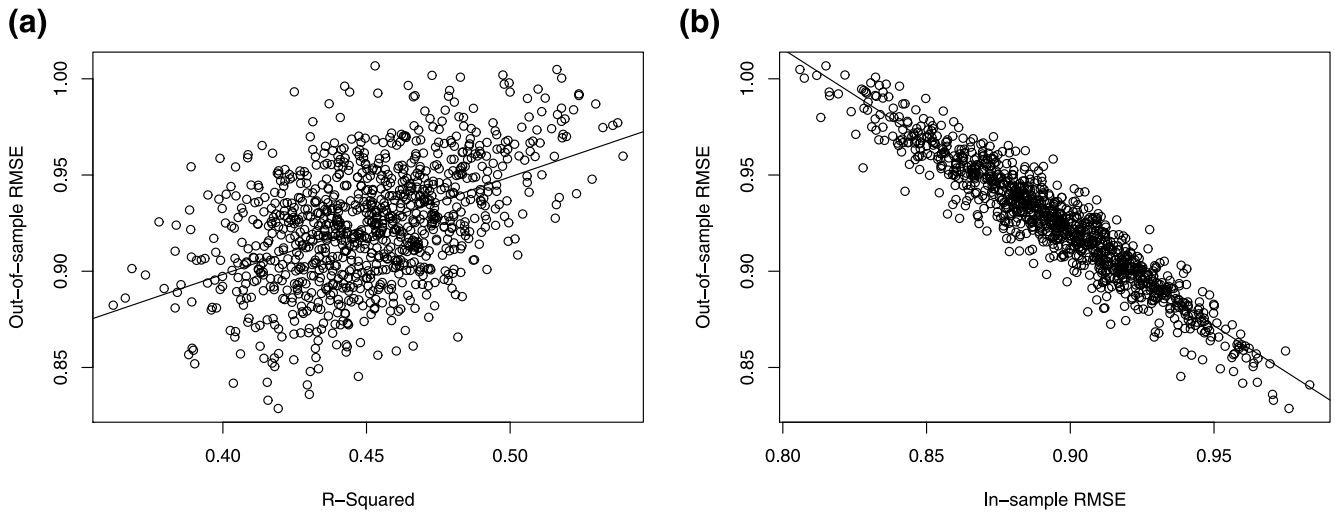

FIGURE C2 The relationship between (a) (in-sample) R^2 and out-of-sample root mean square error (RMSE); and (b) in-sample RMSE and out-of-sample RMSE for $n = 450$