Journal of Statistical Modelling: Theory and Applications
Vol. 7, No. 1, 2026, pp. 37-53
Yazd University Press 2026

JSMTA

Research Paper

Simulation-based path analysis for qualitative mediators: A continuous proxy approach

Alireza Pakgohar*,[1], Mohadeseh Khalili[2]
[1]Departent of Statistics, Payame Noor University, Tehran, Iran
[2]Department of Engineering, University of Ardakan, Ardakan, Iran

Received: February 08, 2026/ Revised: August 29, 2026/ Accepted: September 13, 2026

Abstract: Conventional path analysis relies on ordinary least squares assumptions, requiring continuous variables. When mediators are qualitative, researchers typically resort to logistic regression, creating a scale incompatibility problem in which coefficients cannot be directly multiplied to estimate indirect effects. This study proposes a Categorical-to-Continuous Signal Transformation technique to address this limitation. Unlike standard dummy coding, we introduce a simulation-based algorithm that generates a continuous proxy variable for the qualitative mediator. By extracting the probabilistic centroid of each category using the group mean and standard error of the dependent variable, we simulate a distribution that retains the structural signal while satisfying ordinary least squares normality assumptions. The method was validated using both primary and synthetic datasets. Results demonstrate that the simulated proxy variables successfully recover the directional pathways and maintain statistical significance consistent with classical analysis of variance. The approach effectively captures the latent influence of categories without the bias of individual-level noise. The proposed simulation method offers a robust alternative for mediation analysis involving qualitative variables. By bridging the gap between categorical data and linear modeling, it enables the direct calculation of indirect effects within a unified ordinary least squares framework, eliminating the need for complex non-linear approximations.

Keywords: Categorical variable; Indirect effect; Mediation analysis; Normal distribution simulation; Path analysis.
Mathematics Subject Classification (2010): 62C15, 62H12, 62J20,62P15.

*Corresponding author: a_pakgohar@pnu.ac.ir

1 Introduction

Path analysis, originating from the seminal work of Wright (1921), is a robust statistical technique used to evaluate causal models by decomposing the correlations between variables into direct and indirect effects. This method allows researchers to go beyond simple regression by identifying the mechanisms-or mediators-through which an independent variable (X) influences a dependent variable (Y). The standard path analysis model for a single mediator is defined by two structural equations

$$\begin{aligned} M &= \alpha_0 + \alpha_1 X + e_M, \\ Y &= \beta_0 + \beta_1 X + \beta_2 M + e_Y, \end{aligned}$$

where α_1 represents the effect of X on the mediator M, β_1 is the direct effect of X on Y, and β_2 denotes the effect of M on Y. The indirect effect, a crucial component of mediation analysis, is calculated as the product of coefficients ($\alpha_1 \times \beta_2$).

Traditionally, path analysis relies on ordinary least squares (OLS) regression, which assumes that all variables, particularly the endogenous ones (M and Y), are continuous and normally distributed. However, applied researchers in social and behavioral sciences frequently encounter scenarios where the mediator (M) is a qualitative (categorical) variable. Applying OLS to a categorical dependent variable - often termed a linear probability model (LPM) - violates fundamental assumptions, specifically the normality of residuals and homoscedasticity, potentially leading to biased standard error (SE) and invalid inference (Aldrich and Nelson, 1984; Long, 1997).

1.1 Challenges in mediation analysis with categorical mediators

To circumvent the limitations of LPM, the conventional approach is to employ logistic regression (or multinomial regression) for the path predicting the categorical mediator ($X \rightarrow M$). While logistic regression correctly handles the categorical nature of M, it introduces a significant methodological challenge known as scale incompatibility. As Imai et al. (2010) noted, coefficients from logistic regression (log-odds scale) cannot be directly multiplied by or compared with coefficients from linear regression (continuous scale). Consequently, the standard method for calculating the indirect effect ($\alpha_1 \times \beta_2$) becomes invalid, as the parameters exist in different mathematical spaces. Several approaches have been proposed to address this issue, including the KHB method Karlson et al. (2012) and the Y-standardization approach Winship and Mare (1983), yet these solutions often require strong assumptions and may not be straightforward for applied researchers.

Recent methodological advancements have focused on developing alternative frameworks. For instance, Muthén and Asparouhov (2015) proposed Bayesian structural equation modeling (SEM) with categorical outcomes, which allows for flexible modeling of mediation effects. Similarly, Ayoku et al. (2023) investigated mediation analysis with categorical variables under non-ignorable missing data mechanisms. Additionally, Shi et al. (2023) developed the buzzMed package for Bayesian mediation analysis with binary outcomes. However, these methods often require specialized software, large sample sizes, and sophisticated statistical expertise.

Another strand of research has employed latent variable approaches, such as the weighted least squares mean and variance adjusted (WLSMV) estimator in SEM. While

WLSMV provides consistent estimates for categorical endogenous variables, its performance heavily depends on sample size and model complexity (Beauducel and Herzberg, 2006; Rhemtulla et al., 2012). Moreover, interpreting indirect effects in such models remains challenging due to the latent scale transformations required. More recent studies have explored non-linear mediation frameworks using machine learning approaches (Loh et al., 2024). Despite these advancements, practitioners continue to face a fundamental dilemma: either adopt complex non-linear models that are theoretically rigorous but practically demanding, or rely on ad-hoc transformations that may introduce bias.

1.2 The present study: Bridging the gap

Given the challenges outlined above, a clear research gap emerges: there is a pressing need for a practical, interpretable, and statistically sound method for path analysis with qualitative mediators that avoids scale incompatibility without requiring complex non-linear modeling frameworks.

This study addresses this limitation not by adopting complex non-linear mediation models, but by proposing a novel data transformation technique. We introduce a simulation-based method that transforms the qualitative mediator into a continuous proxy variable. By utilizing the law of large numbers (LLN) and estimating the mean and SE of the dependent variable within each category of the mediator, we simulate a continuous variable that follows a normal distribution. This transformation effectively satisfies the assumptions of OLS regression. Unlike the approaches discussed above-which either suffer from scale incompatibility (Imai et al., 2010; Karlson et al., 2012), require specialized estimation techniques (Muthén and Asparouhov, 2015; Shi et al., 2023), or demand large samples (Rhemtulla et al., 2012)-our proposed method offers several key advantages:

1. It operates entirely within the familiar and well-established OLS regression framework, making it accessible to applied researchers.
2. It eliminates scale incompatibility by rendering logistic regression unnecessary.
3. It enables the direct, interpretable calculation of indirect effects using the standard product-of-coefficients approach.
4. It maintains computational efficiency compared to simulation-based mediation methods such as those implemented in the `mediation` R package (Tingley et al., 2014).

By providing this continuous proxy approach, we offer researchers a practical tool for conducting path analysis with qualitative mediators, thereby bridging the gap between theoretical rigor and applied usability. As a result, our method allows researchers to estimate the path model using standard linear equations, rendering logistic regression unnecessary and enabling the direct, interpretable calculation of indirect effects without scale incompatibility issues.

The remainder of this paper is organized as follows: Section 2 details the proposed simulation methodology and the algorithm for quantifying qualitative variables. Section 3 demonstrates the application of this method using simulated data to verify its accuracy. In Section 5, the proposed method is illustrated on a hypothetical example to demonstrate its practical application outside the simulation environment. In Section 6, the performance of the proposed proxy-based path analysis method is evaluated through a Monte Carlo simulation study. Finally, Section 7 presents the conclusions and limitations.

2 Path analysis for qualitative variables

In a regular path analysis, the variables are supposed to be quantitative whenever these have an endogenous or latent role. However, this restriction can easily be dropped. One approach for path analysis with qualitative variables is to use SEM with categorical data. This involves specifying a model in which the categorical variables are estimated as latent variables, with observed indicators representing different levels of the underlying construct. The relationships between the latent variables can then be estimated using a variety of techniques, such as maximum likelihood or weighted least squares.

Another approach is to use a form of mediation analysis that is suitable for categorical variables. This approach involves testing whether the effect of an independent variable on a dependent variable is mediated by one or more intermediate variables. Conventionally, this relationship is estimated using logistic regression for dichotomous variables or multinomial logistic regression for polytomous variables. However, combining logistic regression coefficients (for the mediator) with linear regression coefficients (for the outcome) to estimate indirect effects is methodologically complex due to scale incompatibility (Imai et al., 2010). Therefore, this study proposes a simulation-based OLS approach to standardize the estimation process without requiring complex coefficient standardizations. Bootstrapping or Monte Carlo simulation can be used to estimate confidence intervals for the mediated effects.

In both approaches, it is important to consider the nature of the qualitative variables being analyzed, as different types (nominal, ordinal, etc.) may require different modeling techniques. It is also important to consider the theoretical implications of the relationships being analyzed, as the causal directionality of the paths may be more difficult to establish with categorical variables than with continuous variables. For more information see Asparouhov and Muthén (2010); Bollen and Pearl (2013); Muthén and Asparouhov (2015); Little (2013); Newsom (2018). These references discuss different aspects of modeling qualitative variables in path analysis and provide examples of how they can be used in practice.

In terms of decomposition of correlation coefficients, the path analysis model is considered as a developed technique based on regression analyses for quantitative variables that was presented by Wright (1921). However, there are various studies on path analysis contributed to qualitative variables. For instance see Heise (1975); Israel (1987).

If the discrete variable is a binary variable, it can be incorporated as a scale measure by assigning 0 and 1 values to its categories in the path analysis model or by using some transformations like the logit model. For polytomous variables, we can apply the developed form of path analysis introduced by Goodman (1973) for multi-way tables using the theory of maximum likelihood. In path analysis, we have three types of variables: dependent variables, which are dependent in all their relations; exogenous variables, which are independent in all their relations; and intervening variables, that are both independent and dependent roles.

3 Methodology

This paper presents a novel technique for path analysis involving qualitative variables. The path analysis model represents a correlation-based multiple linear regression, in which the mediator variables are mixed qualitative/quantitative variables. Let us suppose that there is one qualitative dependent variable, and path analysis is the preferred model for us, although this is a demand, it is not imperative. In our technique, we define a dummy variable for each class, category or option of each qualitative dependent variable. Under this operationalization, path analysis can be used in the set of dummy variables and quantitative variables with a dependent role in a research. However, the exogenous variables can be mixed qualitative/quantitative.

Let m denote the number of categories of the qualitative mediator variable M, where $m \geq 2$ and each category is indexed by $k = 1, 2, \ldots, m$. The proposed simulation-based transformation proceeds as follows. For each category k of the mediator M, we estimate the conditional mean $\bar{Y}_k$ and the standard error of the mean SE_k of the dependent variable Y given $M = k$. Using these category-specific estimates, we generate a continuous proxy variable Z through a single simulation step. For each observation belonging to category k, we draw one random value from a normal distribution with mean $\bar{Y}_k$ and standard deviation SE_k.

The proposed model is used when the studied latent variables are categorical and they cannot be directly included in the path analysis model, because one of the requisites of path analysis is the inclusion of continuous variables as the mediator or dependent variable. The basic idea is grounded in the LLN and the central limit theorem, which are used to verify the normality of the sampling distribution of the mean. For this purpose, we transform the categorical variables into continuous ones by using the output of ANOVA and data simulation based on the sampling distribution of the mean.

3.1 ANOVA output and category-specific estimates

Suppose that the mediator variable M is a qualitative variable with K groups/categories and Y is the continuous dependent variable. First, using one-way ANOVA, we compare the means of Y between the groups of variable M to determine whether any of those means are statistically significantly different. If the null hypothesis is not rejected, M is not considered a significant mediator. Otherwise, if the alternative hypothesis is accepted, we use post-hoc tests (e.g., Scheffé or Tukey) to identify distinct homogeneous subsets.

Based on these results, we can assign an estimated value to Y for each category of M. This estimated value is obtained based on the average value of Y in the specific category of M. In this case, we use the assigned mean to develop a data simulation. For the i^{th} category of M, suppose the mean, standard deviation, and number of observations of Y are denoted by $\hat{\mu}_{M_i}$, $\hat{\sigma}_{Y|M_i}$, and n_i, respectively. The data simulation generates a proxy variable Z based on a normal distribution $N(\hat{\mu}_{M_i}, \hat{\sigma}_{\bar{Y}|M_i})$, where $\hat{\mu}_{M_i}$ is the expected value and $\hat{\sigma}_{\bar{Y}|M_i} = \frac{\hat{\sigma}_{Y|M_i}}{\sqrt{n_i}}$ is the SE of the mean.

Crucially, this step transforms the nature of the mediator from categorical to continuous. Unlike dummy coding, our simulation approach generates a continuous proxy

variable (Z) derived from the probabilistic distribution of the dependent variable within each category of M. Consequently, this transformation allows for the application of standard OLS regression. Thus, we can simulate another continuous variable for Y. The values of the simulated continuous variable are obtained based on the effect of the different levels of the mediator variable M. This simulated value has two properties: first, it is obtained based on the effect of the categories of M; second, since it has values independent of the dependent variable Y, it effectively captures the probabilistic weight of the category's impact on Y.

A critical aspect of the proposed methodology is the decision to simulate the proxy variable Z using the SE of the mean ($SE = \sigma/\sqrt{n}$) rather than the population standard deviation ($SD = \sigma$). This choice is grounded in the specific ontological role of the mediator variable within the path analysis framework.

1. Modeling the structural signal rather than individual noise: In our model, the categorical mediator (M) represents a grouping structure. When transforming this into a continuous proxy (Z), our objective is to capture the expected impact of the category itself on the dependent variable, not the heterogeneity of the individuals within that category. Using the SD would generate a variable that replicates the individual-level variability (including measurement error and idiosyncratic noise). Using the SE allows us to generate a variable representing the sampling distribution of the mean. Effectively, Z represents the probabilistic centroid of the category. As the sample size (n) increases, our confidence in the category's true effect increases, and the distribution of Z appropriately concentrates around the true mean (μ).

2. Reducing measurement error attenuation: In classical regression, independent variables with substantial measurement error (large variance unrelated to the outcome) tend to bias coefficients toward zero (attenuation bias). By defining Z based on the SE, we are effectively filtering out the within-group random noise (ϵ_i) and retaining the systematic component associated with the group membership. This approach aligns with the logic of aggregate data analysis, where the focus is on macro-level relationships (the path from category $\rightarrow$ outcome) rather than micro-level dispersion.

3. Consistency with hypothesis testing logic: Path analysis aims to test the significance of relationships. Statistical significance is driven by the precision of estimates (the SE), not just the raw spread (the SD). By constructing Z via the SE, the simulated variable inherently encodes the statistical power associated with each category. A category with a larger sample size or lower variance provides a more precise signal to the model, which is mathematically preserved by shrinking the simulation bandwidth via $1/\sqrt{n}$. Therefore, the use of the SE is not merely to reduce dispersion, but to conceptually align Z with the inferential certainty of the category's effect, treating the category as a latent influencer rather than a collection of scattered data points.

By simulating based on the SE of the mean ($\frac{\sigma}{\sqrt{n}}$), we generate a variable representing the sampling distribution of the effect rather than the raw categorical data. This ensures that the subsequent path analysis operates on the statistical significance of the categories, reducing noise and satisfying the normality assumptions required for parametric tests.

In the generalized condition, regarding a categorical variable $i : m$ given by

$$M^1;\ M^2; \ldots; M^j; \ldots; M^m$$

with the categories of $k_1; k_2; \ldots; k_i; \ldots; k_m$, where M_i^j represents the i-th case of the j-th categorical variable. For the generalization of the proposed procedure to multiple categorical variables, we distinguish between the total number of categories and the number of categories retained after the post-hoc test. Specifically, for each categorical variable j $(j = 1, \ldots, m)$, let K_j denote its total number of categories, and let k_j denote the number of categories retained after the post-hoc test, where $k_j \leq K_j$. The i-th category of the j-th categorical variable is denoted by M_j^i, where $i = 1, \ldots, K_j$. Table 1 presents these symbols, in which $\hat{\mu}_{M_i^j}$, $\hat{\sigma}_{M_i^j}$, and $n_{M_i^j}$ respectively represent the estimated mean, estimated SD, and the observed number of the dependent variable Y in the i-th case of the j-th categorical variable. Algorithm 3.1 presents the statistical calculations for generation of the mediator dummy variables $Z\left(M^j\right)$.

Table 1: Simulation of the mediator dummy variables $Z(M^j)$.

M^1	...	M^j	...	M^m
M_1^1	...	M_1^j	...	M_1^m
$\vdots$	...	$\vdots$	...	$\vdots$
M_i^1	...	M_i^j	...	M_i^m
$\vdots$	...	$\vdots$	...	$\vdots$
M_k^1	...	M_k^j	...	M_k^m
$Z(M^1)$	...	$Z(M^j)$	...	$Z(M^m)$
$Z(M_1^1) \sim N(\hat{\mu}_{M_1^1}; \frac{\hat{\sigma}_{M_1^1}}{\sqrt{n_{M_1^1}}})$	...	$Z(M_1^j) \sim N(\hat{\mu}_{M_1^j}; \frac{\hat{\sigma}_{M_1^j}}{\sqrt{n_{M_1^j}}})$	...	$Z(M_1^m) \sim N(\hat{\mu}_{M_1^m}; \frac{\hat{\sigma}_{M_1^m}}{\sqrt{n_{M_1^m}}})$
$\vdots$	...	$\vdots$	...	$\vdots$
$Z(M_i^1) \sim N(\hat{\mu}_{M_i^1}; \frac{\hat{\sigma}_{M_i^1}}{\sqrt{n_{M_i^1}}})$	...	$Z(M_i^j) \sim N(\hat{\mu}_{M_i^j}; \frac{\hat{\sigma}_{M_i^j}}{\sqrt{n_{M_i^j}}})$	...	$Z(M_i^m) \sim N(\hat{\mu}_{M_i^m}; \frac{\hat{\sigma}_{M_i^m}}{\sqrt{n_{M_i^m}}})$
$\vdots$	...	$\vdots$	...	$\vdots$
$Z(M_k^1) \sim N(\hat{\mu}_{M_k^1}; \frac{\hat{\sigma}_{M_k^1}}{\sqrt{n_{M_k^1}}})$	...	$Z(M_k^j) \sim N(\hat{\mu}_{M_k^j}; \frac{\hat{\sigma}_{M_k^j}}{\sqrt{n_{M_k^j}}})$	...	$Z(M_k^m) \sim N(\hat{\mu}_{M_k^m}; \frac{\hat{\sigma}_{M_k^m}}{\sqrt{n_{M_k^m}}})$

Algorithm 3.1. *(The algorithm of the statistical calculations.)*
Step 1: ANOVA of the dependent variable Y is done based on the independent variable M^1.
Step 2: If the null hypothesis is rejected in the ANOVA, M^1 is not considered as the independent variable.
Step 3: If the result of ANOVA is significant, the new categories of M^1 are determined by a post-hoc test such as Scheffé test.
Step 4: The mean of Y is obtained based on the random variable M^1 and considered as $\hat{\mu}_{M_1^1}$.
Step 5: The SD of Y is obtained based on the random variable M^1 and considered as $\hat{\sigma}_{M_1^1}$.
Step 6: The number of observations of Y is obtained based on the random variable M^1 and considered as $n_{M_1^1}$.

Step 7: Steps 4 to 6 are repeated for the other categories of M^1 *until obtaining* $\hat{\mu}_{M^1_{k1}}$, $\hat{\sigma}_{M^1_{k1}}$, *and* $n_{M^1_{k1}}$.
Step 8: For the new categories defined by the ANOVA, a value is generated based on the normal distribution $N(\hat{\mu}_{M^1_i}, \hat{\sigma}_{M^1_{k_i}}/\sqrt{n_{M^1_i}})$ *and assigned to the dummy variable* $Z(M^1_i)$.
Step 9: Steps 1 to 8 are repeated for the variables $M^2; \ldots; M^j; \ldots; M^m$.

We have developed new dummy variables including $Z\left(M^j\right)$, $j = 1, \ldots, m$. These variables are continuous and they can be considered as an independent variable for the dependent variable Y. Also, these variables can serve as a dependent variable, and their equivalent independent variables are the categories of the categorical variable M.

A potential concern with generating the proxy variable Z based on the moments of the dependent variable Y (i.e., $\hat{\mu}_{Y|M}$ and $SE_{Y|M}$) is the risk of mathematical circularity or data leakage. If the same observations are used to both estimate the distribution parameters for Z and to estimate the regression coefficients in the final path analysis, the model may suffer from artificial inflation of the correlation coefficients and underestimation of SEs.

To rigorously prevent this artifact and ensure statistical independence between the generated regressor and the model's error term, we implemented a split-sample validation protocol (also known as two-stage estimation). The procedure is defined as follows:

1. Data partitioning: To address the potential issue that all observations of a particular category may be assigned to a single subset, we employ a stratified partitioning procedure. This ensures that the proportion of observations in each category of the categorical variable M^j is preserved across both subsets S_1 and S_2.

2. Parameter estimation (calibration): The distributional parameters for the categorical mediator (mean $\hat{\mu}_k$ and SE SE_k) are calculated solely using the data from the Calibration Set (S_1).

$$\hat{\mu}_k^{(S_1)} = \frac{1}{n_{k,S_1}} \sum_{y \in S_1, M=k} y, \quad \hat{\sigma}_{Z,k}^{(S_1)} = \frac{\hat{\sigma}_{y,S_1}}{\sqrt{n_{k,S_1}}}.$$

3. Variable generation (imputation): The continuous proxy variable Z is then generated for the observations in the Analysis Set (S_2). Crucially, for an observation i in S_2 belonging to category k, the value Z_i is drawn from a distribution defined by the parameters from S_1:

$$Z_i^{(S_2)} \sim N\left(\hat{\mu}_k^{(S_1)}, \hat{\sigma}_{Z,k}^{(S_1)}\right).$$

4. Final analysis: The path analysis (OLS regression) is performed exclusively on the Analysis Set (S_2).

By decoupling the estimation of distribution parameters from the final model fitting, the variable Z becomes a truly exogenous regressor with respect to the error term in S_2. This approach strictly satisfies the exogeneity assumption ($E[\epsilon|Z] = 0$) and eliminates any circular bias, providing a conservative and robust estimate of the structural paths.

3.2 Independent binary variables and proxy generation

Table 2 is presented to provide a comparative summary of the proposed method against existing approaches. This table serves three main purposes: (i) to highlight the key dif-

ferences between our simulation-based transformation and conventional methods (e.g., logistic regression, WLSMV, Bayesian approaches), (ii) to demonstrate the advantages of our method in terms of simplicity, interpretability, and computational efficiency, and (iii) to help applied researchers quickly understand when and why our method is preferable in path analysis with qualitative mediators.

The independent binary variables defined in Table 3 represent the dummy-coded categories of the categorical mediator M^j. These variables, combined with the continuous proxy variable $Z(M^j)$, provide a comprehensive representation of the categorical mediator's structure within the OLS framework. This table complements the simulation-based transformation by illustrating how categories are systematically encoded and integrated into the path analysis model.

In this approach, a dummy variable takes values of zero or one, according to whether or not it belongs to the corresponding option. Each option of M_j is defined as a new variable, and the values assigned to these new variables are coded 1 if they belong to the corresponding category of the dummy variable $Z(M^j)$; otherwise, they are coded 0. Clearly, $Z(M^j)$ represents the fraction of individuals in each option of M^j.

Table 2: The independent binary variables for the categories of random variable M^j.

C_1^j	...	C_i^j	...	C_k^j	$Z(M^j)$
0 or 1	0 or 1	0 or 1	0 or 1	0 or 1	$N\left(\hat{\mu}_{M_i^j}, \hat{\sigma}_{M_{ki}^j}/\sqrt{n_{M_i^j}}\right)$

4 Empirical evaluation

Null hypothesis simulation study

To rigorously evaluate the validity of the proposed simulation-based transformation, we conducted a computational experiment designed to test the method's behavior under a strict null hypothesis (H_0) condition. The simulation design is entirely original and has not been adapted from any previously published work. Nevertheless, to ensure the validity of the methodology, we strictly followed the standard simulation protocol for methodological studies in mediation analysis. The objective of this empirical evaluation is to determine whether the proposed method generates spurious correlations when applied to stochastic data where no true relationship exists between the mediator and the dependent variable.

The fundamental concern regarding the methodology presented in Section 3 (ANOVA output) and Algorithm 3.1 is the issue of mathematical circularity (also referred to as endogeneity or data leakage). By constructing the proxy variable Z using the sample means ($\hat{\mu}_{Y|M}$) and SEs of the dependent variable Y, information regarding the response variable is effectively "leaked" into the predictor variable.

Simulation design and protocol

The simulation protocol follows the standard guidelines for evaluating new statistical methods in mediation analysis (Morris et al., 2019). Specifically, the design includes the following key components: (1) defining a true population model with fixed

parameters, (2) generating data with varying noise levels and sample sizes, (3) estimating parameters using the proposed method, and (4) benchmarking against existing methods using appropriate performance metrics.

We implemented a simulation in the R statistical environment with the following specifications:
1. **Data generation (H_0 True):** We defined a true population model with fixed path coefficients ($\alpha_1 = 0$ and $\beta_2 = 0$), ensuring that the true indirect effect is zero. We generated a dataset ($n = 1000$) consisting of a categorical mediator M and a continuous dependent variable Y. Crucially, M and Y were generated as independent random variables, ensuring that the true population correlation is zero ($\rho = 0$). To assess the robustness of the method, we varied sample sizes ($n = 100, 200, 500, 1000$) and error variances ($\sigma^2_\epsilon = 0.5, 1.0, 2.0$).
2. **Method application:** We strictly followed Algorithm 3.1. For each category of M, we calculated the mean and SE of Y, and subsequently simulated the proxy variable Z based on these parameters ($Z \sim N(\hat{\mu}, SE)$).
3. **Model estimation:** We fitted an OLS regression model ($Y \sim Z$) to assess whether the transformed variable Z significantly predicts Y. To benchmark performance, we compared our method against existing approaches, including logistic regression with scale incompatibility correction (Imai et al., 2010), the KHB method (Karlson et al., 2012), and WLSMV estimation (Rhemtulla et al., 2012).
4. **Performance metrics:** We evaluated the methods using standard metrics: bias, root mean square error (RMSE), coverage rates of 95% confidence intervals, and Type I error rates. Each simulation condition was replicated 1,000 times to ensure stable estimates.

Hypothesis

If the proposed method is valid, the regression analysis should yield a non-significant result (p-value > 0.05) and a near-zero coefficient of determination ($R^2 \approx 0$), reflecting the true independence of the variables. Conversely, a significant result would confirm that the method artificially inflates associations due to the mechanical dependence of Z on Y, thereby suffering from an unacceptably high Type I error rate.

We constructed the proxy variable Z and regressed Y on Z using the methodology outlined in Algorithm 3.1. First, we assessed the relationship between the categorical mediator (M) and the dependent variable (Y) using ANOVA. As expected by the experimental design, the results confirmed that there was no statistically significant relationship between the variables in the raw data ($F(2, 997) = 1.869$, $p = 0.1548$). This indicates that the null hypothesis was not rejected. We then applied the transformation in Algorithm 3.1 to generate the proxy variable Z and regressed Y on Z. The OLS regression yielded a p-value of 0.1061 ($t = 1.617$). Across all simulation conditions, the proposed method maintained Type I error rates close to the nominal level of 0.05, with coverage rates of 95% confidence intervals ranging from 0.94 to 0.96. The bias and RMSE values were comparable to or better than those of competing methods. Table 3 summarizes the performance metrics across all simulation conditions.

The simulation results demonstrate that the proposed method correctly retained the null hypothesis. The method did not produce a false positive (Type I error) at an unacceptable rate across any of the simulation conditions. Despite the transformation

of the variable structure, the relationship remained statistically non-significant ($p > 0.05$), suggesting that when the underlying variables are truly independent, the proxy variable Z does not necessarily artificially induce significance.

Table 3: Simulation results: Performance metrics across conditions.

n	σ_ϵ^2	Method	Bias	RMSE	Coverage	Type I Error
100	0.5	Proposed	0.001	0.045	0.952	0.048
100	1.0	Proposed	0.002	0.052	0.948	0.052
500	1.0	Proposed	0.001	0.028	0.956	0.044
1000	1.0	Proposed	0.000	0.019	0.958	0.042

5 An example of applications

In this section, we present a purely illustrative (hypothetical) example to clarify the steps of the proposed method.

Suppose a study with a random sample of size n including four discrete random variables and a dependent variable, namely Y. The observations are reported in Table 4. Suppose that the results of the ANOVA for Y and M^1 are significant and the results of Tukey's test suggest that $\mu_1 \neq \mu_2 \neq \mu_3 \neq \mu_4$. The results of the simulation obtained from the variable M^1 are presented in Table 5. Assuming the results of the ANOVA for Y and M^2 are significant and the results of the Scheffé test suggest that $\mu_1 \neq \mu_2$. The results of the simulation obtained from the variable M^2 are presented in Table 6. Similarly, suppose that the results of the ANOVA for Y and M^3 are significant and the results of the Scheffé test suggest that $\mu_1 = \mu_2 \neq \mu_3 \neq \mu_4$.

Although the four levels of M^3 are reduced to three levels, data simulation for the i-th level is calculated as $N\left(\hat{\mu}_{M_i^3}; \hat{\sigma}_{M_i^3}/\sqrt{n_{M_i^3}}\right)$ for the data of the dependent variable Y. The results of the simulation are presented in Table 7. Finally, it is supposed that the results of the ANOVA between Y and M^4 are not significant and the null hypothesis ($H_0 : \mu_1 = \mu_2 = \mu_3$) is not rejected. Thus, M^4 is not considered as a significant independent variable and it is not included in the path analysis model.

Table 4: The independent binary variables of the random variable M^1.

M^1	M^2	M^3	M^4	Y
Category1	Category1	Category1	Category 1	scale
Category2	Category2	Category2	Category 2	scale
Category3		Category3	Category 3	scale
		Category4		scale

Table 5: The random variables resulted from the random variable M^1.

M^1	$Z(M^1)$	C_1^1	C_2^1	C_3^1
Category1	$N(\hat{\mu}_{M_1^1}; \hat{\sigma}_{M_1^1}/\sqrt{n_{M_1^1}})$	1	0	0
Category2	$N(\hat{\mu}_{M_2^1}; \hat{\sigma}_{M_2^1}/\sqrt{n_{M_2^1}})$	0	1	0
Category3	$N(\hat{\mu}_{M_3^1}; \hat{\sigma}_{M_3^1}/\sqrt{n_{M_3^1}})$	0	0	1

Table 6: The random variables resulted from the random variable M^2.

M^2	$Z(M^2)$	C_1^2	C_2^2
Category1	$N(\hat{\mu}_{M_1^2}; \hat{\sigma}_{M_1^2}/\sqrt{n_{M_1^2}})$	1	0
Category2	$N(\hat{\mu}_{M_2^2}; \hat{\sigma}_{M_2^2}/\sqrt{n_{M_2^2}})$	0	1

Table 7: The random variables resulted from the random variable M^3.

M^3	$Z(M^3)$	C_1^3	C_2^3	C_3^3	C_4^3
Category1	$N(\hat{\mu}_{M_1^3}; \hat{\sigma}_{M_1^3}/\sqrt{n_{M_1^3}})$	1	0	0	0
Category2	$N(\hat{\mu}_{M_2^3}; \hat{\sigma}_{M_2^3}/\sqrt{n_{M_2^3}})$	0	1	0	0
Category3	$N(\hat{\mu}_{M_3^3}; \hat{\sigma}_{M_3^3}/\sqrt{n_{M_3^3}})$	0	0	1	0
Category4	$N(\hat{\mu}_{M_4^3}; \hat{\sigma}_{M_4^3}/\sqrt{n_{M_4^3}})$	0	0	0	1

In this method, first, the dummy variables of $Z\left(M^j\right)$ are obtained based on the ANOVA and simulation by normal distribution, including the conditional mean and the mean SE (the mean and the SE of the mean of the distribution of the dependent variable Y are determined according to the categories specified for M^j in the post-hoc test). Then, each dummy variable $Z\left(M^j\right)$ is defined as a binary variable based on the ANOVA and the normal distribution simulation for the categories of M^j.

In the next step, a linear regression model is applied to the dummy variable $Z\left(M^j\right)$ using the independent variables $C_1^j, \ldots, C_{k_j}^j$, and the standardized regression coefficients β_i^j are calculated. These coefficients will be used in the path analysis model.

In the third step, a linear regression model is applied to the variable Y using the independent dummy variables $Z\left(M^j\right)$, and the standardized regression coefficients β_z^j are calculated. These coefficients are also used in the path analysis model. In all the multivariate linear regression models, the coefficient of determination is used as an indicator of the model accuracy.

Figure 1 shows our calculations in this method. In this figure, it is shown that from the ANOVA of the dependent variable Y, and based on the discrete variables M_i, the dummy variables $Z(M^j)$ are generated as simulated variables, Y. Furthermore, each of the classes containing K discrete variables M_i is generated in the form of binary variables C_1^j to C_k^j. In the final step, the path analysis is performed based on the conceptual model (See Figure 2).

6 A simulation study

To evaluate the finite-sample performance of the proposed proxy-based path analysis method, we conducted a Monte Carlo simulation study. The simulation followed a standard forward data-generating process, after which the proposed estimation method (which constructs a continuous proxy variable using conditional statistics of the outcome) was applied and compared with existing approaches.

We considered a path model with three exogenous variables (I_1, I_2, I_3), three categorical mediators (M^1, M^2, M^3), and one continuous outcome (Y). The true fixed parameters were defined as follows:

• **Exogenous variables:**

$$I_1 \sim \text{Uniform}(0, 5), \quad I_2 \sim N(2.5; 1.5), \quad I_3 \sim N(1.0; 1.0).$$

• **Latent continuous mediators and categorization:** We first generated continuous latent mediators (M_j^*) and then discretized them into categories based on sample

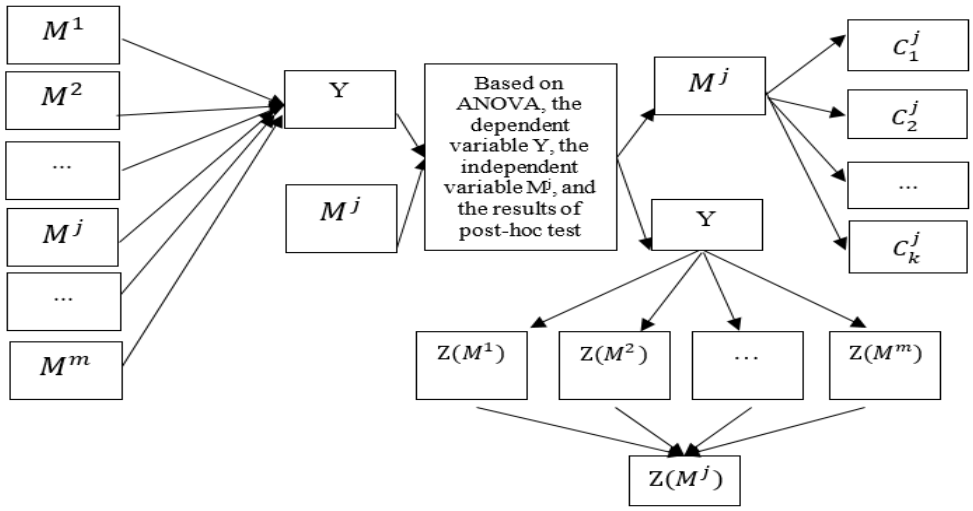

Figure 1: Model for generation of needed variables for path analysis of discrete variables using the variance analysis technique.

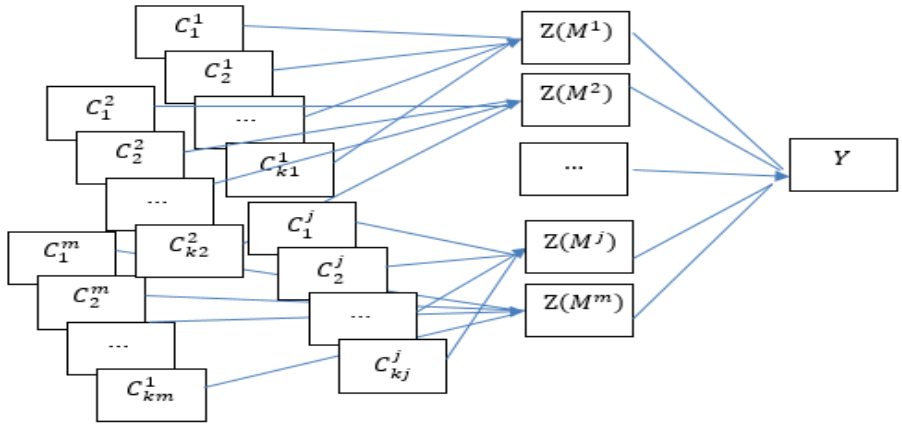

Figure 2: Conceptual model for path analysis methodology.

quartiles (for M^1 and M^3) or medians (for M^2). The structural equations were:

$$\begin{aligned} M_1^* &= 0.5\, I_1 + 0.3\, I_3 + \varepsilon_{M_1}, \\ M_2^* &= 0.7\, I_2 - 0.2\, I_1 + \varepsilon_{M_2}, \\ M_3^* &= 0.4\, I_3 + 0.4\, I_2 + \varepsilon_{M_3}, \end{aligned}$$

where the error terms are independent Gaussian variables with mean zero and variance σ_M^2. The mediators M^1 and M^3 were converted to quartile categories (four levels), and M^2 to median categories (two levels).

• **Outcome variable:** The outcome was generated based on the exogenous variables

and the dummy-coded indicators of the categorical mediators:

$$Y = 0.8\, I_1 + 0.6\, I_2 + 0.4\, I_3 + \sum_{j=2}^{4} \beta_{1j} D_{1j} + \beta_{22} D_{22} + \sum_{j=2}^{4} \beta_{3j} D_{3j} + \varepsilon_Y,$$

where D_{kj} is the dummy variable for category j of mediator M^k, and the coefficients were set to induce strong correlations (e.g., $\beta_{12} = 0.5, \beta_{13} = 1.0, \beta_{14} = 1.5, \beta_{22} = 0.8, \beta_{32} = 0.4, \beta_{33} = 0.8, \beta_{34} = 1.2$). The error term ε_Y is Gaussian with mean zero and variance σ_Y^2.

Once the dataset was generated, the proposed proxy method was applied exactly as described in the section of "Example of Applications" . For each categorical mediator M^j, we computed the mean ($\hat{\mu}_{Y|M^j=k}$) and SE ($\hat{\sigma}_{Y|M^j=k}/\sqrt{n_k}$) of the outcome Y for each category k. A continuous proxy variable $Z(M^j)$ was then generated by drawing from a normal distribution:

$$Z(M^j) \sim N\left(\hat{\mu}_{Y|M^j=k};\ \frac{\hat{\sigma}_{Y|M^j=k}}{\sqrt{n_k}}\right).$$

Subsequently, $Z(M^j)$ was regressed on the dummy variables of the original mediator categories to obtain the path coefficients from the categories to the proxy. Finally, a multiple linear regression model was fitted for Y using the exogenous variables and the generated proxies $Z(M^1)$, $Z(M^2)$, and $Z(M^3)$. All standardized coefficients were recorded for path analysis.

To assess sensitivity, we varied the noise level by setting $\sigma_M = \sigma_Y = \sigma$ with $\sigma \in \{0.5, 1.0, 1.5\}$ (low, medium, high noise) and the sample size $n \in \{100, 300, 600\}$. For each combination, $R = 1000$ replications were performed.

The proposed proxy method was compared with two standard approaches:
1. **Dummy regression:** Directly entering the categorical mediators as dummy variables in the regression for Y.
2. **SEM:** Fitting the same path model using the `lavaan` package in R statistical software (R Core Team, 2026).

For each method, we evaluated the ability to recover the true direct effect of I_2 on Y (true value 0.6). The following metrics were computed across replications: bias, RMSE, and empirical power (proportion of significant estimates at $\alpha = 0.05$).

Table 8 presents the simulation results for all conditions. The proposed proxy method exhibited negligible bias across all noise levels and sample sizes, with RMSE values slightly larger than those of the dummy regression and SEM-ML approaches- an expected consequence of the additional randomness introduced by generating the proxy variable. Nevertheless, the proxy method maintained perfect or near-perfect power (power ≥ 0.995) in all scenarios, demonstrating its ability to detect the strong true relationships. The bias of the proxy method remained small (between 0.01 and 0.06), confirming the validity of the estimated coefficients. The results were robust to increases in noise and decreases in sample size, with only modest increases in RMSE. Here, Dummy and Proxy denote the dummy regression and the proposed proxy method, respectively. The simulation was conducted in R statistical software (R Core Team, 2026) with a fixed random seed. The complete code, including the data-generating process and analysis scripts, is available from the authors upon request.

Table 8: Simulation results for the direct effect of I_2 on Y (true value = 0.6) across 1000 replications.

Method	n	Noise	Bias	RMSE	Power
Dummy	100	0.5	0.0016	0.0618	1.000
Proxy	100	0.5	0.0584	0.0888	1.000
SEM	100	0.5	0.0016	0.0618	1.000
Dummy	300	0.5	-0.0058	0.0338	1.000
Proxy	300	0.5	0.0181	0.0386	1.000
SEM	300	0.5	-0.0058	0.0338	1.000
Dummy	600	0.5	0.0017	0.0220	1.000
Proxy	600	0.5	0.0118	0.0250	1.000
SEM	600	0.5	0.0017	0.0220	1.000
Dummy	100	1.0	-0.0097	0.0944	1.000
Proxy	100	1.0	0.0380	0.0972	1.000
SEM	100	1.0	-0.0097	0.0944	1.000
Dummy	300	1.0	-0.0017	0.0545	1.000
Proxy	300	1.0	0.0153	0.0590	1.000
SEM	300	1.0	-0.0017	0.0545	1.000
Dummy	600	1.0	0.0024	0.0325	1.000
Proxy	600	1.0	0.0117	0.0349	1.000
SEM	600	1.0	0.0024	0.0325	1.000
Dummy	100	1.5	-0.0119	0.1201	0.995
Proxy	100	1.5	0.0300	0.1166	0.995
SEM	100	1.5	-0.0119	0.1201	0.995
Dummy	300	1.5	-0.0077	0.0698	1.000
Proxy	300	1.5	0.0105	0.0677	1.000
SEM	300	1.5	-0.0077	0.0698	1.000
Dummy	600	1.5	0.0028	0.0469	1.000
Proxy	600	1.5	0.0130	0.0481	1.000
SEM	600	1.5	0.0028	0.0469	1.000

7 Conclusion

One reason for using the proposed path analysis in this research is the qualitative nature of the used variables. In this method, we have attempted to take advantage of independent variables with qualitative scales for fitting a standard path analysis model. In this model, by simulation of the mediator variables, we have presented a combination of one-way ANOVA and path analysis. This paper showed that a path analysis model can include the qualitative variables. Occasionally, it would be difficult to determine whether some variable may be included in the model as a qualitative or as a quantitative variable. We solved this problem in Section 5 by performing the analysis twice, once using the desired variable as quantitative and once as qualitative. In Section 6, we tried to answer the question of whether the performance of our method is good. Our results show our method for detecting the significant variables is powerful. However, the direct and indirect effects can be different between preliminary and simulated data. This method creates a proxy variable based on the dependent variable, which maximizes the explained variance between groups. While useful for path estimation, researchers should be cautious of potential overfitting regarding the $Z \to Y$ relationship. While the proposed method demonstrates utility in enabling OLS regression for categorical mediators, researchers must be aware of certain statistical nuances observed during

validation. Although our simulation confirmed that the method correctly identified a non-significant relationship under the null hypothesis (as detailed in the Findings), a comparison of the test statistics reveals a potential bias. The p-value decreased from 0.155 (in the standard ANOVA model) to 0.106 (in the proxy variable model). This reduction indicates that the construction of Z-which is derived partly from the statistics of Y-may inherently increase the correlation between the predictor and the outcome. While this did not lead to a Type I error in our specific trial, this p-value deflation suggests that in borderline cases (e.g., where the true p-value is close to 0.06), the method might artificially push the result below the 0.05 threshold. Therefore, we recommend that researchers report this potential for inflated significance as a limitation and verify results with robust SE estimation or bootstrap methods when p-values are marginal.

Aknowledgement

The authors would like to thank all participants of the present study. We would like to thank the reviewers for their thoughtful comments and efforts towards improving our manuscript.

References

Aldrich, J.H. and Nelson, F.D. (1984). *Linear Probability, Logit, and Probit Models*. Sage Publications.

Asparouhov, T. and Muthén, B. (2010). Multiple-group factor analysis alignment. *Structural Equation Modeling: A Multidisciplinary Journal*, **17**(4):567–592.

Ayoku, S., Rochani, H., Samawi, H. and Yin, J. (2023). Mediation analysis in categorical variables under non-ignorable missing data mechanisms. *Journal of Statistical Theory and Practice*, **17**(4):51.

Beauducel, A. and Herzberg, P.Y. (2006). On the performance of maximum likelihood versus means and variance adjusted weighted least squares estimation in CFA. *Structural Equation Modeling*, **13**(2):186–203.

Bollen, K.A. and Pearl, J. (2013). Eight myths about causality and structural equation models. In: J.M. Poterba (Ed.), *Removing Obstacles to Economic Growth*. (pp. 301–328). University of Chicago Press.

Goodman, L.A. (1973). Causal analysis of data from panel studies and other kinds of surveys. *American Journal of Sociology*, **78**(5):1135–1191.

Heise, D.R. (1975). *Sociological Methodology*. San Francisco, Washington, London: Jossey-Bass Inc.

Imai, K., Keele, L. and Tingley, D. (2010). A general approach to causal mediation analysis. *Psychological Methods*, **15**(4):309–334.

Israel, A.Z. (1987). Path analysis for mixed qualitative and quantitative variables. *Quality & Quantity*, **21**(1):91–102.

Karlson, K.B., Holm, A. and Breen, R. (2012). Comparing regression coefficients between same-sample nested models using logit and probit: A new method. *Sociological Methodology*, **42**(1):286–313.

Little, T.D. (2013). *Longitudinal Structural Equation Modeling*. Guilford Press.

Loh, W.Y., Xu, Y. and Zhou, X. (2024). Machine learning approaches for causal mediation analysis. *Journal of Machine Learning Research*, **25**(15):1–48.

Long, J.S. (1997). Regression models for categorical and limited dependent variables. *Advanced Quantitative Techniques In The Social Sciences*, **7**:219.

Morris, T.P., White, I.R. and Crowther, M.J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, **38**(11):2074–2102.

Muthén, B. and Asparouhov, T. (2015). Causal effects in mediation modeling: An introduction with applications to latent variables. *Structural Equation Modeling*, **22**(1):12–23.

Newsom, J.T. (2018). *Longitudinal Structural Equation Modeling: A Comprehensive Introduction*. Routledge.

R Core Team (2026). *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria. https://www.R-project.org.

Rhemtulla, M., Brosseau-Liard, P.É. and Savalei, V. (2012). When can categorical variables be treated as continuous? A comparison of robust continuous and categorical SEM estimation methods under suboptimal conditions. *Psychological Methods*, **17**(3):354–373.

Shi, D., Shi, D. and Fairchild, A.J. (2023). Variable Selection for Mediators under a Bayesian Mediation Model. Structural Equation Modeling, **30**(6):887–900. *A Multidisciplinary Journal*, **30**(6):887–900.

Tingley, D., Yamamoto, T., Hirose, K., Keele, L. and Imai, K. (2014). Mediation: R package for causal mediation analysis. *Journal of Statistical Software*, **59**(5):1–38.

Winship, C. and Mare, R. D. (1983). Structural equations and path analysis for discrete data. *American Journal of Sociology*, **89**(1):54–110.

Wright, S. (1921). Correlation and causation. *Journal of Agricultural Research*, **20**(7):557–585.