Prediction in Human Recourse Managment

A gab between rhetoric and reality

Abhinna Shah Ariel Jeremi Wowor Bianca Ricci
Marcel Dieti

2025-07-08

we will write a abstract here at the very end.

Table of contents

1 Introduction

Statistical modeling in the social sciences serves two distinct but complementary purposes: explanation and prediction. Explanatory modeling, the dominant paradigm in Human Resource Management (HRM), focuses on testing theoretical propositions by estimating associations among constructs and evaluating in-sample fit via statistics such as R^2, F-tests, and SEM indices (e.g., CFI, RMSEA)[1] (Shmueli, 2010). In contrast, predictive modeling assesses a model's capacity to generate accurate forecasts for new, unseen observations using out-of-sample performance metrics like root mean squared error (RMSE) and mean absolute error (MAE) (Shmueli & Koppius, 2011). The distinction has profound methodological implications; explanatory analyses prioritize causal inference and parameter significance, whereas predictive analyses emphasize generalizability and error minimization, often employing train–test splits or cross-validation techniques (Hastie et al., 2013). Scholars have long cautioned that conflating explanation with prediction may lead to overfitting [2] and misleading inferences when models optimized for in-sample performance fail to generalize beyond the estimation dataset (Forster, 2002; Forster & Sober, 1994).

This distinction is especially critical in an applied field like HRM, where research often culminates in managerial recommendations. Such prescriptions—implying that "If an organization implements practice X, then outcome Y will improve"—are inherently predictive. They presuppose that the underlying statistical model can reliably forecast how Y will change when X is altered. Yet, a significant gap exists between this predictive goal and the field's methodological practices. In a landmark review, Sarstedt and Danks (2021) demonstrated that while 99% of HRM studies advance prescriptive claims, they rely solely on explanatory metrics for model validation. This creates a fundamental "gap between rhetoric and reality," calling into question the real-world utility of many research-based recommendations (Sarstedt & Danks, 2021).

The present thesis aims to replicate and extend Sarstedt and Danks's (2021) investigation, empirically demonstrating the gap between explanatory and predictive modeling in HRM. In particular, we are showing that models with similar degrees of explanatory power can perform very differently in term of predictive power. In addition, we are going to demonstrate that a model developed in one context, for example in a country, can not necessarly generalize well if applied to other contexts. Using the International Social Survey Programme (ISSP) 2015 Work Orientation dataset, we re-estimate a job satisfaction model originally proposed by Drabe et al. (2015) across three distinct national contexts: the United States, Germany, and Japan. Through a systematic evaluation of in-sample R^2 and out-of-sample RMSE and MAE—employing train–hold-out splits and k-fold cross-validation—we seek to illustrate the

[1] *Structural Equation Modeling (SEM) fit indices, such as the Comparative Fit Index (CFI) and Root Mean Square Error of Approximation (RMSEA), evaluate how well a proposed model reproduces observed data patterns. Acceptable model fit is typically indicated by CFI values .90 and RMSEA values .08*

[2] *Overfitting occurs when a model captures noise unique to the estimation dataset rather than the true underlying patterns. Predictive validation techniques, such as train/test splits or k-fold cross-validation (CV), help detect overfitting and mitigate the risk of generating misleading managerial insights.*

variability of predictive performance relative to explanatory fit. Furthermore, we integrate Necessary Condition Analysis (NCA) to identify conditions that must be present for high job satisfaction, thereby enriching explanatory insights with boundary constraints on outcome attainment.

Problem Statement

2 Theory and Hypothesis

Evaluating Predictive Models: Justification for Key Metrics

In assessing the predictive utility of a model, it is crucial to select metrics that capture different dimensions of performance. Among the various available metrics, R-squared (R^2), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE) are particularly valuable because they provide complementary insights into goodness-of-fit, error magnitude, and interpretability.

R-squared (R^2)

R-squared quantifies the proportion of variance in the dependent variable that is explained by the independent variables. It serves as a central indicator of in-sample goodness-of-fit and helps assess the strength of association within the estimation dataset (Sarstedt & Danks, 2021). R^2 is widely used due to its intuitive, scale-free interpretation as a percentage (ranging from 0 to 100), making it accessible for communicating how much of the outcome variance is accounted for by the model. However, a high in-sample R^2 is not necessarily indicative of strong predictive power. Models can exhibit high R^2 values by overfitting to sample-specific noise, which undermines generalizability to new data (Sarstedt & Danks, 2021).

Root Mean Square Error (RMSE)

RMSE is a commonly used metric for assessing predictive accuracy and reflects the standard deviation of residuals, i.e., the differences between predicted and observed values (Sarstedt & Danks, 2021). As a quadratic scoring rule, RMSE disproportionately penalizes larger errors, making it especially sensitive to extreme deviations. One of RMSE's primary advantages is that it is expressed in the same units as the dependent variable, which enhances its practical interpretability. For example, in a model predicting job satisfaction on a 1–7 scale, an out-of-sample RMSE of 0.9 conveys the typical magnitude of prediction error in real terms (Sarstedt & Danks, 2021). This makes RMSE a particularly informative metric for evaluating predictive utility on holdout data.

Mean Absolute Error (MAE)

MAE measures the average absolute difference between predicted and observed values. Unlike RMSE, it applies equal weight to all errors, regardless of their size, offering a linear and less distortion-prone assessment of predictive accuracy.MAE is valued for its simplicity and intuitive appeal. It directly communicates how far, on average, the model's predictions deviate from actual outcomes (Sarstedt & Danks, 2021). For instance, a MAE of 0.7 in a job satisfaction model suggests that predictions are typically off by 0.7 points, a figure that is easily interpretable for both technical and non-technical audiences.

Understanding Downturns in Predictive Accuracy

A frequent observation in predictive modeling is that performance often degrades when models are applied to new data or different contexts. This reduction in predictive accuracy can be attributed to both statistical factors and real-world contextual changes.

Statistical Explanations

Overfitting:

Overfitting occurs when a model is excessively tailored to the training data, capturing random noise alongside meaningful patterns. This results in high explanatory power (e.g., high in-sample R^2) but poor generalization to unseen data (Sarstedt & Danks, 2021). A strong in-sample fit does not guarantee low prediction error on new data, as models may fail to reproduce their performance beyond the original sample.

Sample-Specific Variance:

Every dataset represents a random sample from a broader population and will therefore contain sample-specific idiosyncrasies. A model optimized on one sample may perform poorly on another due to sampling variability alone. Evidence of this effect can be seen in studies where models with similar R^2 scores display wide variation in out-of-sample RMSE values across multiple resampled datasets.

Substantive Explanations in Human Resource Management (HRM)

Limited Generalizability Across Contexts:

Predictive models in HRM often struggle to generalize across national, organizational, or cultural contexts. Social, institutional, and cultural variables influence employee attitudes and behaviors, and a model trained in one context (e.g., Japan) may not perform well in

another (e.g., the United States) due to differences in underlying drivers (Sarstedt & Danks, 2021).

Temporal Instability:

The factors influencing outcomes like job satisfaction are not static. Changes in the economy, labor market conditions, technological developments, or generational shifts can alter the relevance or strength of predictors over time. Empirical research has shown that models estimated using data from one time period (e.g., 2005) often perform poorly when applied to later periods (e.g., 2015), suggesting a decline in predictive relevance due to evolving conditions (Sarstedt & Danks, 2021).

2.3 Research Extension

After replication, this research is extended with a detailed analysis of the covariates to identify the most crucial determinants. To compare variables with different units of measurement, they must be placed on a common scale. Therefore, the standardized coefficient is used to estimate the crucial determinants in the job satisfaction model, as it can assess the relative magnitude of each independent variable and enables an assessment of the relative predictive power of these variables.(Hair, 2018).

However, standardized coefficients indicate which factors are influential but not whether a certain condition must be present for an outcome to occur. Relying solely on a traditional coefficient analysis with standardized coefficients would mean classifying the variable with the highest coefficient as the most important determinant and risking the neglect of other variables with lower values. This is where Necessary Condition Analysis (NCA) comes in.

NCA is a relatively recent analytical method introduced by (Dul, 2016) to identify necessary conditions within datasets. These are conditions that must be present for an outcome to occur. In 2020, Richter et al. extended NCA by integrating it with regression-based analysis using PLS-SEM (Richter & Sarstedt, 2020). In this study, instead of applying PLS-SEM, NCA is combined with a simple multiple regression model following the guidelines proposed by Richter et al. Details about the NCA methodology will be discussed in Chapter 3.

3 Method

Data Source and Variable Description

ISSP 2015 Work Orientation Dataset

The ISSP 2015 Work Orientation module provides harmonized, publicly available survey data from multiple countries. We selected three national subsamples—United States (USA; n = 915), Germany (GER; n = 899), and Japan (JPN; n = 635). The dependent variable, job satisfaction (job_sat), was measured on a 1–7 Likert scale. Independent variables matched those used by Drabe et al. These are as follows: (Drabe et al., 2015).

Key Constructs and Measurement

Table 1: Variables defining Job satisfaction

Variable	Intrinsic/Extrinsic
Gender (sex)	Intrinsic demographic
Age Category (Age_cat: 35, 36–49, 50)	Intrinsic demographic
Education (educyrs)	Intrinsic human-capital
Income	Extrinsic reward
Job security (security)	Extrinsic reward
Job interest (interesting)	Intrinsic job characteristic
Autonomy (independent)	Intrinsic job characteristic
Relationship with management (rel_mgmt)	Social work context
Relationship with colleagues (rel_clgs)	Social work context
Advancement opportunities (advancement)	Extrinsic reward
Job satisfaction (job_sat)	Outcome Variable

Data Preprocessing and Preparation

Handling Missing Data and Case-Wise Deletion

Survey items exhibited uneven missingness across countries (up to 35% for some items). In line with Sarstedt and Danks (2021), we applied case-wise deletion to maintain consistent sample composition for comparative modeling, accepting potential reductions in statistical power to preserve internal validity.

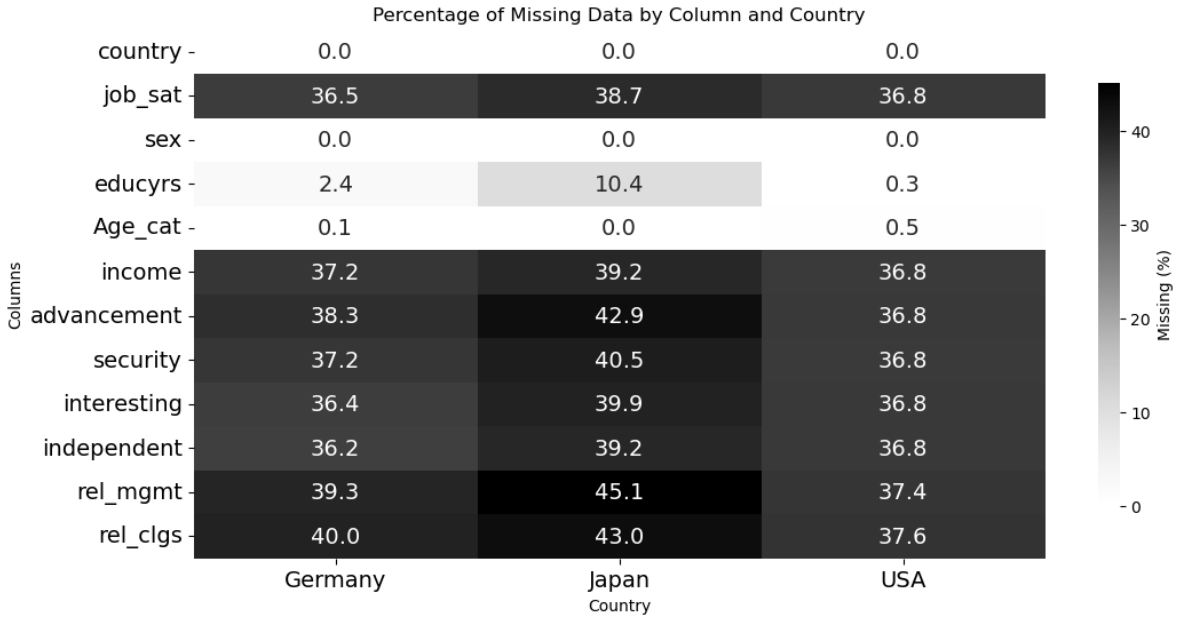

Figure 1: Heatmap of percentage Missing Data by column.

Reverse Coding

In the ISSP survey, higher numerical responses sometimes indicate less favorable conditions (e.g., "1 = Very satisfied" to "7 = Very dissatisfied"). To ensure that all predictors align directionally with job satisfaction (higher = better), we reversed scales so that greater values consistently denote more positive levels. This simplifies interpretation: a positive regression coefficient uniformly implies that increases in the predictor raise satisfaction.

3.3 Determinants of Job Satisfaction

After examining the relationship between dependent and independent variables using coefficients, the next step with NCA is to identify areas in scatter plots that indicate the presence of a necessary condition (Richter & Sarstedt, 2020). In our model, the independent variables (X) presented in **Table XX** are treated as potential necessary conditions for Job Satisfaction (Y). This implies that if a necessary condition is not met, failure to achieve the desired outcome is guaranteed. However, NCA examines each variable individually, treating the necessary condition as operating in isolation and independently of context (Richter & Sarstedt, 2020). Therefore, the absence of a necessary condition cannot be compensated for by other conditions or determinants.

NCA uses scatter plots to visualize the necessity relationship between an independent variable and an outcome, dividing the plot into two distinct areas, as shown in **Figure xx**. The area where observations occur is called the total zone (scope), while the area without observations is referred to as the empty zone (ceiling).

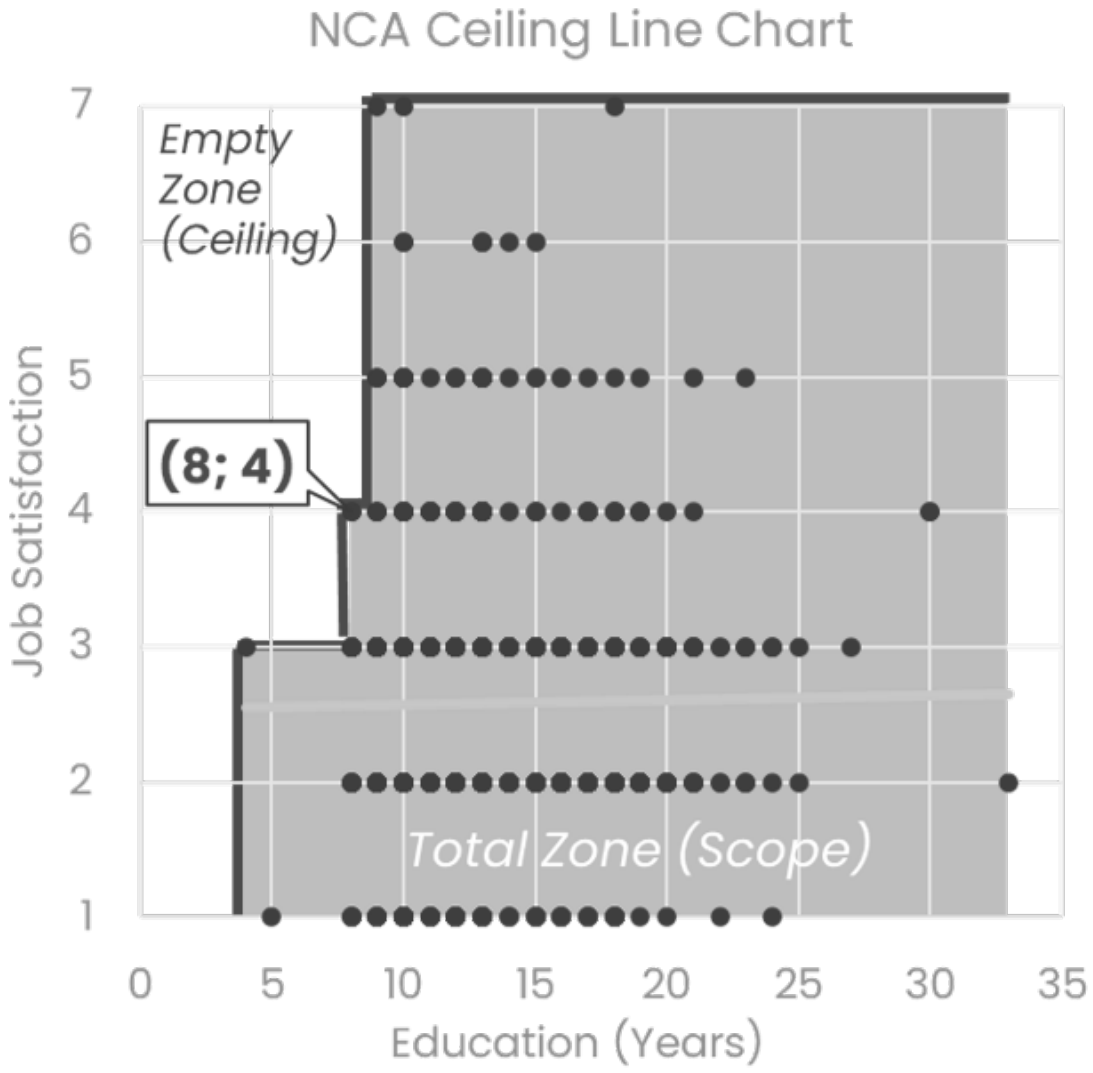

Figure 2: Ceiling Line Chart for Education (Years) as Predictor of Job Satisfaction

An empty zone indicates a necessary condition, with its size providing a way to assess the strength of that requirement. To quantify this strength, (Richter & Sarstedt, 2020) describe two key NCA parameters, which are the ceiling accuracy and the necessity effect size d. The necessity effect size d measures how much of the outcome space is constrained by a necessary

condition, ranging from 0 to 1. Values of d are interpreted as small (0–0.1), medium (0.1–0.3), large (0.3–0.5), and very large (0.5) effects (Dul, 2016). While d 0.1 is often used to support necessity hypotheses, its value indicates the substantive importance of the condition. Ceiling accuracy reflects the percentage of observations on or below the ceiling line. The CE-FDH line always achieves 100% accuracy, while lines like CR-FDH may have lower accuracy. Although no strict threshold exists, benchmarks (e.g., 95%) can help evaluate solution quality (Dul, 2016). However, because the data in this model are discrete rather than continuous, only the CE-FDH line is relevant, with 100% accuracy expected. Therefore, the necessity effect size measurement is the primary focus in this NCA.

4 Results -> Findings/Analysis/Interpretation

Diagnostics and Assumption Checks

Linearity, Homoscedasticity, Normality

Collinearity Diagnostics and VIF

Replication Results

In this chapter, we build on the models developed in the previous sections using the ISSP dataset to explore two key ideas.
First, we examine whether a model that fits its training data well (i.e., with high explanatory power) also performs well in predicting new data. Second, we assess whether a model trained in one specific context (such as one country) can generalize to different contexts.
All analyses were conducted in R and later replicated in Python to ensure robustness. Since results were consistent across both environments, we report only the R-based plots here for clarity.

Model Variability within Context

To address the first question, we focused on the German dataset and built 1000 models using repeated random sampling. In each replication, we drew a sample of 500 observations, then split it into two equally sized subsets: one for training and one as a holdout set.
Each model was trained on its training set, and we computed two in-sample metrics to evaluate explanatory power: R-squared and root mean square error (RMSE). We then used the same model to generate predictions on the holdout set and calculated the out-of-sample RMSE as a measure of predictive power.
Figures 1 and 2 illustrate how these two types of performance relate. In Figure 1, we plot R-squared (x-axis) against out-of-sample RMSE (y-axis); in Figure 2, the x-axis is the in-sample

RMSE.
In both plots, we observe that models with similar in-sample performance can vary widely in their out-of-sample predictive accuracy. This means that high explanatory power does not guarantee high predictive power.
Interestingly, Figure 1 shows that higher R-squared values tend to be associated with higher prediction error, suggesting possible overfitting. Similarly, Figure 2 shows a negative trend between in-sample and out-of-sample RMSE: models that fit the training data very well often perform worse on new data.

To explore this further, we zoomed in on the models with an R-squared close to that of the original German model (R^2 between 0.366 and 0.386). Although these models all explain roughly the same proportion of variance, their prediction errors still vary greatly, from 0.75 to 0.94, meaning a 25% increase in error between the best and worst cases.

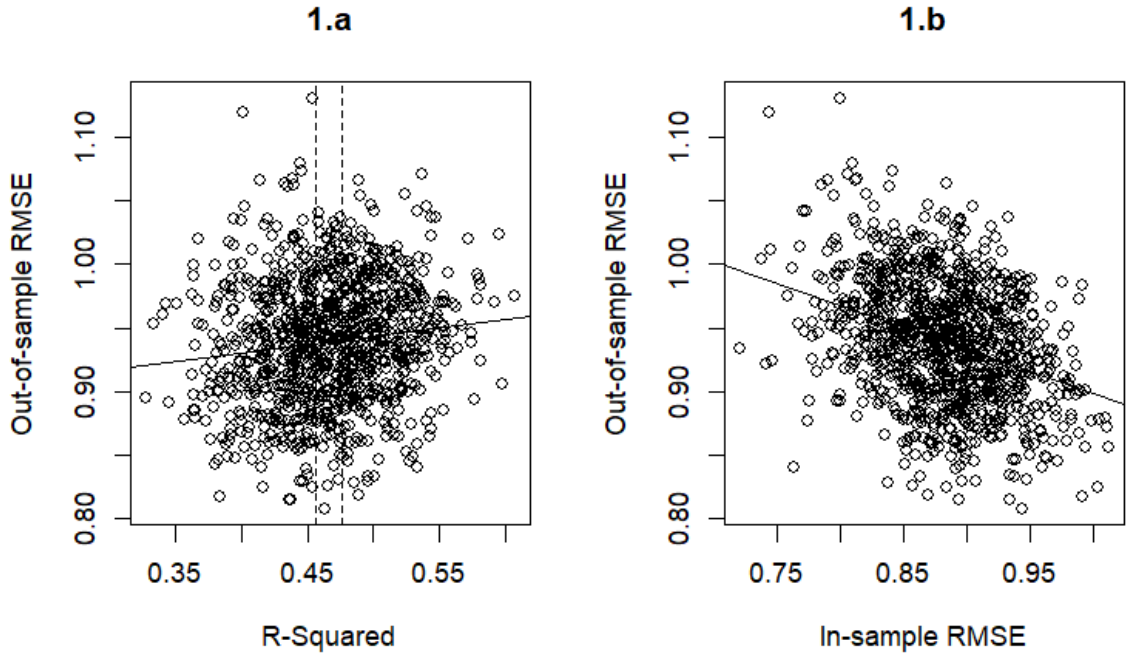

Figure 3: [1a] Relationship between (in-sample) R^2 and out-of-sample root mean square error (RMSE) [1.b] Relationship between in-sample RMSE and out-of-sample RMSE

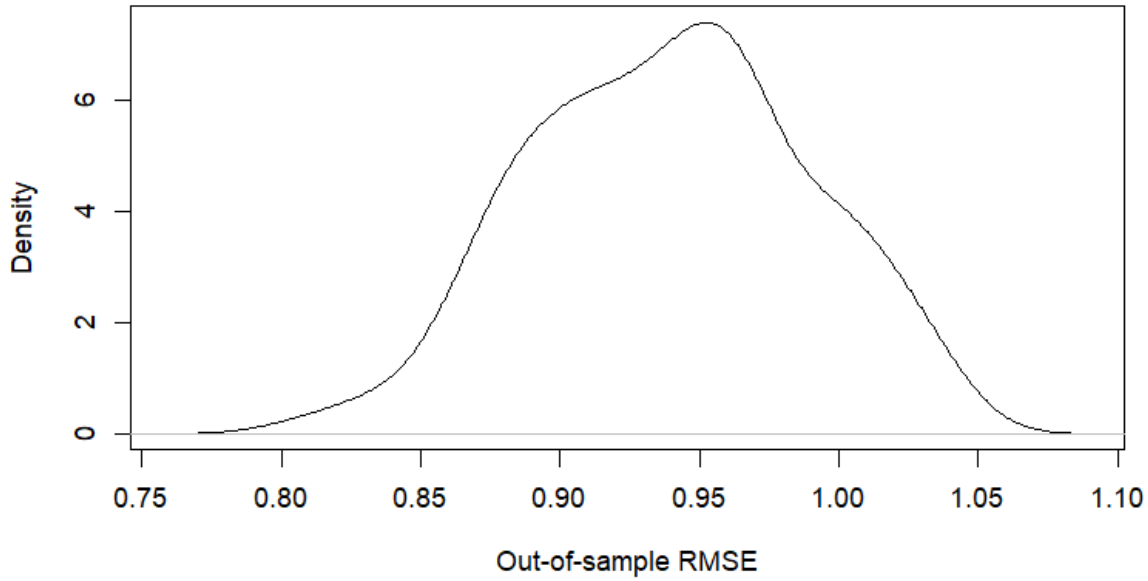

Figure 4: Density plot of predictive power for subsamples with $0.366 < R^2 < 0.386$. RMSE, root mean square error

This first analysis highlights the importance of jointly evaluating both explanatory and predictive power when assessing a model's performance. Our goal is to caution researchers against drawing prescriptive conclusions based solely on a model's in-sample fit, such as R-squared or in-sample RMSE, without verifying how well the model performs on unseen data.

As we have shown, models that explain a large proportion of variance in the training data do not necessarily perform well in predicting new observations. In fact, we observed that some models with high explanatory power exhibited relatively poor predictive performance, indicating a risk of overfitting.

A well-designed model should not only provide a good fit to its own data but also be able to generalize to new samples. However, there is no fixed relationship between in-sample metrics (e.g., R-squared, RMSE) and out-of-sample performance, which makes the predictive evaluation an essential step.

While explanatory modelling focuses on understanding relationships between variables, through the significance, direction, and size of coefficients, predictive modelling aims to accurately forecast new outcomes. These two goals are distinct, but not mutually exclusive.

By adopting an EP (Explanatory and Predictive) approach, researchers can strike a balance: developing models that both explain the data and support reliable prescriptive insights.

Ultimately, when the objective is to derive actionable conclusions, predictive power must be explicitly assessed, explanatory power alone is not enough.

Cross Country Generalizability of Predictive Power

In the second part of our analysis, we explore whether a model trained in one country can predict outcomes in other countries.
We used the full ISSP data for Germany, the USA, and Japan, building one model per country. The bar chart in Figure 4 shows the R-squared of each model on its own country's data, indicating how well it explains variation within its own context.
The heatmap in Figure 5 shows the out-of-sample RMSE when each country-specific model is used to predict data from the other countries. This allows us to compare generalizability across contexts.
From the results, the Japanese model shows to have the highest R-squared (0.508), meaning it fits its own data well. However, it performs poorly when predicting the USA data, suggesting it may be overfitting to context-specific patterns that don't transfer.
Furthermore, no other model predicts the Japanese data well, indicating that job satisfaction in Japan may follow idiosyncratic patterns not captured by models trained elsewhere.
The German model has the lowest explanatory power ($R^2 = 0.384$), but its data are reasonably well predicted by both the USA and Japan models. This may mean that the German model omits some relevant predictors or suffers from multicollinearity. At the same time, the predictors that matter in the USA and Japan might partially capture variation in the German context.

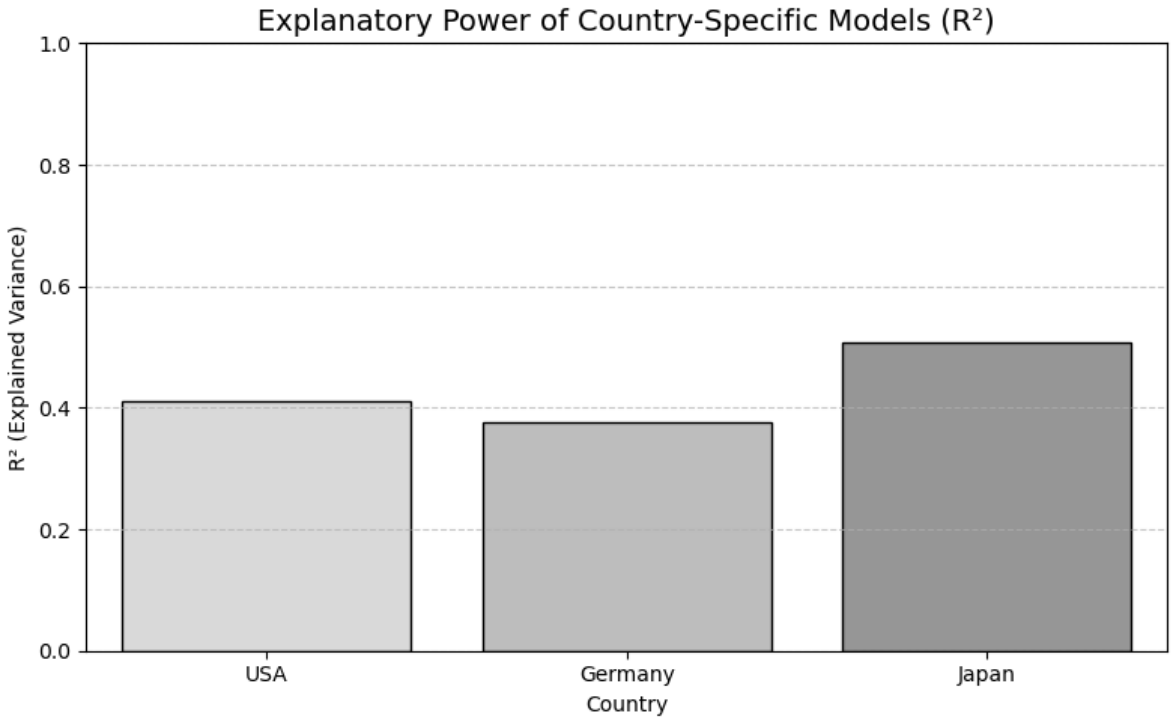

Figure 5: Ability of each country-specific model to fit its own country data

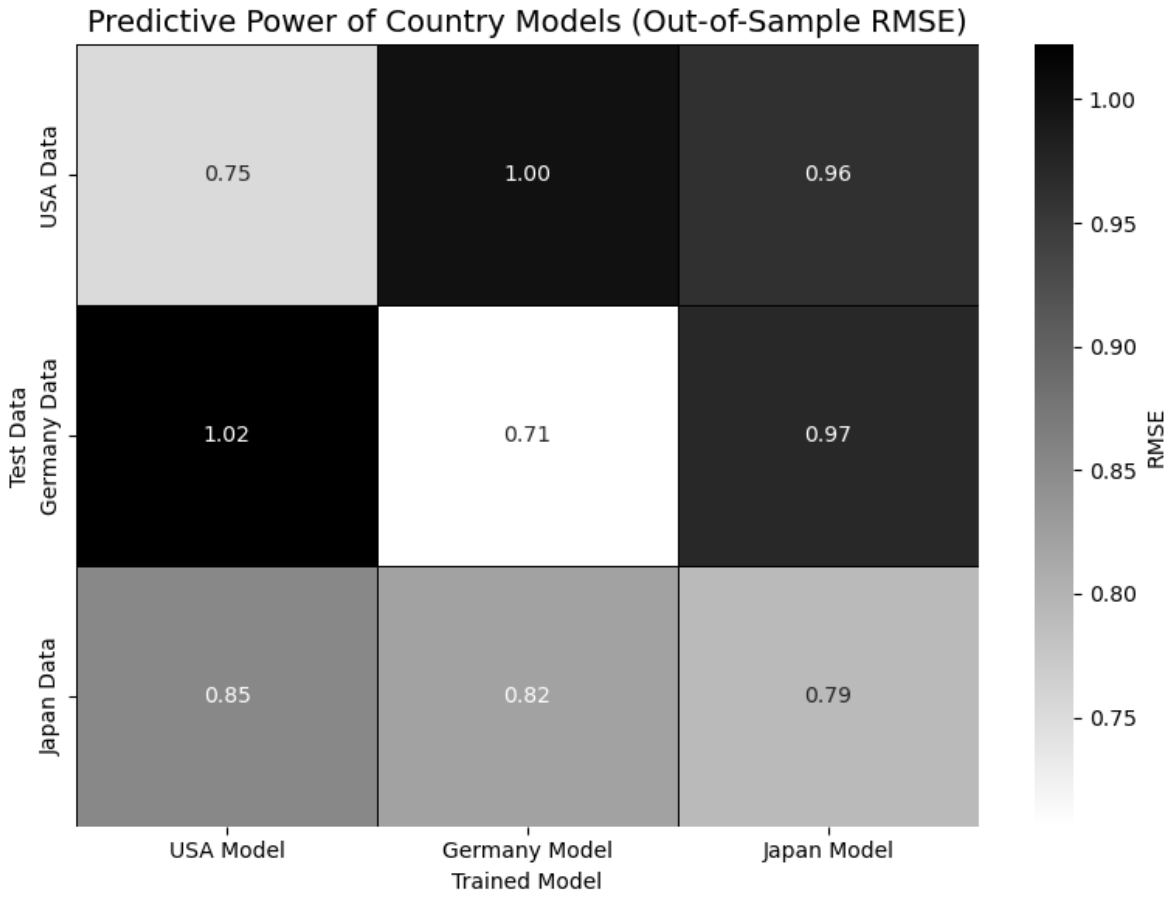

Figure 6: Predictive power of country-specific models

This second analysis shows that models trained in one country do not necessarily generalize well to other countries. Patterns that hold in one cultural or institutional context may not apply elsewhere.
We emphasize the importance of validating predictive performance across time, populations, or geographies, especially when a model is intended for broader use. Testing models on external data is key to understanding their generalizability and reliability.

4.3 Extension Results

Figure xx illustrates the main determinants of job satisfaction, ranked from highest to lowest. Having an interesting job and maintaining a good relationship with management emerge as the strongest determinants, while education ranks among the weaker predictors. The overall ordering of predictors aligns closely with the original model from [paper xyz], showing only

minor differences in the coefficient values. Although the order and direction of some predictors may vary across countries, these patterns remain consistent with those observed in the original model from [paper xyz]. The full set of coefficients across different languages (Python, R, and SmartPLS) is provided in **Table xy** in the appendix.

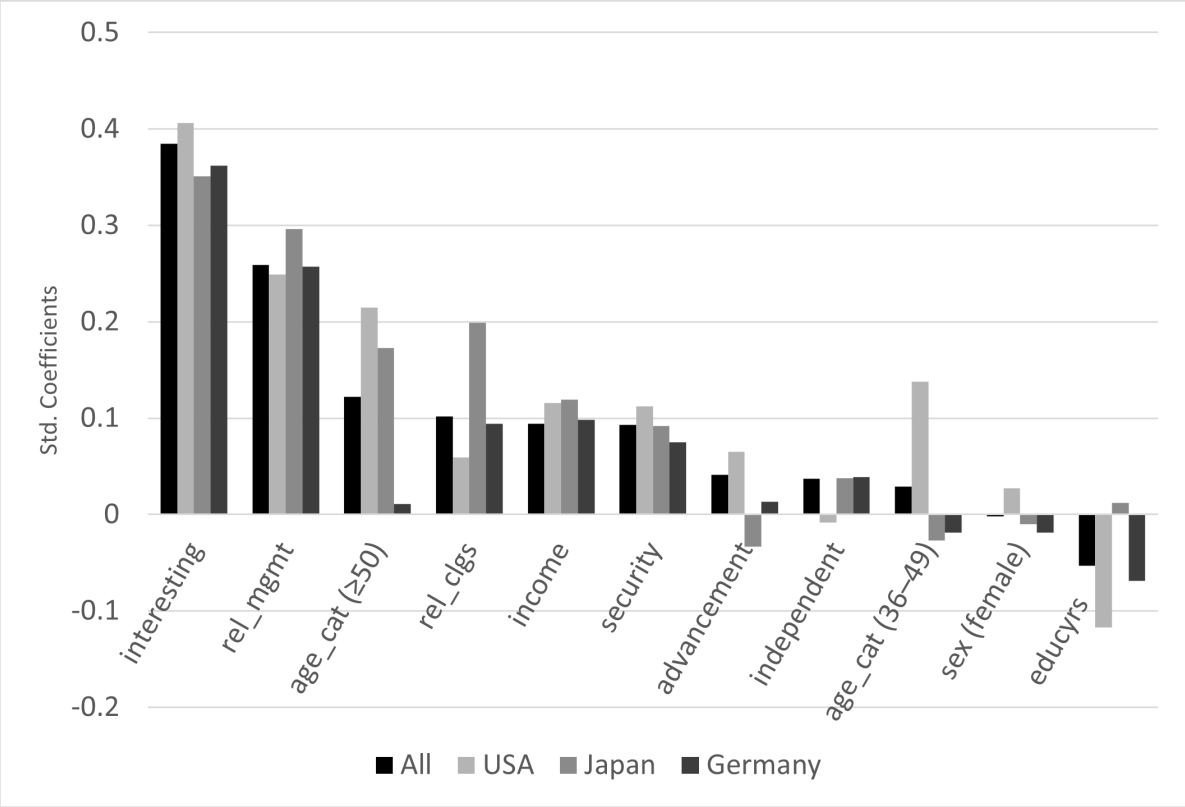

Figure 7: Cross-National Comparison of Standardized Regression Coefficients

For the NCA analysis, control variables such as age group and gender are excluded because they are categorical. The variables *interesting work*, *income*, *job security*, *working independently*, and *opportunities for advancement* show no empty zone in the scatter plots, with observations spread across all attribute levels. This indicates that there is no necessary condition for these independent variables. A summary table of all results is provided in the appendix.

In contrast, *good relationship with management*, *good relationship with colleagues*, and *years of education* emerge as variables with necessary conditions. All of these variables have 100% ceiling accuracy and effect sizes greater than zero. However, the effect size for *relationship with colleagues* is 0.042, which is considered too small to be meaningful and can therefore be ignored. *Relationship with management* shows the highest effect size at 0.125 and is classified as the strongest determinant of job satisfaction. *Years of education* has an effect size of 0.109, ranking second in NCA.

Both *relationship with management* and *years of education* demonstrate necessary conditions in the NCA. However, when compared with the standard regression analysis, they are classified differently based on their standardized coefficients. *Relationship with management* is supported by its high standardized coefficient and is classified as a strong determinant. In contrast, *years of education* is considered an insignificant determinant in the standard regression analysis due to its low standardized coefficient, but NCA provides additional insight by revealing it as a necessary condition.

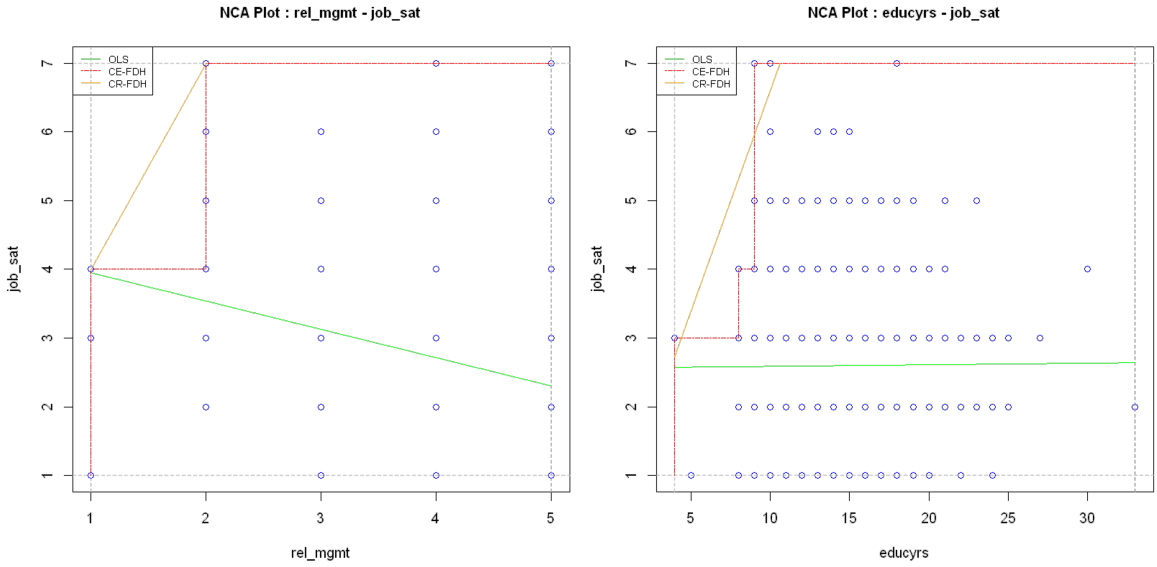

Figure 8: NCA Plots

This extension enables us not only to identify the significance of determinants based on sufficiency but also to explore whether a condition must be present at all—through necessity. This approach highlights that significant determinants can have a necessary condition or not, and the same applies to nonsignificant determinants.

Table 2: Comparison of OLS Regression Results and NCA Findings

Variable	OLS results	NCA results	Std. Coefficient	Effect size
rel_mgmt	significant determinant	necessary condition with medium effect	0.343	0.125
educyrs	nonsignificant determinant	necessary condition with medium effect	-0.018	0.109
rel_clgs	slightly significant determinant	necessary condition with small effect	0.143	0.042

Variable	OLS results	NCA results	Std. Coefficient	Effect size
interesting	significant determinant	no necessary condition	0.433	0
income	nonsignificant determinant	no necessary condition	0.089	0
security	slightly significant determinant	no necessary condition	0.075	0
independent	nonsignificant determinant	no necessary condition	0.070	0
advancement	nonsignificant determinant	no necessary condition	0.021	0

5 Implications

5.1 Discussion

Implications for HRM Theory and Practice

Challenges in Reproducing Published Analyses

Comparison of Python, R/JASP, and SmartPLS Outcomes

5.2 Conclussion and Future Research

Summary of Key Findings

Limitations of the Current Study

Suggestions for Advancing Predictive Rigour in HRM

Appendix

Use of AI

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.

Dul, J. (2016). Necessary condition analysis (NCA): Logic and methodology of "necessary but not sufficient" causality. *Sage Journals*, *19*(1), 10–52. https://doi.org/10.1177/1094428115584005

Forster, M. R. (2002). Predictive accuracy as an achievable goal of science. *Philosophy of Science*, *69*(3), S124–S134.

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

Hair, B., J. F. (2018). *Multivariate data analysis*. 259–370.

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

Richter, S., N. F., & Sarstedt, M. (2020). When predictors of outcomes are necessary: Guidelines for the combined use of PLS-SEM and NCA. *Industrial Management & Data Systems*, *80*(12), 2243–2267. https://doi.org/10.1108/IMDS-11-2019-0638

Sarstedt, M., & Danks, N. P. (2021). 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

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Shmueli, G., & Koppius, O. R. (2011). Predictive analytics in information systems research. *MIS Quarterly*, *35*(3), 553–572.