Psychometrika (2026), **0**, 1–22
doi:10.1017/psy.2026.10109

PSYCHOMETRIC SOCIETY

APPLICATION AND CASE STUDIES - ORIGINAL

Variable Selection via Knockoffs in Missing Data Settings with Categorical Predictors

Silvia Bacci, Emanuela Dreassi, Leonardo Grilli and Carla Rampichini

Department of Statistics, Computer Science, Applications, Università degli Studi di Firenze, Italy

Corresponding author: Silvia Bacci; Email: silvia.bacci@unifi.it

(Received 29 January 2025; revised 26 March 2026; accepted 1 April 2026)

Abstract

Large-scale assessment data typically include numerous variables, often affected by missing values. Motivated by the challenges arising in this framework, we extend the knockoffs method for selecting predictors to settings with missing values. Our proposal relies on a preliminary phase of multiple imputation (MI) of missing values. Each imputed dataset is then processed using a suitable knockoff filter. We evaluate the performance of the proposed method through simulation studies, showing satisfactory results consistent with a recently advocated cutting-edge method. We apply the method to large-scale assessment data collected by INVALSI on test scores of Italian students in grade 5, including many background variables. This case study is challenging, as most predictors have unordered categories, a setting not considered by traditional knockoff methods. In addition, some of the key predictors are affected by missing values. Our proposal to implement the knockoffs method within an MI framework is feasible, flexible, and effective.

Keywords: derandomized knockoffs; large-scale assessment data; multiple imputation; sequential knockoffs

1. Introduction

Selecting predictors with a relevant effect on an outcome is fundamental in assessing a statistical model. It is particularly challenging when numerous predictors are available. Indeed, different selection strategies may lead to different results with the risk of including in the model *null variables*, that is, predictors with a regression coefficient equal to 0 in the true model or, on the opposite, excluding *non-null variables*.

This issue is encountered in many contexts, including education, where large-scale assessment data are routinely collected to measure students' proficiency levels. In Italy, the National Institute for the Evaluation of the Education and Training System (INVALSI, https://www.invalsiopen.it/) administers standardized tests to assess students' abilities in Italian, Mathematics, and English across grades from primary school to the end of high school. INVALSI collects many variables on students and their families. Some variables, such as parents' education, have been widely recognized as determinants of student achievement. In contrast, other variables are just proxies and should be included in the model only if they are predictive of the outcome (e.g., the number of books at home or the availability of a room for studying). It is therefore essential to employ effective variable selection methods. The challenge is to apply those methods in this setting where most of the variables are affected by missing values, with rates as high as 26%.

Many methods for variable selection in regression models have been proposed in the literature, ranging from the classical forward, backward, and stepwise selection to modern regularization approaches, such as the lasso (Tibshirani, 1996) and the knockoff filter. The knockoffs framework (Barber & Candès, 2015) is based on the general idea of randomly generating variables (i.e., knockoffs) that mimic the original predictors in terms of distributive properties, but are conditionally independent of the response variable given the original variables. Then, the importance of these copies is compared with that of the original variables using a suitable statistic based on the corresponding regression coefficients. As any other selection method, the knockoff filter aims at identifying the relevant (i.e., *non-null*) variables, but it has the peculiarity of controlling for the false discovery rate (FDR), that is, the expected proportion of false discoveries (i.e., null variables wrongly declared as non-null) or, alternatively, the expected number of false discoveries, that is, the per family error rate (PFER). We stress that controlling FDR (or PFER) is particularly important in high-dimensional settings, where the large number of available predictors poses a severe risk of selecting non-relevant (i.e., *null*) variables.

Numerous contributions have followed in the knockoff literature within a few years. However, even if the theory underlying the knockoffs is valid for any joint distribution of the predictors, traditional methods for generating knockoffs are limited to continuous predictors, specifically multivariate normal distributions; furthermore, they require complete data. Therefore, they are not fully appropriate in many real-world applications. A recent extension to treat mixed types of variables is due to Kormaksson et al. (2021), which, however, still requires complete data. To our knowledge, the only approach that handles both categorical predictors and missing values is the one proposed by Xie et al. (2023). These authors consider missing values depending only on non-null variables, according to the assumption of strong missing at random (SMAR), which is more restrictive than the standard missing at random (MAR; Little & Rubin, 2002). Their approach relies on a latent variable model in which the predictors are assumed to follow a multivariate normal distribution, which is then categorized to handle binary and ordinal observed variables. The latent variable model is estimated through maximum likelihood, and the missing values are integrated out via an expectation–maximization (EM) algorithm (Dempster et al., 1977). In summary, even if Xie et al. (2023) specify a complex model, their approach relies on restrictive assumptions about both the missingness mechanism and the type of predictors. In particular, their approach can handle only continuous and binary/ordered predictors, a limitation in settings with unordered categorical predictors, such as the INVALSI data. Note that an unordered predictor with K levels can be represented by a set of $K - 1$ binary variables. However, the method of Xie et al. (2023)—as well as the standard lasso—treats these dummy variables as independent predictors, so that some may be selected and others not, depending on the arbitrary choice of the reference category (Huang et al., 2024). This approach conflates two distinct problems: deciding whether a predictor should be included in the model, and determining whether some of its categories should be collapsed. In our view, these are conceptually different steps. In the variable selection phase, the primary goal is to assess the relevance of the predictor as a whole, whereas decisions about collapsing categories should be addressed only after establishing that the variable is relevant.

To overcome the mentioned limitations, we propose to embed the knockoffs method into a multiple imputation (MI; Rubin, 1987, 1996) framework, separating the phase of treatment of missing data from the phase of variable selection. MI is a general and well-established approach to handling missing values under an MAR mechanism. Since MI produces a set of complete datasets, applying any standard variable selection method to each imputed dataset is straightforward. The challenge is to decide how to combine the results from the imputed datasets. In fact, the results can be combined with Rubin's rules for inference on model parameters, ensuring well-defined properties. However, for variable selection methods, there are no theoretical results on the properties of the combined results (see Grilli et al., 2022, and the references therein). Therefore, the choice should be guided by evidence from simulation studies.

The rest of the article is organized as follows. Section 2 illustrates the proposed knockoffs method with MI. Section 3 summarizes the results of a simulation study to assess the performance of the new method compared with the approach of Xie et al. (2023). Section 4 describes the INVALSI large-scale

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

assessment data and reports the results of the analysis based on the proposed method. Section 5 offers some final remarks. Finally, Appendix A reports the results of a data-driven simulation study tailored to an INVALSI-type setting.

2. Variable selection via knockoffs in the presence of missing values

In this section, we review the fundamentals of variable selection using the knockoff filter, starting with the original approach designed for continuous variables, then outlining a sequential knockoffs method that also properly handles categorical variables. Then, we illustrate our proposal to exploit MI to implement the knockoff filter in the presence of missing values.

2.1. The knockoff filter: Methodological background

The knockoff filter has a very recent history among methods for selecting the set of *non-null* variables, namely, the predictors that affect the outcome in the population model. Its introduction dates back to the work of Barber and Candès (2015), then extended by Candès et al. (2018). Unlike traditional variable selection methods (e.g., stepwise selection and lasso), the knockoff filter explicitly controls the PFER or FDR. Let S_D be the set of discoveries (i.e., selected variables), and S_F be the set of false discoveries (i.e., selected variables with a null effect), then

$$PFER = E(|S_F|) \tag{1}$$

and

$$FDR = E\left(\frac{|S_F|}{|S_D|}\right). \tag{2}$$

Given a regression model for the outcome Y on the observed p predictors $\mathbf{X} = [X_1, \ldots, X_p]$, the knockoffs $\tilde{\mathbf{X}} = [\tilde{X}_1, \ldots, \tilde{X}_p]$ are randomly generated variables independent of Y conditionally on $\mathbf{X}$ and with the following swapping property, holding for any subset S of predictors:

$$[\mathbf{X}, \tilde{\mathbf{X}}]_{swap(S)} \stackrel{d}{=} [\mathbf{X}, \tilde{\mathbf{X}}], \tag{3}$$

where $S \subseteq \{1, \ldots, p\}$ and $\stackrel{d}{=}$ denotes the equality in distribution. The vector $[\mathbf{X}, \tilde{\mathbf{X}}]_{swap(S)}$ is obtained by swapping the entries X_j and $\tilde{X}_j$ for any $j \in S \subseteq \{1, \ldots, p\}$, that is, the distribution of the augmented vector $[\mathbf{X}, \tilde{\mathbf{X}}]$ does not change if the original variables are swapped with the corresponding knockoff copies.

To carry out variable selection, the outcome Y is regressed on the augmented vector of predictors $[\mathbf{X}, \tilde{\mathbf{X}}]$ using a regularization procedure such as the lasso. Then, for each element X_j in $\mathbf{X}$, a statistic W_j is computed, satisfying the flip-sign property, saying that swapping the jth variable with its knockoff copy has the effect of changing the sign of W_j (Barber & Candès, 2015). A commonly used statistic W_j is the difference between the absolute values of the estimated regression coefficients $\hat{\beta}_j$ and $\hat{\tilde{\beta}}_j$ for X_j and its knockoff copy $\tilde{X}_j$, respectively,

$$W_j = |\hat{\beta}_j| - |\hat{\tilde{\beta}}_j|. \tag{4}$$

The larger the value of W_j, the stronger the evidence in favor of the relevance of X_j. Specifically, the variable X_j is selected if $W_j > \tau$, where τ is a constant chosen to ensure that the PFER or the FDR is not greater than the desired level (see Candès et al., 2018, Equation (3.9)).

Since the set of selected variables according to the original knockoffs approach may depend on the specific values randomly generated for the knockoffs $\tilde{\mathbf{X}}$, in the derandomized knockoffs approach (Ren et al., 2023), the non-uniqueness of the knockoff solution is addressed by repeating the described procedure a given number of times. Then, variables chosen in at least a given proportion of cases are selected. Usually, a *threshold* equal to 0.50 is used over 31 replications.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

The cited knockoff methods implicitly assume continuous predictors. To broaden the applicability of the knockoffs filter approach, Kormaksson et al. (2021) proposed an algorithm that handles both continuous and categorical variables. The proposed sequential knockoff method differs in the way knockoffs are generated. Specifically, the knockoff $\tilde{X}_j$ is randomly sampled from a normal distribution when X_j is a continuous variable and from a multinomial distribution when X_j is a categorical variable. The characteristic parameters of these distributions (i.e., mean and variance in the normal case and vector of probabilities in the multinomial case) are estimated through a sequential procedure: for $j = 1,\ldots,p$, the predictor X_j is regressed on the remaining predictors $\mathbf{X}_{-j}$ and on the knockoffs generated in the previous steps $[\tilde{X}_1,\ldots,\tilde{X}_{j-1}]$. The model for knockoff generation is a penalized linear model when X_j is continuous and a penalized multinomial logit model when X_j is categorical.

After the generation of knockoffs, the comparison among regression coefficients $\hat{\beta}_j$ and $\hat{\tilde{\beta}}_j$ proceeds along the same lines as Candès et al. (2018), with multiple knockoffs generation and, consequently, multiple selections as in Ren et al. (2023).

It is worth noting that when a categorical variable with K levels is used as a predictor, it is represented in the model by $K-1$ dummy variables. The standard lasso treats those dummy variables as distinct predictors, so it may happen that only some of them are selected, depending on the arbitrary choice of the reference category. To avoid this issue, one can rely on a grouped lasso (Yuan & Lin, 2006) that treats the set of dummies as a whole. Under grouped lasso, the statistic W_j in (4) modifies as follows:

$$W_j = \max_k \left\{ \left| |\hat{\beta}_{jk}| - |\hat{\tilde{\beta}}_{jk}| \right| \right\}, \tag{5}$$

where $\hat{\beta}_{jk}$ and $\hat{\tilde{\beta}}_{jk}$ are the estimated regression coefficients of the dummy for category k ($k = 2,\ldots,K$) of variable j and its knockoff copy, respectively. Other approaches to handling grouped variables in the knockoffs framework are described in Dai and Barber (2016) and Katsevich and Sabatti (2019), though they consider only continuous variables.

The sequential knockoffs procedure was recently optimized computationally by Zimmermann et al. (2024) through the development of the sparse sequential knockoff algorithm, which introduces a preliminary phase to pre-select predictors. This phase relies on the graphical lasso (Friedman et al., 2008) to estimate the precision matrix of the original set of predictors. Zero entries in the precision matrix are used to identify predictors that are conditionally independent of each variable X_j, and these predictors do not enter in the subsequent step of generating $\tilde{X}_j$. Since the graphical lasso is designed for continuous variables, Zimmermann et al. (2024) address the issue of categorical variables by employing a dummy encoding scheme. In this approach, conditional dependence between two categorical variables is assumed if at least one non-zero element appears in the precision matrix of the dummy-encoded representation.

All the methods described so far require complete datasets. To overcome this limitation, Xie et al. (2023) proposed a general setting that extends the derandomized approach of Ren et al. (2023), allowing for the accommodation of both missing values and observed variables of mixed type (continuous, binary, and ordered). Their proposal relies on a restrictive version of the MAR assumption (Little & Rubin, 2002), known as SMAR, according to which the probability of missingness depends only on non-null variables. To accommodate continuous, binary, and ordinal variables in a single model, the approach exploits a latent variable model in which the predictors are assumed to be jointly distributed as an underlying multivariate normal distribution. This model is estimated by maximum likelihood, integrating out missing values via an EM algorithm (Dempster et al., 1977). Currently, the approach does not allow for unordered categorical and count variables or less restrictive assumptions about missingness. This approach, which simultaneously specifies a model for the outcome and integrates out missing data via the EM algorithm, is very efficient when the underlying assumptions hold, but is difficult to extend to other settings.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

2.2. Implementing the knockoff filter with multiple imputation

To extend the range of possible models and enable flexible handling of missing data, we propose a multi-step procedure for variable selection using knockoffs, based on the following steps:

(i) MI of the missing data;
(ii) application of the knockoff filter to each imputed dataset;
(iii) identification of a unified set of selected variables.

Once the variable selection process is concluded, the model of interest can be fitted on the imputed datasets using the selected variables.

The first step of the proposed procedure consists of imputing missing values through MI (Rubin, 1987, 1996), a general and well-established approach to handling missing values under an MAR mechanism. Implementing the MI procedure involves several critical choices. Among these, the selection of the imputation method and the number of imputed datasets are crucial. We adopt MI by chained equations (MICE; van Buuren, 2018; van Buuren & Groothuis-Oudshoorn, 2011), a widely used and flexible approach that accommodates various types of variables and complex relationships among them. MICE is particularly advantageous in settings with both categorical and continuous variables, as it sequentially imputes missing values for each variable conditional on all others, using tailored imputation models. This flexibility makes MICE suitable for INVALSI data, characterized by a set of ordered and unordered categorical variables with varying degrees of missingness.

Another critical decision concerns the number of imputed datasets. A small number of imputed datasets, say 5, is often enough to appropriately account for the uncertainty introduced by the imputation process. However, the optimal number of imputed datasets depends on various factors, including the proportion of missing data, the mechanism underlying the missingness, and the desired precision of estimates (Graham et al., 2007). The search for the optimal number of imputed datasets is beyond the scope of this article. In the simulation studies of Section 3 and Appendix A and in the application of Section 4, we use 10 imputed datasets to get a balance between computational burden and efficiency (Bodner, 2008; White et al., 2011).

After imputing the missing values using MI, variable selection is performed separately for each imputed dataset using the knockoff approach. It is important to note that any knockoff method can be applied. In the simulation study of Section 3, we compare the performance of the proposed MI-based procedure using two types of knockoffs: (i) the knockoffs procedure of Ren et al. (2023) that is the one embedded within the algorithm of Xie et al. (2023), enabling a comparison with our proposed method, under the same conditions and (ii) the sequential knockoffs procedure of Kormaksson et al. (2021) and Zimmermann et al. (2024) that properly handles categorical variables, making it particularly suitable for application to INVALSI data.

The knockoffs-based variable selection is applied to each imputed dataset separately. In this way, the FDR is controlled within each imputed dataset, but there is no guarantee to control the *overall* FDR. Each variable can be selected in all imputed datasets, in none, or in a proportion of them. Thus, the third step involves choosing the *selection proportion*, namely, the minimal proportion for a variable to be ultimately selected. A higher selection proportion reduces the probability that a variable is ultimately selected, thereby decreasing the expected number of false discoveries (i.e., non-null variables wrongly selected) and, thus, the overall FDR. Thus, a conservative choice would be to set the minimal proportion to 1, that is, ultimately select only variables selected in all the imputed datasets. However, this choice could lead to a final FDR much lower than the target one at the cost of a reduction of the ability of the procedure to correctly identify non-null variables, leading to a potentially low true positive rate (TPR):

$$TPR = E\left(\frac{|S_D \cap S^*|}{|S^*|}\right), \tag{6}$$

where $S^* \subseteq \{1, \ldots, p\}$ is the subset of the true non-null predictors.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

In the context of variable selection via stepwise, Wood et al. (2008) suggest the "majority rule," that is, selecting a variable whenever it is chosen in the majority of the imputed datasets. However, their study does not investigate the impact of this strategy on the overall FDR. As there are no theoretical guidelines for determining the optimal selection proportion, in the simulation study presented in the next section, we examine results for different selection proportions.

3. Simulation study: Alternative methods for variable selection with missing values

We perform a Monte Carlo simulation study to investigate the properties of our proposal to embed the knockoffs method into an MI procedure. The main features of the simulation study are reported in Table 1.

We implement the MI-based knockoff approach using both the derandomized knockoff filter proposed by Ren et al. (2023) (called MI-RWC) and the sparse sequential knockoff filter introduced by Zimmermann et al. (2024) (called MI-seq). It is worth noting that the selection procedure is nested in both cases. First, the multiple repetition of the knockoff process (31 runs) selects the variables *within* each of the 10 imputed datasets, adopting a selection threshold equal to 0.5, as suggested in the literature (see, for instance, Ren et al., 2023). Then, these selections are combined across the imputed datasets according to a given selection proportion to obtain a unique set of selected variables. In the following, we explore how the performance of the approach changes with the selection proportion.

The results obtained by these MI-based knockoff methods (i.e., MI-RWC and MI-seq) are compared with those obtained using the approach of Xie et al. (2023), here called XCDW, which is currently

Table 1. Overview of the simulation study

Selection methods	- XCDW: Maximum likelihood with EM, missing values integrated out (Xie et al., 2023) - `R` scripts made available by Xie et al. (2023) - MI-lasso: MICE + grouped lasso (Yuan & Lin, 2006) - `R` packages `mice` (van Buuren & Groothuis-Oudshoorn, 2011) and `grpreg` (Breheny & Huang, 2015) - MI-RWC: MICE + derandomized knockoffs (Ren et al., 2023) - `R` packages `mice` and `derandomKnock` (Ren et al., 2023) - MI-seq: MICE + sparse sequential knockoffs (Zimmermann et al., 2024) - `R` packages `mice` and `knockofftools` (Zimmermann et al., 2024)
Tuning parameters	- PFER = 2, except for MI-lasso that is based on minimum BIC - Multiple runs (except for MI-lasso): 31 runs with selection threshold 0.5
Sample size	1,000 observations
Variables	- 100 (50 binary and 50 continuous) divided into five blocks of 20 variables with the correlation structure of Xie et al. (2023) - Non-null variables: 2 for each block (one binary and one continuous), for a total of 10 (five binary and five continuous)
Missing data mechanism	- SMAR: missingness only depending on non-null observed variables (one binary and one continuous for each individual) - MAR: missingness depending on both non-null and null observed variables (one binary non-null, one continuous non-null, one binary null, and one continuous null, for each individual)
Rate of missing values	0% (complete data), 10%, 25%, 32%, and 45%
Number of imputed datasets	10 (not applicable to XCDW)
Monte Carlo replications	100 runs

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

the only well-established method for handling knockoffs with missing data. To this end, we devise a simulation study closest to that of Xie et al. (2023), starting from their scripts.

To make the three approaches comparable, we set the nominal value of PFER defined in Equation (1), which is the only metric that all three approaches can control. For details on how the algorithms control PFER, see the specific references (Kormaksson et al., 2021; Ren et al., 2023; Xie et al., 2023).

In the simulation study, we also apply a standard selection procedure based on the lasso (without a knockoff filter) to the imputed datasets, denoted as MI-lasso.

The simulated scenario consists of 50 binary variables, 5 of which are non-null, and 50 continuous variables, 5 of which are non-null. The correlation structure is specified as in Xie et al. (2023). In detail, the 100 variables are equally divided into five blocks (each block contains 10 binary variables and 10 continuous variables). The correlation within each block is: 0.60 for the first, second, and third blocks, and 0.30 for the fourth and fifth blocks; the correlation between pairs of blocks ranges from 0.10 to 0.30. For further details, see Xie et al. (2023, Section 5, Figure 4). The response variable is generated by a linear regression model with normal errors and coefficients for non-null variables equal to +0.5 for binary variables and −0.5 for continuous variables.

After response generation, missing values on predictors are generated using the SMAR mechanism (as described in Xie et al., 2023), where missingness depends solely on non-null observed variables associated with the outcome. Specifically, in the simulation scheme of Xie et al. (2023), for each individual i, 2 of the 10 non-null variables, say X_{ij_1} and X_{ij_2}, are assumed to be observed and affect the probability of missing value of the remaining 98 ones, as follows:

$$P(X_{ij} = NA) = \frac{1}{1 + e^{-\left[h + \left(X_{ij_1} + X_{ij_2}\right)/2\right)]}}, \quad j \neq j_1, j_2, \tag{7}$$

where h is a constant that drives the rate of missing values.

Additionally, since MI is valid under the more general MAR assumption, we also consider an MAR scenario by allowing missingness on predictors to depend on two null observed variables, say X_{ij_3} and X_{ij_4}, in addition to the two non-null ones. Formula (7) modifies accordingly:

$$P(X_{ij} = NA) = \frac{1}{1 + e^{-\left[h + \left(X_{ij_1} + X_{ij_2} + X_{ij_3} + X_{ij_4}\right)/4)\right]}}, \quad j \neq j_1, j_2, j_3, j_4. \tag{8}$$

Setting $h = -1$ in formulas (7) and (8) as in Xie et al. (2023), the resulting rate of missing values in the simulations is approximately 32%. In the following, we present the results obtained adopting $h = -2.4, -1.4, -1.0, -0.4$ corresponding to about 10%, 25%, 32%, and 45% of missing values, respectively (Section 3.1). Then, we provide further insights focusing on the scenario with $h = -1$ (Section 3.2).

The simulation results are summarized by the following indices, where "discoveries" stands for "selected variables" and MC for Monte Carlo mean:

$$\widehat{PFER} = MC(\text{number of false discoveries})$$
$$\widehat{FDR} = MC\left(\frac{\text{number of false discoveries}}{\text{number of discoveries}}\right)$$
$$\widehat{TPR} = MC\left(\frac{\text{number of true discoveries}}{\text{number of non-null variables}}\right).$$

The three indices $\widehat{PFER}$, $\widehat{FDR}$, and $\widehat{TPR}$ represent the empirical counterparts of the theoretical ones, defined in Equations (1), (2), and (6).

3.1. Results of Monte Carlo simulations for varying rates of missing values

We start performing a preliminary check for the robustness of the proposed knockoffs methods in the absence of missing values (i.e., with complete data), within the proposed simulation scenario.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 2. Simulation study: variable selection using XCDW, MI-lasso, MI-RWC, and MI-seq, for increasing rates of missing values (about 10%, 25%, 32%, and 45%), by selection proportion (MAR assumption; nominal PFER=2)

Method	$\widehat{PFER}$				$\widehat{FDR}$				$\widehat{TPR}$			
	10%	25%	32%	45%	10%	25%	32%	45%	10%	25%	32%	45%
XCDW	*0.11*	*0.66*	2.51	56.53	0.00	0.06	0.20	0.86	0.91	0.90	0.90	0.93
Selection prop.: ≥ 0.6												
MI-lasso	14.07	19.33	24.37	80.58	0.58	0.66	0.70	0.89	1.00	1.00	0.99	1.00
MI-RWC	2.09	6.14	18.06	79.56	0.16	0.38	0.61	0.89	0.99	0.98	0.98	0.98
MI-seq	2.18	4.13	6.76	12.62	0.17	0.29	0.40	0.55	0.99	0.97	0.95	0.82
Selection prop.: ≥ 0.7												
MI-lasso	11.60	14.89	17.91	70.96	0.53	0.59	0.63	0.87	1.00	0.99	0.99	1.00
MI-RWC	*1.61*	4.43	12.56	68.55	0.13	0.29	0.52	0.88	0.99	0.98	0.97	0.95
MI-seq	*1.65*	2.96	4.58	7.44	0.13	0.22	0.31	0.44	0.99	0.97	0.93	0.77
Selection prop.: ≥ 0.8												
MI-lasso	9.41	10.70	12.81	56.25	0.48	0.51	0.55	0.83	1.00	0.99	0.98	0.99
MI-RWC	*1.19*	2.91	8.40	51.46	0.10	0.21	0.43	0.84	0.98	0.97	0.95	0.90
MI-seq	*1.21*	2.04	2.99	4.04	0.10	0.16	0.23	0.30	0.98	0.95	0.91	0.72
Selection prop.: ≥ 0.9												
MI-lasso	7.03	7.24	8.21	37.39	0.40	0.41	0.44	0.74	1.00	0.99	0.97	0.97
MI-RWC	*0.82*	*1.81*	5.14	31.81	0.07	0.15	0.31	0.77	0.98	0.96	0.93	0.82
MI-seq	*0.84*	*1.18*	*1.67*	*2.00*	0.07	0.10	0.14	0.19	0.98	0.94	0.88	0.65
Selection prop.: 1.0												
MI-lasso	4.91	3.91	4.43	16.38	0.32	0.27	0.30	0.56	1.00	0.98	0.95	0.86
MI-RWC	*0.48*	*0.91*	2.39	12.11	0.04	0.08	0.18	0.60	0.97	0.93	0.89	0.63
MI-seq	*0.47*	*0.52*	*0.70*	*0.59*	0.04	0.05	0.07	0.08	0.97	0.90	0.83	0.53

Note: Monte Carlo means of empirical PFER (in italic if ≤ 2), FDR, and TPR.

Indeed, as widely documented in the literature, we expect that both derandomized knockoffs and sparse sequential knockoffs show a satisfactorily behavior. In this setting and with a nominal PFER set to 2, both derandomized and sequential knockoff filters effectively control the PFER, yielding empirical values of 1.37 and 1.51, respectively (FDR: 0.11 and 0.12; TPR: 0.99 and 1.00).

Then, we investigate the performance of MI-based approaches compared to XCDW and MI-lasso under varying rates of missing data. Table 2 reports the values of empirical PFER, FDR, and TPR under MAR for increasing rates of missing values: 10%, 25%, 32%, and 45%, generated according to Equations (7) and (8). Results under SMAR (not shown for the sake of parsimony) are similar. Note that for MI-based approaches (i.e., MI-lasso, MI-RWC, and MI-seq), performance is evaluated for different selection proportions over the 10 imputed datasets.

The simulation results highlight several key patterns. Overall, as expected, the performance of all compared methods deteriorates as the proportion of missing data increases. When the amount of missing data is low, the XCDW performs exceptionally well, while the two MI-based knockoff approaches yield comparable results.

However, when missing values increase, a clear performance gap emerges between MI-RWC and MI-seq, with the latter showing a noticeable advantage. For instance, at 32% missing values, MI-RWC never controls the PFER, even when the selection proportion equals 1 (empirical PFER settles at 2.39).

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

In contrast, the MI-seq ensures an empirical PFER below 2 (i.e., 1.67) with a selection proportion of 0.9. In this scenario, the difference between XCDW and MI-seq also tends to diminish.

At the highest level of missing data considered (45%), the only method that still manages to control the PFER is the MI-seq—provided a high selection proportion (i.e., ≥ 0.9) is used—although this comes at the cost of a considerable drop in TPR, reaching a value of 0.65. On the other hand, with such high missing value rates, XCDW performs worse than the MI-based knockoff procedures, with an empirical PFER of 56.53.

These findings suggest that the optimal selection proportion for MI-based knockoff procedures is strongly influenced by the extent of missing data: the higher the proportion of missing values, the more stringent the selection rule must be. In general, the commonly recommended majority rule for MI (as suggested, for instance, by Wood et al., 2008) appears insufficient in our setting. Namely, in our simulation framework, the majority rule corresponds to a selection proportion of at least 0.6, which does not adequately control the error rate under medium–high levels of missingness.

3.2. More results on simulations with 32% of missing values

Here, we focus on a specific percentage of missing values in order to further explore the performance of the various approaches under study. Specifically, we refer to simulated data with approximately 32% of missing values ($h = -1$ in Equations (7) and (8)) to facilitate a direct comparison with our benchmark approach (i.e., XCDW). Table 3 reports the values of the empirical PFER, FDR, and TPR and their standard deviations for the considered methods, under SMAR and MAR assumptions.

The XCDW method has an empirical PFER of 2.12 under SMAR and 2.51 under MAR, close enough to the nominal value of 2. On the other hand, for MI-based approaches, the empirical PFER depends on the selection proportion over the imputed datasets, as already discussed in Section 3.1: specifically, for MI-RWC, the empirical PFER gets close to 2 when the selection proportion is 1, whereas for MI-seq, the optimal selection proportion is between 0.8 and 0.9. The MI-lasso yields empirical PFER values much greater than 2, even when the selection proportion is 1. This last result is expected, since the MI-lasso procedure chooses the tuning parameter by minimizing the BIC, and it does not explicitly control the PFER.

Figure 1 displays the box plots based on the quartiles for empirical FDR and empirical TPR, while Figure 2 depicts the trends for an increasing selection proportion in the mean Monte Carlo values for the two indices.

The reference values for evaluating the performance of the MI-based approaches are those achieved by XCDW: an FDR of 0.17 under SMAR and 0.20 under MAR, and a TPR of 0.91 under SMAR and 0.90 under MAR. Notably, although the XCDW approach is valid only under SMAR (i.e., missingness depending only on non-null variables), its performance is minimally affected when applied to MAR data. Across all approaches, switching from SMAR to the more general MAR results in a modest increase in FDR, while TPR remains unchanged.

Focusing on MI-based knockoff approaches, MI-RWC shows a performance comparable to that of XCDW only when variables are retained if selected in all imputed datasets (i.e., selection proportion equal to 1). Under this condition, MI-RWC achieves a Monte Carlo mean FDR of 0.17–0.18 and a mean PFER close to 2, albeit with a slight reduction in TPR (0.88–0.89). In contrast, with a selection proportion of at least 0.9, the FDR rises to approximately 0.30, and the PFER increases to over 4, with both metrics worsening further at smaller selection proportions.

A markedly different scenario emerges when evaluating the performance of MI-seq. Specifically, when variables are retained if selected in at least 0.8 of the imputed datasets, the FDR and TPR are very close to those achieved by XCDW. Increasing the selection proportion further reduces the FDR, but comes at the cost of a slight decrease in TPR, which declines from 0.91 (for a selection proportion ≥ 0.8) to 0.88 (if the selection proportion is ≥ 0.9), reaching a minimum of 0.83 when variables are retained only if selected in all imputed datasets.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 3. Simulation study with 32% missing values: variable selection using XCDW, MI-lasso, MI-RWC, and MI-seq, for increasing selection proportion (nominal PFER = 2)

Method	Missing mechanism	$\widehat{PFER}$		$\widehat{FDR}$		$\widehat{TPR}$	
		Mean	SD	Mean	SD	Mean	SD
XCDW	SMAR	2.12	1.87	0.17	0.13	0.91	0.08
	MAR	2.51	1.80	0.20	0.12	0.90	0.08
Selection prop.: ≥ 0.6							
MI-lasso	SMAR	24.04	6.15	0.70	0.05	1.00	0.02
	MAR	24.37	6.34	0.70	0.05	0.99	0.03
MI-RWC	SMAR	16.06	8.74	0.58	0.13	0.97	0.05
	MAR	18.06	10.22	0.61	0.11	0.98	0.05
MI-seq	SMAR	6.07	2.91	0.37	0.12	0.94	0.07
	MAR	6.76	3.26	0.40	0.11	0.95	0.06
Selection prop.: ≥ 0.7							
MI-lasso	SMAR	17.30	5.22	0.62	0.07	0.99	0.03
	MAR	17.91	4.76	0.63	0.06	0.99	0.03
MI-RWC	SMAR	11.20	7.09	0.49	0.16	0.96	0.05
	MAR	12.56	7.75	0.52	0.13	0.97	0.05
MI-seq	SMAR	3.93	2.35	0.28	0.12	0.92	0.07
	MAR	4.58	2.75	0.31	0.12	0.93	0.07
Selection prop.: ≥ 0.8							
MI-lasso	SMAR	12.23	4.36	0.54	0.09	0.99	0.03
	MAR	12.81	4.04	0.55	0.07	0.98	0.04
MI-RWC	SMAR	7.44	5.17	0.40	0.16	0.95	0.05
	MAR	8.40	5.72	0.43	0.15	0.95	0.06
MI-seq	SMAR	2.41	1.88	0.19	0.12	0.91	0.08
	MAR	2.99	2.28	0.23	0.12	0.91	0.07
Selection prop.: ≥ 0.9							
MI-lasso	SMAR	7.46	3.52	0.41	0.12	0.98	0.04
	MAR	8.21	3.35	0.44	0.10	0.97	0.05
MI-RWC	SMAR	4.35	3.67	0.28	0.16	0.93	0.07
	MAR	5.14	4.15	0.31	0.16	0.93	0.07
MI-seq	SMAR	*1.31*	1.24	0.12	0.10	0.88	0.08
	MAR	*1.67*	1.58	0.14	0.11	0.88	0.08
Selection prop.: 1							
MI-lasso	SMAR	3.87	2.20	0.27	0.12	0.96	0.06
	MAR	4.43	2.39	0.30	0.11	0.95	0.06
MI-RWC	SMAR	2.04	2.00	0.17	0.13	0.88	0.09
	MAR	2.39	2.43	0.18	0.14	0.89	0.08
MI-seq	SMAR	*0.53*	0.69	0.05	0.07	0.83	0.10
	MAR	*0.70*	0.86	0.07	0.08	0.83	0.10

Note: Monte Carlo means and standard deviations of empirical PFER (in italic if ≤ 2), FDR, and TPR.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

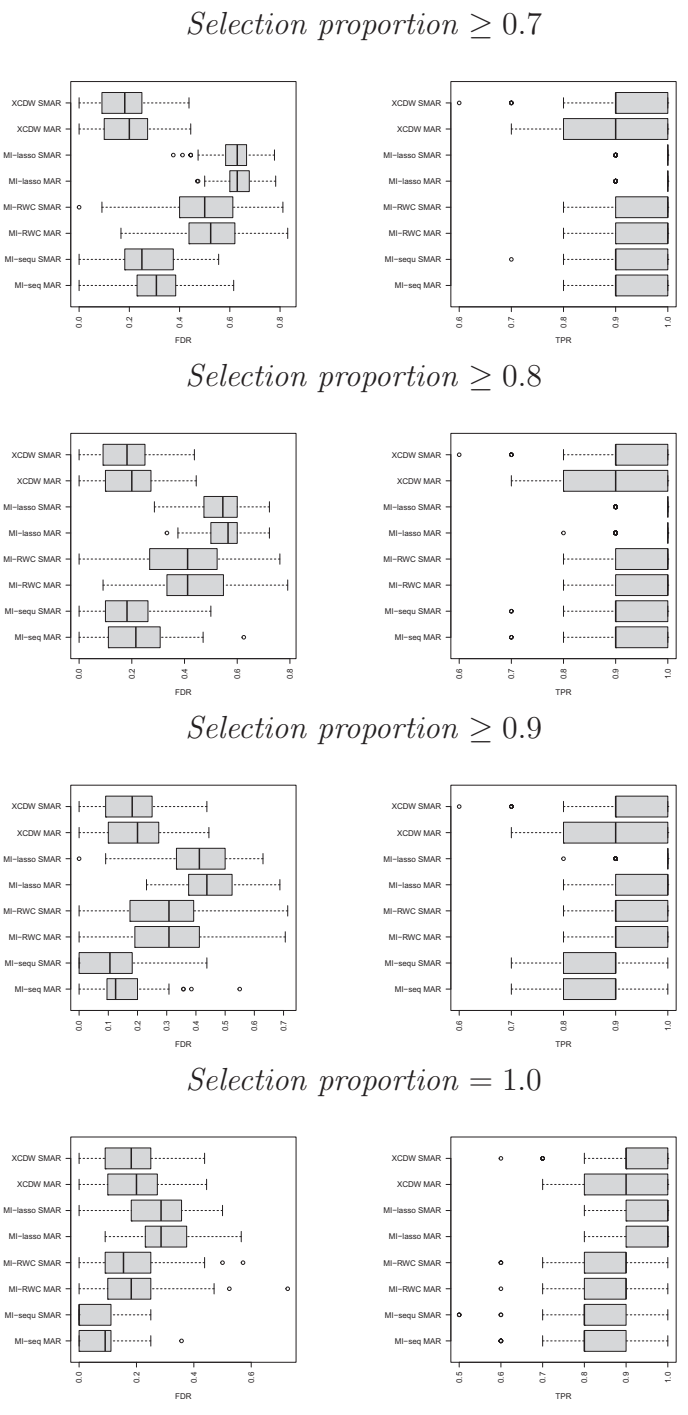

Figure 1. Distribution of empirical FDR (left) and empirical TPR (right) under XCDW, MI-lasso, MI-RWC, and MI-seq, for an increasing selection proportion (up to bottom panels: ≥ 0.7, ≥ 0.8, ≥ 0.9, and 1): box plots based on Monte Carlo values for minimum, 1st quartile, median, 3rd quartile, and maximum.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

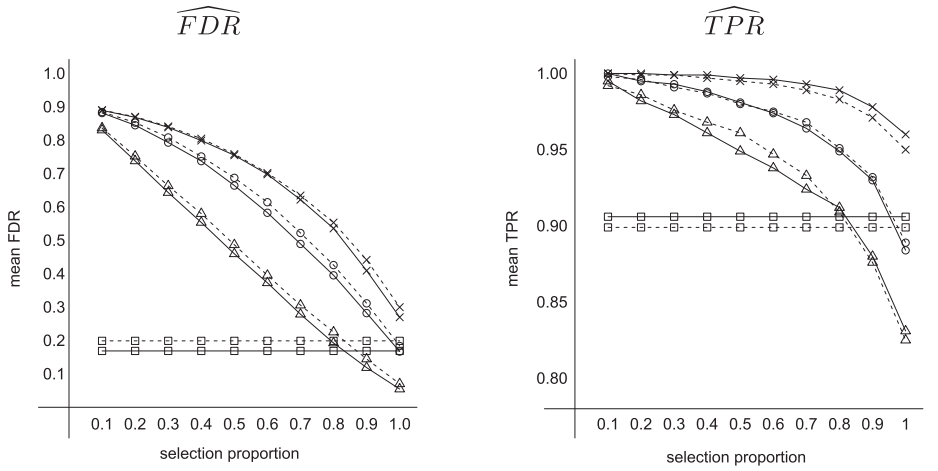

Figure 2. Variable selection using XCDW, MI-lasso, MI-RWC, and MI-seq, for an increasing selection proportion for ultimately selecting the variables over the imputed datasets: Monte Carlo means of FDR (left) and TPR (right). Dashed lines for MAR and solid lines for SMAR: □ XCDW, × MI-lasso, ○ MI-RWC, and △ MI-seq.

3.3. Remarks on simulation results

In summary, the MI-seq approach represents a valuable alternative to the XCDW method. Based on the simulation results described above, there is evidence that the selection proportion depends on the rate of missing values. For a low missing rate (approximately 10%), retaining a variable if selected in at least 0.7 of the imputed datasets is sufficient to control the PFER while maintaining high TPR values. When the missing rate increases to 25%–32%, this selection proportion rises to 0.9 (0.8 is almost sufficient with 25% of missingness), while the empirical TPR remains high (0.94 with 25% of missingness and 0.88 with 32% of missingness). For very high missing rates (around 45%), the general recommendation is to retain variables selected in at least 0.9 of the imputed datasets; however, this comes at the cost of a substantial reduction in TPR.

It is worth noting that the simulation design used here is the one proposed by Xie et al. (2023) to validate their method. In particular, the continuous variables are assumed to be normally distributed, which is important for the XCDW method based on likelihood. Indeed, deviations from normality could negatively affect the XCDW approach. On the other hand, the normality assumption is irrelevant for the MI-based approaches, since the imputation phase for continuous variables exploits least-squares linear models. Moreover, the simulation study considers only binary and continuous variables, neglecting ordered and unordered categorical variables. Ordinal variables are admitted in XCDW, even if the selection procedure should be adapted to properly account for the ordering of categories. On the other hand, unordered categorical variables are not handled by XCDW, which is a limitation in some applications, like the INVALSI case on math test scores presented in the next section.

In order to evaluate the performance of the proposed MI-seq procedure in the INVALSI setting, which is quite different in terms of variables and missingness pattern, we conduct a further simulation study tailored to the INVALSI case. The specific simulation design and main results are illustrated in Appendix 1, showing that the MI-seq method also performs well in this setting.

4. Variable selection via knockoffs for INVALSI data on mathematics test scores

The INVALSI is the national institute that evaluates the Italian school system. Specifically, it evaluates student achievement in Mathematics, Italian, and English for different grades each year. The evaluation is based on multiple-choice questions. The selection of the set of items relies on internationally validated methods based on the Rasch model (Rasch, 1960), similarly to TIMSS (von Davier et al., 2024). The

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

evaluation test is administered to all students across all schools to assess school performance. An additional background questionnaire is administered to students in a sample of schools to examine the relationship between achievement and students' background characteristics. We focus on data from the INVALSI sample survey for 2022–2023 on pupils in grade 5, corresponding to the last year of primary school. The dataset includes 16,828 students nested in 501 schools.

Specifically, we consider achievement in mathematics, a latent ability measured by a set of dichotomously scored items (1: correct answer; 0: wrong answer). INVALSI provides the item response patterns and a Rasch-based estimate of ability. The average sample score in mathematics is 191.82 points, with a standard deviation of 41.07.

The dataset includes 30 student-level variables of different types:

- binary (e.g., gender and availability of a personal computer);
- ordinal (e.g., Mother's education);
- unordered categorical (e.g., Mother's occupation).

The dataset also includes the school's geographical area, which is not included in the variable selection process because it is used as a stratification variable.

To study the relationship between background characteristics and pupils' achievement while accounting for clustering within schools, we specify a random intercept model (Snijders & Bosker, 2011). Denoting the pupils with i (level 1) and the schools with s (level 2), the model is specified as

$$y_{is} = \mathbf{x}_{is}'\boldsymbol{\beta} + \mathbf{z}_s'\boldsymbol{\gamma} + u_s + e_{is}, \tag{9}$$

where y_{is} is the mathematics score of pupil i of school s, $\mathbf{x}_{is}$ is the vector of level 1 predictors (including the constant) with coefficients $\boldsymbol{\beta}$ (including the intercept), and $\mathbf{z}_s$ is the vector of level 2 predictors with coefficients $\boldsymbol{\gamma}$. The level 1 errors are assumed to be independent and identically distributed, $e_{is} \sim N(0, \sigma_e^2)$, and the level 2 errors (random effects) are assumed to be independent and identically distributed, $u_s \sim N(0, \sigma_u^2)$. Moreover, the errors at the two levels are assumed to be independent.

The background variables are highly correlated, and there are no theoretical guidelines for choosing among them, which calls for a data-driven selection procedure.

Moreover, a critical issue of the data is the relevant rate of missing values for many student-level variables, ranging from 1.6% to 25.5%. Specifically, the variables with the highest rates of missing values are Father's and Mother's education and occupation (ranging from 23.4% to 25.5%), followed by Italian language at home, Dialect spoken at home, Italian used with friends, and Nursery school (about 10%). A complete-case analysis would halve the number of students, leading to a large loss of power and, especially, potentially high bias, given that the missing mechanism cannot be assumed to be missing completely at random. Rather, an MAR mechanism with missingness depending on observed values may be plausible, so MI is a viable approach to preserve the original sample size and avoid bias in the estimators.

To perform variable selection while handling missing data, we follow the proposed procedure described in Section 2.2, namely, MI-seq. It is worth noting that the other procedures (i.e., XCDW and MI-RWC) are not suitable, as they do not handle variables with unordered categories, such as Immigrant status, Father's occupation, and Mother's occupation.

The MI-seq is applied to the INVALSI case study, implementing MICE with 10 imputed datasets using the `mice` package of `R` (van Buuren & Groothuis-Oudshoorn, 2011). Each variable is imputed with a model suitable for its measurement scale: logit model for binary variables, cumulative logit model for ordinal variables, and multinomial logit model for unordered categorical variables.

For each imputed dataset, the predictors are selected using the sequential knockoffs procedure of Kormaksson et al. (2021), as implemented in the `R` package `knockofftools` (Zimmermann et al., 2024) with options `PFER = 2` and `method= "sparseseq"`. Unordered categorical and ordinal variables are coded using a set of dummy variables with a baseline category. In the knockoffs filter, we use group lasso (Yuan & Lin, 2006) instead of standard lasso to prevent the selection from depending

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

on the arbitrary choice of the baseline category. In this way, the corresponding set of dummy variables is jointly selected or discarded for each predictor. The procedure runs 31 times for each imputed dataset, and a predictor is retained if it is selected in at least 50% of the runs.

Table 4 shows the predictors with the proportion selected over the 10 imputed datasets by the sequential knockoffs procedure. A notable feature of the results is the clear separation in selection proportions: predictors are either selected in nearly all imputed datasets (selection proportion 0.9 or 1) or rarely selected (with proportions less than or equal to 0.3). This pattern indicates a high degree of stability in the selection results across the imputation process, suggesting that the contributions of most predictors are clearly identified despite missing data. Therefore, the choice of which variables to retain does not appear particularly problematic in this situation. Predictors included in the final model are denoted by a check mark in Table 4. Overall, 22 out of 30 variables have been selected.

The sequential knockoff procedure gives parameter estimates based on the group lasso for a standard linear model, that is, without random effects. This is not a limitation for selecting the predictors, since in the linear model, there is an equivalence between the regression coefficients conditional on the random effects and the marginal ones (Ritz & Spiegelman, 2004). Notwithstanding, we need to fit a random effect model with the selected variables for two reasons: (i) to obtain standard errors properly accounting for the school clustering and (ii) to avoid the downward bias implicit in lasso-based procedures (e.g., Belloni & Chernozhukov, 2013). For these reasons, we fit the random intercept model (9) with the selected predictors to each imputed dataset via maximum likelihood using the `R` package `lme4` (Bates et al., 2015). The results across the 10 imputed datasets are combined using Rubin's rules (Rubin, 1987) to obtain point estimates and standard errors.

Table 5 reports the residual variances and the intraclass correlation coefficient (ICC) for the model without predictors (null model) and for the model with the predictors selected by sequential knockoffs (full model; see Table 4). The null model shows that the variability in the mathematics scores due to school clustering is 13.8%. The predictors explain about 23% of the school-level variance and about 15% of the pupil-level variance. Given the included predictors, the residual ICC in the final model is 12.6%, indicating that school clustering remains relevant.

Table 6 presents the point estimates and corresponding standard errors for the full model, obtained through Rubin's rules. It is worth noting that the standard errors account for the variability due to MI; however, they ignore the uncertainty implicit in the variable selection process through knockoffs and, thus, are likely to be underestimated, raising a problem for post-selection inference. Although formal inference is therefore limited, the selected variables and their estimated effects still provide valuable insights into the relationship between students' background characteristics and achievement. In particular, the direction (i.e., the sign) and relative magnitude of the regression coefficients allow for a meaningful interpretation of the main determinants of performance. This is in line with previous findings showing that knockoff-based procedures are effective in identifying the direction of the effects (Barber & Candès, 2019) and is further supported in our data setting by the simulation results reported in Appendix 1. In the following, we therefore focus on the interpretation of the point estimates.

In general, the estimated effects are consistent with expectations and with the existing literature on educational achievement. For example, the gender gap in favor of males is confirmed (Contini et al., 2017).

In analyses of large-scale assessment data, families' cultural background and socioeconomic status (SES) are typically associated with an advantage in educational achievement. Our results confirm this relationship. In particular, the largest effect pertains to the number of books at home, a well-established proxy of the cultural environment (Heppt et al., 2022). As for SES, it is often measured through an indicator that summarizes parental education and occupation (e.g., Cascella, 2025). Here, we keep these components separate in order to estimate their specific effects. It turns out that the effects of parental education are similar for mothers and fathers and increase monotonically from compulsory schooling to a master's degree or higher, in line with a large body of literature documenting the role of family

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 4. Variable selection via MI-seq: proportion selected over 10 imputed datasets and retained status (✓ if proportion ≥ 0.8)

Variable	*Levels*	*Prop.*	*Retained*
Male	1=yes, 0=no	1	✓
No. of brothers	ordinal, 5 levels	1	✓
No. of sisters	ordinal, 5 levels	1	✓
At home, you have:			
quiet space	1=yes, 0=no	0.9	✓
computer	1=yes, 0=no	1	✓
personal desk	1=yes, 0=no	0.3	
software	1=yes, 0=no	0	
internet connection	1=yes, 0=no	1	✓
personal room	1=yes, 0=no	1	✓
classical books	1=yes, 0=no	0	
artworks	1=yes, 0=no	0	
technical manuals	1=yes, 0=no	0	
dictionary	1=yes, 0=no	1	✓
personal tablet	1=yes, 0=no	1	✓
smartphone	1=yes, 0=no	1	✓
No. of books	ordinal, 5 levels	1	✓
Student origin and language			
Student born in Italy	1=yes, 0=no	1	✓
Immigrant status	1=native; 2=1st gen immigrant; 3=2nd gen imm.	0	
Italian language at home	1=yes, 0=no	1	✓
Dialect spoken at home	1=yes, 0=no	1	✓
Italian used with friends	1=yes, 0=no	1	✓
Parents			
Father born in Italy	1=yes, 0=no	0.2	
Mother born in Italy	1=yes, 0=no	1	✓
Father's education	1=compulsory, 2=high sch., 3=college, 4=master+	1	✓
Mother's education	1=compulsory, 2=high sch., 3=college, 4=master+	1	✓
Father's occupation	1=unenmployed 2=manager, clerk 3=self-employed 4=blue collar 5=professional 6=housewife, retired	1	✓
Mother's occupation	1=unenmployed 2=manager, clerk 3=self-employed 4=blue collar 5=professional 6=housewife, retired	1	✓
Student career			
Nursery school	(for pupils aged 3–5 years) 1=yes, 0=no	1	✓
Regular student	1=yes, 0=no	1	✓
School weekly hours >30	1=yes, 0=no	0	
School area	1=NW; 2=NE; 3=Center; 4= South; 5= S-Islands	*retained by default*	

Note: INVALSI grade 5 sample survey, year 2022–2023.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 5. Random intercept null model and full model with selected predictors (Table 4): combined estimates of the residual variances based on 10 imputed datasets

Residual variances	Null model	Full model	% Explained
Level 2 (school)	233.30	178.94	23.30
Level 1 (pupil)	1,461.94	1,239.97	15.18
Total	1,695.24	1,418.91	16.30
ICC	13.76%	12.61%	

Note: INVALSI grade 5 sample survey, year 2022–2023.

background in shaping educational outcomes. On the other hand, the effect of the Father's occupation appears stronger than that of the mother, which is not statistically significant.

Other findings provide interesting additional insights. In particular, personal tablet or smartphone ownership is associated with lower mathematics scores. Since the test evaluates mathematics proficiency acquired during primary school, this result may reflect substitution effects between screen time and study-related activities or, more broadly, differences in the family environment (Beland & Murphy, 2016). While causal interpretations are beyond the scope of this analysis, this finding highlights potentially relevant learning mechanisms that deserve further investigation.

Finally, it is worth noting that the selected variables can be used to predict a student's response. Indeed, the results of the data-driven simulation study presented in Appendix 1 show that models based on variables selected by the MI-seq procedure achieve predictive performance essentially indistinguishable from that obtained using the true underlying predictors. This evidence suggests that the selection procedure is not only effective at identifying relevant variables but also preserves the model's predictive capability in realistic scenarios.

5. Final remarks

In this contribution, we analyzed the effect of student characteristics on mathematics ability in the Italian primary school. These data call for a variable selection method that handles missing values. To this aim, we proposed an approach, called MI-seq, based on knockoffs. The proposed method consists of the following sequential steps: first, missing values are addressed through MI techniques; second, for each imputed dataset, variables are selected using a knockoffs method, and only those variables selected in at least a predefined minimal proportion of datasets are retained. This multi-step structure provides substantial flexibility, distinguishing the MI-seq from the recent contribution by Xie et al. (2023), called here XCDW, which, to the best of our knowledge, is the only one that applies the knockoff method in the presence of missing data. Unlike XCDW, MI-seq does not require assumptions about the distributions of continuous variables and can handle unordered categorical variables alongside continuous, binary, and ordinal variables. Moreover, MI-seq yields results valid using the standard MAR assumption instead of the more restrictive SMAR assumption of XCWD, where the missing data mechanism depends only on non-null variables. Finally, separating the imputation and modeling steps makes it straightforward to specify any kind of model, whereas, in XCWD, using a different model would require a non-trivial modification of the likelihood function and the EM algorithm.

These advantages of MI-seq are made possible because of the following features: (i) the use of MI to deal with missing values, allowing to deal with any variables under the MAR assumption and (ii) the selection of variables based on the sequential knockoffs procedure of Kormaksson et al. (2021) and Zimmermann et al. (2024) that allows to generate knockoff copies of unordered categorical variables based on the multinomial logit model, differently from standard knockoffs procedures (see, e.g., Ren et al., 2023), which are designed for continuous variables.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 6. Random intercept model with selected predictors: combined estimates of the regression coefficients and standard errors based on 10 imputed datasets

Variable	Estimate	Std. error
Intercept	124.26	4.016
Male	9.55	0.568
No. of brothers (ref. 0)		
1	−4.33	0.652
2	−6.98	1.082
3	−11.92	1.961
≥ 4	−12.70	2.929
No. of sisters (ref. 0)		
1	−3.00	0.645
2	−5.38	1.109
3	−9.55	2.472
≥ 4	−11.28	3.776
At home, you have:		
quiet space	1.20	0.832
computer	3.39	0.612
internet connection	4.49	0.862
personal room	−2.58	0.613
dictionary	8.47	1.003
personal tablet	−4.11	0.586
smartphone	−4.29	0.661
No. of books (ref. 0–10)		
11–25	3.96	0.933
26–100	12.34	0.966
101–200	15.39	1.126
>200	17.33	1.262
Student origin and language:		
Student born in Italy	6.31	1.673
Italian language at home	4.56	1.071
Dialect spoken at home	−1.91	0.617
Italian used with friends	9.87	1.540
Mother born in Italy	−1.69	0.975
Father's education (ref. compulsory)		
high school	5.31	0.921
college	8.72	1.541
master or more	10.49	1.333

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table 6. Continued

Variable	Estimate	Std. error
Mother's education (ref. compulsory)		
high school	5.54	0.885
college	6.75	1.342
master or more	10.36	1.316
Father's occupation (ref. unempl.)		
manager, clerk	7.11	1.857
self-employed	5.60	1.858
blue collar	2.90	1.828
professional	5.58	1.922
housewife, retired	2.68	3.546
Mother's occupation (ref. unemployed)		
manager, clerk	2.11	1.483
self-employed	1.32	1.659
blue collar	−0.76	1.522
professional	2.80	1.637
housewife, retired	−0.62	1.495
Student career:		
Nursery school	8.90	1.520
Regular student	5.18	2.287
School area (ref. North-West)		
North-East	2.96	2.147
Center	5.59	2.161
South	8.65	2.089
South-Islands	−1.39	2.151

Note: INVALSI grade 5 sample survey, year 2022–2023.

To evaluate the performance of the MI-seq, we conducted a simulation study under the scenario proposed by Xie et al. (2023), obtaining similar performance for a rate of missing values around 32% and better performance for a rate of 45% of missingness. Interestingly, MI-seq performed better than the approach that integrates MI with the derandomized knockoff procedure of Ren et al. (2023) (referred to in the text as MI-RWC), likely because half of the variables are binary in the simulation. We checked the performance of the proposed method in the context of the INVALSI case study, characterized by several unordered categorical variables affected by missing values, by means of an additional data-driven simulation study illustrated in Appendix 1. This simulation proved the validity of the proposed MI-seq method in the INVALSI scenario.

A limitation of the proposed MI-seq approach is that it does not strictly control the PFER or the FDR, contrary to knockoff filters for complete data and the XCDW method of Xie et al. (2023) for SMAR data. In fact, any MI-based procedure requires choosing the minimal proportion of imputed datasets for ultimately retaining a variable, and the PFER is inversely related to this proportion. Furthermore, the simulation study shows that the minimal selection proportion required to control the PFER increases with the rate of missing values. The "majority rule" is generally inadequate even with a low missingness

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

rate (10%), where a selection proportion of at least 0.7 is preferable. Remarkably, for higher levels of missingness, the optimal selection proportion approaches 0.9. Anyway, it is worth noting that the choice of this proportion is not always critical, as outlined by the data-driven simulation study reported in Appendix 1, since, in applications, it may happen that the selection is not very sensitive to this parameter. Remarkably, in our case study, the non-null variables were clearly identified since the actual proportions were either ≥ 0.9 or ≤ 0.3. As a general strategy, we recommend monitoring the set of retained variables as the selection proportion decreases from 1 to 0.5: if the set is relatively stable, the selection is trustworthy; otherwise, it is advisable to set up a simulation study tailored on the empirical setting, like the one presented in Appendix 1.

The implementation of MI-based methods via MICE may raise computational issues in high-dimensional data. Possible solutions include dividing variables into blocks or using regularization techniques in the imputation phase (van Buuren, 2018).

A further challenging issue concerns the post-selection inference in the presence of missing data. Indeed, the standard errors obtained from fitting the model with selected predictors are underestimated because they ignore the uncertainty introduced by the selection process. A general solution to this issue is data splitting, which involves randomly partitioning the data into two sets, one for variable selection and the other for model fitting and inference (see, for instance, García Rasines & Young, 2023; Wasserman & Roeder, 2009). In case of missing values, MI should be run on each of the two sets separately to avoid generating dependence between them.

Finally, a general issue with knockoffs, unrelated to missing data, concerns their performance in selecting cluster-level predictors in multilevel analysis. In fact, the available methods to generate the knockoffs do not account for the levels of the predictors, so a copy of a cluster-level variable is produced as if it were an individual-level variable, ignoring the constraint that it has to be constant within the clusters. This limitation does not apply in the INVALSI case study because all predictors are at the student level (except for the geographical area, which, however, is not included in the selection procedure). However, applications of multilevel models often involve many cluster-level predictors; thus, it is essential to investigate the extent to which the selection process for cluster-level predictors is negatively affected by the use of knockoffs designed for individual-level predictors and to devise effective solutions.

Supplementary material. The supplementary material for this article can be found at https://doi.org/10.1017/psy.2026.10109.

Data availability statement. The data used in the article are the property of INVALSI. In the Supplementary Material, we provide the R codes to replicate the simulation study described in Section 3 (under MAR and using the sparse sequential knockoffs method) and to select variables on INVALSI-type simulated data.

Acknowledgments. We are grateful to the National Institute for the Evaluation of the Italian School System (INVALSI), which provided the data for the case study reported in the article. INVALSI is not responsible for the analyses and data processing reported in the article. This manuscript is part of the special section, Integrating and Analyzing Complex High-Dimensional Data in Social and Behavioral Sciences Research. We extend our heartfelt gratitude to the co-Guest Editors, Drs Katrijn Van Deun and Eric Lock, as well as the reviewers for their invaluable and insightful feedback, which significantly enhanced this article.

Author contributions. All authors contributed equally to this work.

Funding statement. This work was supported by the Italian Ministry of University and Research (MUR) under grant Department of Excellence project 2023–2027 ReDS "Rethinking Data Science"—Department of Statistics, Computer Science, Applications—University of Florence (Cup B13C22004500001), and Italian Ministry of University and Research (MUR) under grant PRIN 2022 Call (D.D. 104/2022), project "Latent variable models and dimensionality reduction methods for complex data" (Project No. 20224CRB9E, CUP B53C24006320006).

Competing interests. The authors declare none.

References

Barber, R. F., & Candès, E. J. (2015). Controlling the false discovery rate via knockoffs. *The Annals of Statistics*, *43*(5), 2055–2085. https://doi.org/10.1214/18-AOS1755

Barber, R. F., & Candès, E. J. (2019). A knockoff filter for high-dimensional selective inference. *The Annals of Statistics*, *47*(5), 2504–2537. https://doi.org/10.1214/18-AOS1755

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Beland, L.-P., & Murphy, R. (2016). Ill communication: Technology, distraction & student performance. *Labour Economics*, *41*, 61–76. https://doi.org/10.1016/j.labeco.2016.04.004

Belloni, A., & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*, 521–547. https://doi.org/10.2139/ssrn.1582594

Bodner, T. E. (2008). What improves with increased missing data imputations? *Structural Equation Modeling: A Multidisciplinary Journal*, *15*(4), 651–675. https://doi.org/10.1080/10705510802339072

Breheny, P., & Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, *25*(2), 173–187. https://doi.org/10.1007/s11222-013-9424-2

Candès, E., Fan, Y., Janson, L., & Lv, J. (2018). Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *Journal of the Royal Statistical Society, Series B*, *80*, 551–577. https://doi.org/10.1111/rssb.12265

Cascella, C. (2025). Measuring the effect of the sociocultural background on learning outcomes by a simplified and effective index. *International Journal of Social Research Methodology*, *28*(5), 557–572. https://doi.org/10.1080/13645579.2024.2432270

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Contini, D., Di Tommaso, A. L., & Mendolia, S. (2017). The gender gap in mathematics achievement: Evidence from Italian data. *Economics of Education Review*, *58*, 32–42. https://doi.org/10.1016/j.econedurev.2017.03.001

Dai, R., & Barber, R. (2016). The knockoff filter for FDR control in group-sparse and multitask regression. In M. F. Balcan, & K. Q. Weinberger (Eds.), *Proceedings of the 33rd international conference on machine learning* (Vol. *48*, pp. 1851–1859). PMLR.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *Journal of the Royal Statistical Society, Series B*, *39*, 1–38. https://doi.org/10.1111/j.2517-6161.1977.tb01600.x

Friedman, J., Hastie, T., & Tibshirani, R. (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*(3), 432–441. https://doi.org/10.1093/biostatistics/kxm045

García Rasines, D., & Young, G. A. (2023). Splitting strategies for post-selection inference. *Biometrika*, *110*(3), 597–614. https://doi.org/10.1093/biomet/asac070

Graham, J. W., Olchowski, A. E., & Gilreath, T. (2007). How many imputations are really needed? Some practical clarifications of multiple imputation theory. *Prevention Science*, *8*(3), 206–213. https://doi.org/10.1007/s11121-007-0070-9

Grilli, L., Marino, F., Paccagnella, O., & Rampichini, C. (2022). Multiple imputation and selection of ordinal level 2 predictors in multilevel models: An analysis of the relationship between student ratings and teacher practices and attitudes. *Statistical Modelling*, *22*(3), 221–238. https://doi.org/10.1177/1471082X20949710

Heppt, B., Olczyk, M., & Volodina, A. (2022). Number of books at home as an indicator of socioeconomic status: Examining its extensions and their incremental validity for academic achievement. *Social Psychology of Education*, *25*, 903–928. https://doi.org/10.1007/s11218-022-09704-8

Huang, Y., Tibbe, T., Tang, A., & Montoya, A. (2024). Lasso and group lasso with categorical predictors: Impact of coding strategy on variable selection and prediction. *Journal of Behavioral Data Science*, *3*(2), 15–42. https://doi.org/10.35566/jbds/v3n2/montoya

Katsevich, E., & Sabatti, C. (2019). Multilayer knockoff filter: Controlled variable selection at multiple resolutions. *The Annals of Applied Statistics*, *13*(1), 1–33. https://doi.org/10.1214/18-AOAS1185

Kormaksson, M., Kelly, L. J., Zhu, X., Haemmerle, S., Pricop, L., & Ohlssen, D. (2021). Sequential knockoffs for continuous and categorical predictors: With application to a large psoriatic arthritis clinical trial pool. *Statistics in Medicine*, *40*, 3313–3328. https://doi.org/10.1002/sim.8955

Little, R. J. A., & Rubin, D. B. (2002). *Statistical analysis with missing data*. (2nd ed.) Wiley.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Danish Institute for Educational Research.

Ren, Z., Wei, Y., & Candès, E. (2023). Derandomizing knockoffs. *Journal of the American Statistical Association*, *118*, 948–958. https://doi.org/10.1080/01621459.2021.1962720

Ritz, J., & Spiegelman, D. (2004). Equivalence of conditional and marginal regression models for clustered and longitudinal data. *Statistical Methods in Medical Research*, *13*(4), 309–323. https://doi.org/10.1191/0962280204sm368ra

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. Wiley.

Rubin, D. B. (1996). Multiple imputation after 18+ years. *Journal of the American Statistical Association*, *91*(434), 473–489. https://doi.org/10.1080/01621459.1996.10476908

Snijders, T., & Bosker, R. (2011). *Multilevel analysis: An introduction to basic and advanced multilevel modeling*. (2nd ed.) Sage.

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, *58*, 267–288. https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

van Buuren, S. (2018). *Flexible imputation of missing data*. CRC Press.

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. *Journal of Statistical Software*, *45*(3), 1–67. https://doi.org/10.18637/jss.v045.i03

von Davier, M., Fishbein, B., & Kennedy, A. E. (2024). *TIMSS 2023 technical report (methods and procedures)*. TIMSS & PIRLS International Study Center.

Wasserman, L., & Roeder, K. (2009). High-dimensional variable selection. *The Annals of Statistics*, *37*(5A), 2178–2201. https://doi.org/10.1214/08-AOS646

White, I., Royston, P., & Wood, A. M. (2011). Multiple imputation using chained equations: Issues and guidance for practice. *Statistics in Medicine*, *30*(4), 377–399. https://doi.org/10.1002/sim.4067

Wood, A. M., White, I. R., & Royston, P. (2008). How should variable selection be performed with multiply imputed data? *Statistics in Medicine*, *27*(17), 3227–3246. https://doi.org/10.1002/sim.3177

Xie, Z., Chen, Y., von Davier, M., & Weng, H. (2023). Variable selection in latent variable models via knockoffs: An application to international large-scale assessment in education. *Journal of the Royal Statistical Society, Series A*, 1–25. https://doi.org/10.1093/jrsssa/qnad137

Yuan, M., & Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *68*(1), 49–67. https://doi.org/10.1111/j.1467-9868.2005.00532.x

Zimmermann, M. R., Baillie, M., Kormaksson, M., Ohlssen, D., & Sechidis, K. (2024). All that glitters is not gold: Type-I error controlled variable selection from clinical trial data. *Clinical Pharmacology & Therapeutics*, *115*(4), 774–785. https://doi.org/10.1002/cpt.3211

Appendix A. Data-driven simulation study based on the INVALSI data structure

To investigate the appropriateness of the MI-seq approach for addressing variable selection in the INVALSI setting, characterized by unordered categorical predictors and missing data, we present an additional simulation study specifically designed to closely mimic the INVALSI application. This simulation employs a data-driven strategy, where the joint distribution of the predictors and the missingness pattern are directly inherited from the observed INVALSI data.

Let $\mathbf{X}$ denote the covariate matrix from the INVALSI dataset, which includes binary, ordered, and unordered categorical predictors with multiple levels, as well as non-negligible amounts of missing data (see Table 4 for the list of variables). To sample the response variable according to a known data-generating model, we proceed as follows:

1. The missing values of the matrix $\mathbf{X}$ are imputed once using MICE to get a complete matrix $\mathbf{X}_{imp}$.
2. To obtain realistic effect sizes, the observed INVALSI response vector $\mathbf{y}$ is regressed on the imputed matrix $\mathbf{X}_{imp}$ using lasso (with grouped lasso penalty for unordered categorical predictors).
3. The vector of the lasso estimates of the regression coefficients, denoted by $\boldsymbol{\beta}_{true}$, defines the data-generating model and is used to generate the new response vector following the linear regression model $\mathbf{y}_{gen} = \mathbf{X}_{imp}\boldsymbol{\beta}_{true} + \boldsymbol{\varepsilon}$, where the errors are independently sampled from a normal distribution with mean 0 and standard deviation set on the basis of the lasso residuals.

The dataset used for the simulation consists of the generated response values $\mathbf{y}_{gen}$ and the original covariate matrix $\mathbf{X}$ (i.e., the original observed data with missing values). From this dataset, $R = 50$ random subsamples of half of the 16,828 observations are drawn without replacement and used as a training set, while the remaining observations form a corresponding test set. This sampling scheme substantially reduces computational costs while enabling evaluation of variable selection performance, the correctness of regression coefficients' signs, and the out-of-sample prediction performance.

For each training set, the MI-seq procedure is applied as described in Section 2.2, that is, imputing missing covariates with MICE and selecting predictors using the sparse sequential knockoff approach. Variable selection performance is evaluated across the replications by the empirical PFER, FDR, and TPR. Table A1 reports the distributions of $\widehat{PFER}$, $\widehat{FDR}$, and $\widehat{TPR}$ for different selection proportions.

Table A1 shows that the MI-seq approach works well in the proposed setting. In fact, the PFER is controlled, whereas the TPR (proportion of truly non-null variables selected) is satisfactory. Furthermore, in the particular setting at issue, the results are not influenced by the selection proportion.

To further assess the practical relevance of MI combined with knockoffs, we also evaluate the approach's ability to correctly identify the sign of the predictors' effects. To this end, the set of variables selected on the training data is used to fit the model on the test data.

Each test set is multiply imputed using the MICE procedure, generating 10 complete versions of the test data.

For each imputed test set, a linear regression model is fitted using the variables selected in the corresponding training set, yielding a set of estimated coefficient vectors. We set the selection proportion to 0.7, though this choice has little consequence

https://doi.org/10.1017/psy.2026.10109 Published online by Cambridge University Press

Table A1. Data-driven simulation study—Variable selection with MI-seq, for increasing selection proportion (nominal PFER = 2): Monte Carlo means and standard deviations of empirical PFER, FDR, and TPR

Selection proportion	$\widehat{PFER}$		$\widehat{FDR}$		$\widehat{TPR}$	
	Mean	SD	Mean	SD	Mean	SD
≥ 0.1	0.618	0.850	0.027	0.035	0.858	0.055
≥ 0.2	0.600	0.784	0.027	0.034	0.856	0.052
≥ 0.3	0.600	0.784	0.027	0.034	0.856	0.052
≥ 0.4	0.582	0.738	0.026	0.032	0.856	0.052
≥ 0.5	0.582	0.738	0.026	0.033	0.855	0.051
≥ 0.6	0.582	0.738	0.026	0.033	0.855	0.051
≥ 0.7	0.582	0.738	0.026	0.033	0.855	0.051
≥ 0.8	0.582	0.738	0.026	0.033	0.855	0.051
≥ 0.9	0.582	0.738	0.026	0.033	0.855	0.051
1.0	0.582	0.738	0.026	0.033	0.855	0.051

in this application. These coefficient estimates are then combined using Rubin's rules, resulting in a single pooled coefficient vector for each test set. By repeating this procedure on all $R = 50$ test sets, we obtain 50 estimates of the regression coefficients to compare with the true coefficient vector $\boldsymbol{\beta}_{true}$.

To evaluate the ability of the selection method to correctly identify the sign of the regression coefficients, we classify the coefficients as negative, null, and positive, then compute the agreement between true and estimated coefficients using Cohen's Kappa (Cohen, 1960). We obtain satisfactory results, as Cohen's Kappa coefficient lies in the range $[0.677, 0.939]$, with an average value of 0.808 (median of 0.820).

Furthermore, even if the primary aim of the knockoff method is selecting non-null variables, it is of interest to assess the predictive ability of the model based on the selected variables. For this aim, we first select the variables and estimate the regression coefficients on each training set; the resulting coefficients are then used to predict the outcome on the corresponding test set. The MC mean of the mean absolute error (MAE) is 8.006 and of the mean squared relative error (MSRE) is 0.003. As a benchmark, we also compute predictions using the true non-null variables, obtaining MAE = 8.001 and MSRE = 0.003. These results indicate that, in the setting considered here, the selection process based on MI-seq preserves the model's predictive capability.

Cite this article: Bacci, S., Dreassi, E., Grilli, L. and Rampichini, C. (2026). Variable Selection via Knockoffs in Missing Data Settings with Categorical Predictors. *Psychometrika*, 1–22. https://doi.org/10.1017/psy.2026.10109