Prediction in Human Recourse Managment

A gab between rhetoric and reality

Abhinna Shah Ariel Jeremi Wowor Bianca Ricci
Marcel Dieti

2025-07-07

we will write a abstract here at the very end.

Table of contents

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]. 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) [2]. 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 [3]. Scholars have long cautioned that conflating explanation with prediction may lead to overfitting [1] and misleading inferences when models optimized for in-sample performance fail to generalize beyond the estimation dataset [4], [5].

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

[1] Overfitting occurs when a model captures random noise specific to the estimation dataset rather than the underlying phenomenon. Predictive validation (e.g., train/test split, k-fold CV) quantifies and often reveals such overfitting, safeguarding against spurious managerial guidance.

model validation. This creates a fundamental "gap between rhetoric and reality," calling into question the real-world utility of many research-based recommendations [6].

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. 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 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.

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: [7].

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. All continuous predictors were standardized (z-scores) to facilitate coefficient comparability across models.

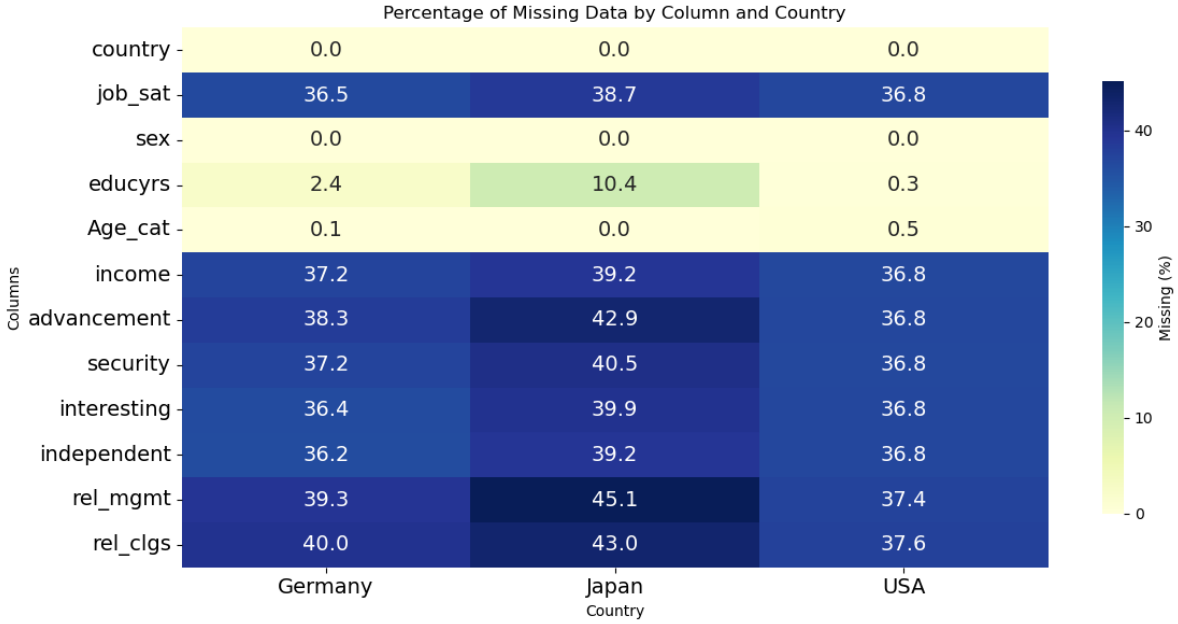

Figure 1: Heatmap of percentage Missing Data by column.

Reverse Coding and Scale Alignment

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.

Methodological Framework

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 [6]. 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 [6].

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 [6]. 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 [6]. 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 [6]. 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 [6]. 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 [6].

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 [6].

Necessary Condition Analysis (NCA)

NCA identifies variables that constitute **prerequisites** for achieving a certain outcome level. A necessary condition is one without which the outcome cannot occur, even if it alone is insufficient to guarantee the outcome[8]. NCA detects "empty zones" in scatterplots—regions above a ceiling line where no observations exist—indicating that below a threshold of the predictor, the outcome never reaches the target.

If years of education (X) is a necessary condition for job satisfaction 6 (Y), NCA might show that no respondents with fewer than 8 years of education report Y 6. Despite other favorable conditions, education below this threshold precludes high satisfaction.

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.

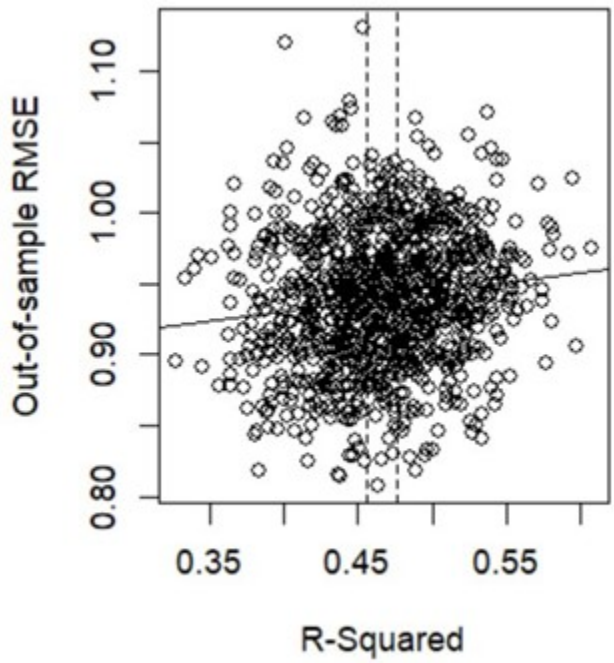

Figure 2: Relationship between (in-sample) R^2 and out-of-sample root mean square error (RMSE)

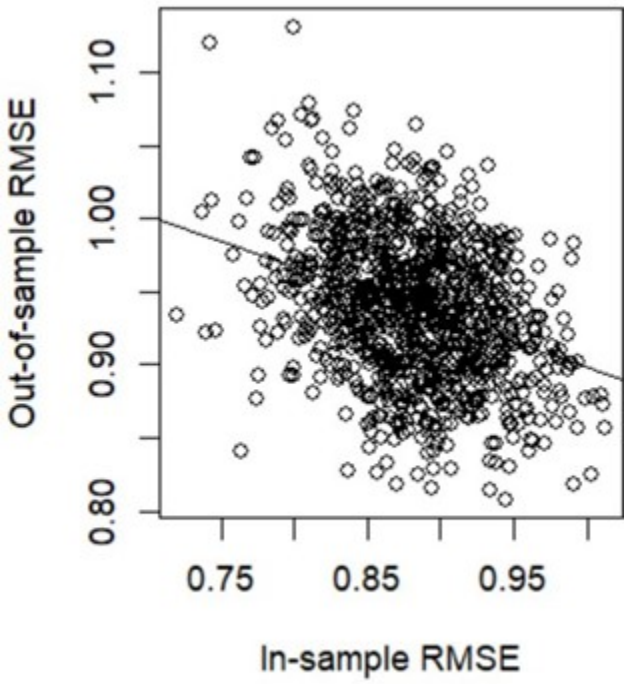

Figure 3: Relationship between in-sample RMSE and out-of-sample RMSE

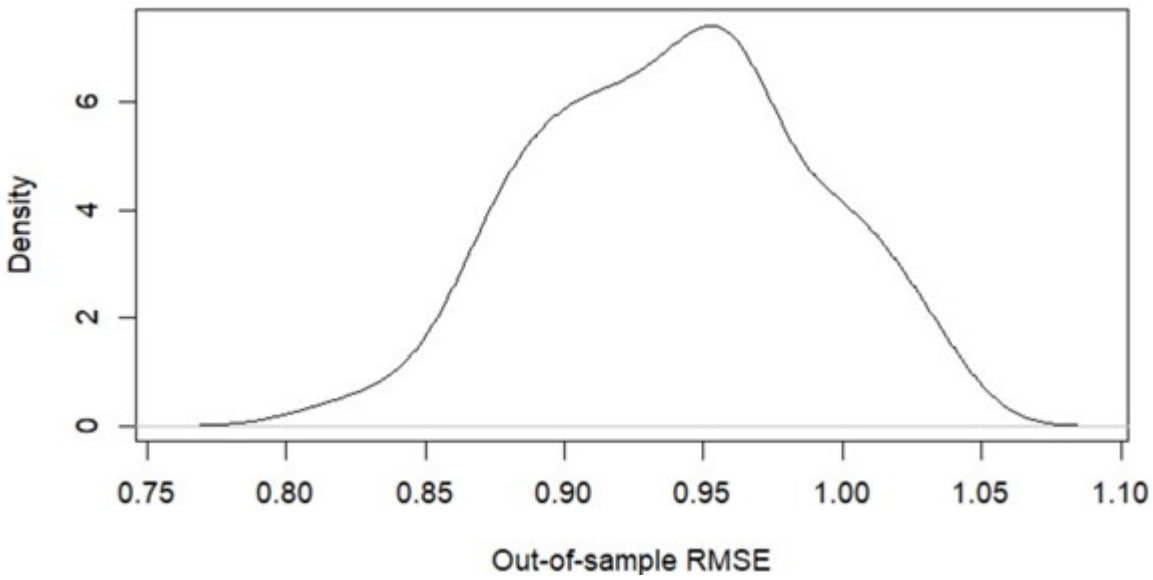

Figure 4: Density plot of predictive power for subsamples with $0.366 < R^2 < 0.386$. RMSE, root mean square error

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.

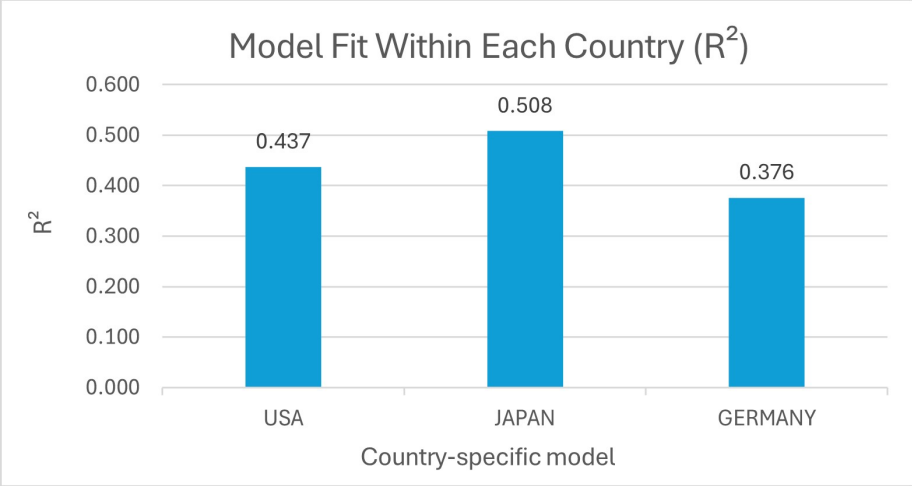

Figure 5: Ability of each country-specific model to fit its own country data

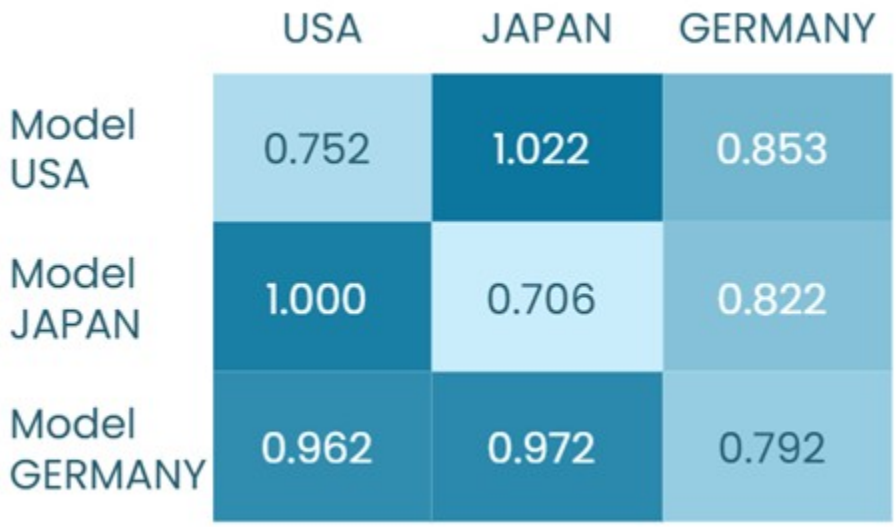

Figure 6: Predictive power of country-specific models

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.

Extension: Determinants of Job Satisfaction

Standardized Coefficients across Countries

Identifying Necessary Conditions with NCA

Discussion

Implications for HRM Theory and Practice

Challenges in Reproducing Published Analyses

Comparison of Python, R/JASP, and SmartPLS Outcomes

Conclusion and Future Research

Summary of Key Findings

Limitations of the Current Study

Suggestions for Advancing Predictive Rigour in HRM

[1] G. Shmueli, "To explain or to predict?" *Statistical Science*, vol. 25, no. 3, pp. 289–310, 2010.

[2] G. Shmueli and O. R. Koppius, "Predictive analytics in information systems research," *MIS Quarterly*, vol. 35, no. 3, pp. 553–572, 2011.

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

[4] M. R. Forster, "Predictive accuracy as an achievable goal of science," *Philosophy of Science*, vol. 69, no. 3, pp. S124–S134, 2002.

[5] M. R. Forster and E. Sober, "How to tell when simpler, more unified, or less ad hoc theories will provide more accurate predictions," *British Journal for the Philosophy of Science*, vol. 45, no. 1, pp. 1–35, 1994.

[6] M. Sarstedt and N. P. Danks, "Prediction in HRM research—a gap between rhetoric and reality," *Human Resource Management Journal*, vol. 32, no. 2, pp. 485–513, 2021, doi: 10.1111/1748-8583.12400.

[7] D. Drabe, S. Hauff, and N. F. Richter, "Job satisfaction in aging workforces: An analysis of the USA, japan and germany," *International Journal of Human Resource Management*, vol. 26, no. 6, pp. 783–805, 2015.

[8] C. Richter and L. Zheng, "Revealing prerequisites with necessary condition analysis," *Journal of Organizational Analysis*, 2025.