Psychometrika (2025), **0**, 1–25
doi:10.1017/psy.2025.13

PSYCHOMETRIC SOCIETY

THEORY AND METHODS

Testing of Reverse Causality Using Semi-Supervised Machine Learning

Nan Zhang[1], Heng Xu[1], Manuel J. Vaulont[2] and Zhen Zhang[3]

[1]Department of Management, Warrington College of Business, University of Florida, Gainesville, FL, USA; [2]Management and Organizational Development Group, D'Amore-McKim School of Business, Northeastern University, Boston, MA, USA; [3]Department of Management, Strategy and Entrepreneurship, Edwin L. Cox School of Business, Southern Methodist University, Dallas, TX, USA
Corresponding author: Nan Zhang; Email: zhang.nan@ufl.edu

(Received 2 January 2025; accepted 27 March 2025)

Abstract
Two potential obstacles stand between the observation of a statistical correlation and the design (and deployment) of an effective intervention, *omitted variable bias* and *reverse causality*. Whereas the former has received ample attention, comparably scant focus has been devoted to the latter in the methodological literature. Many existing methods for reverse causality testing commence by postulating a structural model that may suffer from widely recognized issues such as the difficulty of properly setting temporal lags, which are critical to model validity. In this article, we draw upon advances in machine learning, specifically the recently established link between causal direction and the effectiveness of semi-supervised learning algorithms, to develop a novel method for reverse causality testing that circumvents many of the assumptions required by traditional methods. Mathematical analysis and simulation studies were carried out to demonstrate the effectiveness of our method. We also performed tests over a real-world dataset to show how our method may be used to identify causal relationships in practice.

Keywords: machine learning; reverse causality; semi-supervised learning

1. Introduction

A fundamental purpose of research in psychology—and many other disciplines in social sciences for that matter—is to understand causal relationships between variables. In particular, it is both theoretically and practically important to distinguish between the mere observation of associations (between variables) and cases where causality can be inferred. When controlled randomized experiments are impractical, how to properly do so has garnered considerable attention in multiple disciplines (e.g., Pearl, 1998) including psychology (Rogosa, 1980). The practical significance of this inquiry becomes apparent considering the pivotal role that causal inference plays in formulating effective intervention strategies. For example, research has consistently shown that individuals who occupy central positions in their social networks (i.e., *centrality*) tend to receive more favorable assessment of *charisma* from their connections (Balkundi et al., 2011). Yet, the mere identification of this correlation[1] does not permit the conclusion that enhancing an individual's social-network centrality would be an effective intervention for boosting their charisma. As a case in point, whereas some scholars view charisma as an outcome

[1]Throughout this article, we use "correlation" in the broader sense to refer to any (i.e., linear or nonlinear) co-variation (of two or more variables, constructs, etc.) that manifests in an observed dataset.

of social-network patterns (Pastor et al., 2002), others contend that charisma is what attracts followers and allows an individual to occupy a central position in the first place (Shils, 1965), suggesting that the aforementioned intervention might be less effective.

Two potential obstacles stand between (a) the observation of a robust correlation between two constructs X and Y, like the centrality-charisma correlation, and (b) the effectiveness of deploying an intervention on X to change Y, like an attempt to boost centrality in order to increase charisma (Bollen, 1989, p. 41). The first is *omitted variable bias* (Mauro, 1990), meaning that an unobserved confounder induces the co-variation of both X and Y. For example, individuals high on extraversion may excel on both social-network centrality and charisma. The other is *reverse causality* (Leszczensky & Wolbring, 2022), meaning that the causal influence flows in the reverse direction (i.e., $Y \rightarrow X$) from what is required by the intended intervention—e.g., charisma influences centrality but not the other way around. There is a substantial body of methodological research on assessing the magnitude of omitted variable bias in a relationship (e.g., Busenbark et al., 2022; Cinelli & Hazlett, 2020; Harring et al., 2017; Mauro, 1990). Yet comparatively little attention has been directed toward testing for reverse causality (Leszczensky & Wolbring, 2022). Admittedly, the issue may not be applicable in situations where a clear causal direction is obvious (e.g., cancer → age is implausible). Nonetheless, for psychological constructs, concerns on reverse causality are prevalent, especially when theoretical arguments could be made for both directions. The focus of this work is to develop a method for testing reverse causality in observational data, with an emphasis on panel data.

To empirically investigate the issue of reverse causality in observational data, some existing methods, like cross-lagged panel models (CLPM; Hamaker et al., 2015), commence by presuming a reciprocal relationship, before postulating a structural model based on the assumption and empirically estimating the model to assess the magnitude in both directions while considering auto-regressive effects. Some other methods infer the direction of causality by exploiting certain distributional and/or functional features assumed of the underlying data-generating process, such as non-normality in a linear model (e.g., Shimizu et al., 2006; von Eye & DeShon, 2012), statistical independence between input variables and additive noise (e.g., Hoyer et al., 2008; Rosenström et al., 2023), etc. These methods pose two unsolved issues. First, from a theoretical perspective, given that many psychological theories (and theories in other fields) stipulate unidirectional (instead of reciprocal) relationships, there exists substantial theoretical interest in *testing* the direction of causality rather than assuming it away with the adoption of a reciprocal model like CLPM. For example, the structural advantage theory of social networks (Brass, 1984; Burt, 1992) posits a unidirectional effect of one's social network characteristics (e.g., centrality) on individual behaviors (e.g., charismatic leadership). Given the robust correlation reported in the literature for centrality and charisma (e.g., Balkundi et al., 2011), a well-powered empirical study that shows the absence of a reverse effect (i.e., charisma $\not\rightarrow$ centrality) would provide strong evidential value for supporting the structural advantage theory. Second, from a methodological perspective, formulating an appropriate model—or accurately specifying its distributional/functional features (e.g., whether it is linear)—can be particularly challenging when dealing with panel data (Hamaker, 2024; Lucas, 2023). For example, the validity of many existing panel models (or even the notion of Granger causality itself) is known to break down when the sampling frequency is improperly specified (Shojaie & Fox, 2022; Vaisey & Miles, 2017), when there are confounders that are unaccounted for (e.g., Hamaker et al., 2015), etc.

A promising avenue for addressing these issues of existing methods arises from causal learning (Peters et al., 2017), a branch of machine learning that injects causal inference into the design of learning algorithms. Central to this approach is a fundamental question: if we were to train a machine learning model that predicts Y from X (based on a limited number of training data points $\langle X, Y \rangle$), could knowledge about the probability distribution of X (i.e., $P(X)$) help us improve the predictive accuracy of the trained model? Schölkopf et al. (2012) show that the answer is positive *if and only if* the causal direction flows from Y to X. This suggests that, by demonstrating a machine learning model's ability to capitalize on $P(X)$ for enhancing predictive accuracy toward Y, we would identify the existence of reverse causality. Even more importantly, this proposition would hold irrespective of the functional form of the X–Y relationship.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Whereas Schölkopf et al. (2012) establish the dependence between causal direction and the predictive accuracy of semi-supervised learning algorithms, they leverage this finding to understand why semi-supervised learning works over some datasets but not others, instead of developing a concrete method for testing reverse causality. In the current research, we address this gap by developing a novel method for testing reverse causality over panel data. The contribution of our work is two-fold. First, our method for reverse causality testing allows researchers to rigorously test their causal theories by addressing two issues that researchers face. On the one hand, our method allows researchers to directly test unidirectional relationships, instead of assuming these relationships are reciprocal in nature. On the other hand, our method does not require the specification of distributional or functional features such as the sampling frequency or the shape of the proposed relationship. By demonstrating the effectiveness of our method using simulation studies and a case study, we show its applicability for psychology researchers in theory building and testing. Second, our work pioneers the integration of advancements in causal learning into the methodological arsenal of psychology for causal inference. We present conceptual arguments and mathematical formalization that link reverse causality testing with the predictive accuracy of semi-supervised learning (Van Engelen & Hoos, 2020). In doing so, we enrich the understanding of how machine learning could contribute to the psychological methods literature (e.g., Sterner et al., 2023; Wilcox et al., 2023; Zimmer & Debelak, 2023) and open up future avenues of inquiry at the intersection of psychological research and computer science.

2. Literature review

In this section, we briefly review the existing literature on reverse causality testing and the machine learning method we propose to use (i.e., semi-supervised learning).

2.1. Reverse causality testing

Compared with the rich and growing literature on omitted variable bias—in psychology (Harring et al., 2017), sociology (Halaby, 2004) and economics (Wüthrich & Zhu, 2023)—researchers across disciplines have made relatively limited progress on the testing of reverse causality over panel data. As summarized by Leszczensky & Wolbring (2022), a common approach is to specify causal direction in a panel model by applying *temporal lags* on variables that represent the "cause" after partialing out auto-regressive effects. For example, a causal direction of $X \rightarrow Y$ would be reflected by setting a lagged (e.g., previous-wave) value of X and the contemporary value of Y as independent and dependent variables, respectively, in the panel model, suggesting that X has a causal, lagged, effect on Y. A variety of panel models follow this idea (Orth et al., 2021), such as lagged first-difference (LFD) models (Vaisey & Miles, 2017), CLPM (Hamaker et al., 2015), etc.

As these existing panel models rely on temporal lags to identify the causal direction, how to specify the amount of this temporal lag becomes a prominent question. Theoretically deriving the "correct" temporal lag is obviously even more challenging than discerning the causal direction, suggesting the need for methods to be robust to misspecified temporal lags. Unfortunately, Vaisey & Miles (2017) show that panel models such as LFDs can be highly sensitive to the misspecification of temporal lags. Leszczensky & Wolbring (2022, Figure 2) also demonstrate that, when a contemporaneous effect is mischaracterized as lagged in a CLPM, the model can produce highly biased estimates. More fundamentally, the very notion of Granger causality, which underpins all panel models, is known to break down with a mis-specified temporal lag (Shojaie & Fox, 2022).

Beyond panel models, other existing methods for reverse causality testing rely on certain distributional/functional features assumed of the underlying data-generating process. Some—like Direction Dependence Analysis (DDA; Li & Wiedermann, 2020; Pornprasertmanit & Little, 2012; von Eye & DeShon, 2012; Wiedermann & Li, 2018) and Linear Non-Gaussian Acyclic Models (LiNGAM; Shimizu et al., 2006)—exploit the non-normality of data distributions in a linear model to infer causal direction.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

In the case of DDA, for example, the causal direction may be inferred by comparing the degrees of departure[2] from normality across different variables. For longitudinal data, extensions of these methods (e.g., Bauer et al., 2016; Geiger et al., 2015; Hyvärinen et al., 2010) leverage the non-normality of noise to identify causal direction in multivariate time series. Additionally, methods like Rosenström et al.'s (2023) directional analysis approach, additive noise models (Hoyer et al., 2008; Peters et al., 2014) and (more generally) post-nonlinear models (Zhang & Hyvärinen, 2009) infer causal direction by assuming statistical independence between the input variables and an additive noise component within linear or nonlinear models. Yet, like panel models' reliance on properly specified temporal lags, these methods also hinge on accurately defining certain distributional or functional features of the relationship between variables—knowledge that may not be available *a priori*.

These issues of existing methods raise an important question: Can we test for reverse causality *without* constraining the functional form of the data-generating process? Doing so would not only circumvent the complexities of specifying the temporal lag, but also relax the linearity or additive noise assumptions that permeate existing methods. We seek to answer this question in the current work by integrating recent advances in machine learning and the broader causal inference literature.

2.2. Semi-supervised learning

The objective of semi-supervised learning is to approximate the function f that links independent variables X to a dependent variable Y such that $Y = f(X)$. This is done by learning from two datasets containing i.i.d. samples. The first dataset, typically known as *labeled set*, provides $\langle x_i, y_i \rangle$ (i.e., paired X–Y values) for n_1 data points. The second dataset provides only the values of X, but *not* Y, for other n_2 data points, and is therefore known as the *unlabeled set*. In machine learning, it is generally assumed that n_1 is much smaller than n_2, due mainly to the high cost of acquiring Y in practice. For example, a task for which semi-supervised learning has shown remarkable success is image classification (Xie et al., 2020). With this task, X is an image and Y is its category (e.g., landscape, portrait). It is virtually cost-less to collect millions of images from the web, but considerably more expensive to hire human workers to properly label the collected images. In this case, we may opt to manually label only a few observations of X, leading to the assumption that $n_1 \ll n_2$.

The limited size of the labeled set means that, if we were to launch a canonical supervised learning algorithm (e.g., OLS or logistic regression), which can only learn from the labeled set, we would not be able to obtain an accurate prediction of Y. The uniqueness of semi-supervised learning lies in its ability to leverage the n_2 unlabeled, X-only, data points to significantly improve the predictive accuracy of the learned $\hat{f}$. To this end, numerous methods have been proposed for semi-supervised learning (Van Engelen & Hoos, 2020). A famous example, which we will further elaborate later in the article, is self-training (e.g., Sohn et al., 2020), which can be readily integrated with many supervised learning algorithms such as logistic regression. With self-training, we start by running a supervised learning algorithm (e.g., logistic regression) over the n_1 labeled data points to generate an approximation of f, denoted by $\hat{f}_1$, which predicts Y based on an input X. Then, we apply $\hat{f}_1$ over each of the n_2 unlabeled data points to predict its *pseudo-label* (i.e., an estimate of Y). Note that when a machine learning model is used for prediction, it may generate not only a point-estimate (e.g., binary prediction in logistic regression) but also a confidence level associated with the estimate (e.g., log-odds in logistic regression). Leveraging this, we select from the n_2 pseudo-labels those with confidence above a pre-determined threshold, add their X values paired with pseudo-labels (i.e.,) $\langle x_i, \hat{f}_1(x_i) \rangle$ to the labeled set, before running the supervised learning algorithm again to update our approximation of f. This process can continue iteratively until no more pseudo-labels can be added.

Whereas the efficacy of semi-supervised learning has long been established (Van Engelen & Hoos, 2020), there has been limited research probing *why* unlabeled data, which lack any information about Y, can bolster our understanding of the X–Y relationship. To this end, Buja et al. (2019) establish

[2]Such degrees of departure are typically quantified using higher-order central moments, such as skewness and kurtosis.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

that the distribution of a variable X may convey rich information about the X–Y relationship when the relationship is not strictly linear. For the high-dimensional case where X comprises multiple variables, Niyogi (2013) attributes the success of semi-supervised learning to the concept of manifold regularization (Belkin et al., 2006). At its core, Niyogi (2013) posits that almost all machine learning algorithms predicate their predictions on a smoothness assumption: if two data points are similar in X, then they also tend to be analogous in Y. Unfortunately, defining "similar" in the context of multi-dimensional X is challenging—and requires structural insights into the multivariate distribution of X—because common similarity measures like Euclidean distance are known to become meaningless in high-dimensional spaces (Aggarwal et al., 2001). By providing a more refined depiction of the multi-dimensional distribution of X, unlabeled data allow us to more accurately estimate the similarity of two data points in X, thereby enhancing our predictive precision for Y (Niyogi, 2013).

While this line of research sheds light on the mechanics of semi-supervised learning, it does not address why semi-supervised learning achieves high predictive accuracy with certain datasets but performs poorly with others (Schölkopf et al., 2012). The missing piece—the key to understanding the conditions under which semi-supervised learning succeeds—lies in the causal effects driving the generation of the observed data, as we will elaborate in the next section.

3. Linking causality with semi-supervised learning

In this section, we draw upon insights from causal learning (Peters et al., 2017)—specifically Janzing & Schölkopf's (2015) mathematical characterisation of Schölkopf et al.'s (2012) seminal finding on the working condition for semi-supervised learning—to elucidate how the predictive accuracy of semi-supervised learning over panel data may reveal the existence of reverse causality in the underlying data-generating process. Our explanation unfolds in three steps. First, we offer conceptual arguments through an illustrative example. Following this, we delve into mathematical formalization, analyzing a special case where the data-generating process features linear, cross-lagged, effects. Finally, we expand our discussions to justify the reasoning behind a test that can be applied to any arbitrary data-generating process. The specific computational algorithms for semi-supervised learning will be described in the next section.

3.1. Conceptual illustration

In what follows, we offer an intuitive explanation for the key insight of Schölkopf et al. (2012) through two observations: (1) $P(X)$ is useless for predicting Y if $X \to Y$, and (2) if $Y \to X$, $P(X)$ (when combined with a small labeled set) may lead to an accurate prediction model toward Y. The first is obvious: When $X \to Y$ fully characterizes the X–Y relationship, we can represent their relationship as a stochastic function f such as $Y = f(X)$. Changing f clearly has no impact on $P(X)$. Consequently, knowledge of $P(X)$ reveals no information about f, making it useless for the prediction of Y (Janzing & Schölkopf, 2010).

For the second observation, consider a simple example of $Y \to X$ depicted in Figure 1, where $Y \in \{0,1\}$ is binary and $X = \alpha Y + \varepsilon$ with $\varepsilon \sim N(0,\sigma^2)$. In this case, altering the causal mechanism (i.e., α) obviously modifies $P(X)$. From the perspective of machine learning, this means that $P(X)$ now encapsulates certain cues regarding α, offering the potential for building an accurate prediction model toward Y. As can be seen from Figure 1, for this specific example, knowledge of $P(X)$ alone permits an exact inference[3] of α, which is all that is required to build a Bayes-optimal predictior of Y.

[3]While outside the scope of this article, this exact inference also works when Y is not binary but follows other distributions such as Poisson (due to Linnik's theorem; Linnik, 1957). Interested readers could refer to the rich literature of random variable decomposition (Lukacs, 1970) and the more recent literature of additive noise model (Hoyer et al., 2008; Peters et al., 2014) which are both intimately related to the unique identification of $P(Y|X)$ (and thereby α and σ) given $P(X)$.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

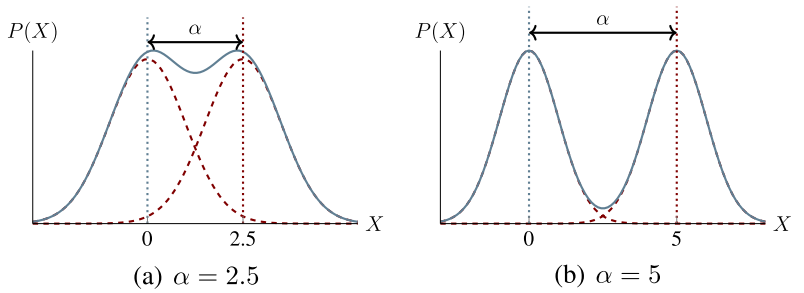

Figure 1. Illustrative example for $Y \to X$.
Note: Both panels depict the probability density function of $P(X)$ when $X = \alpha Y + \varepsilon$, where Y follows Bernoulli distribution with $p = 0.5$ and $\varepsilon \sim N(0,1)$. Note that, in either case, $P(X)$ follows a Gaussian mixture distribution with two equal-weight components, which are illustrated in red dashed lines. The mean difference between the two Gaussian components is always equal to α, suggesting that the functional relationship between X and Y can be precisely inferred from $P(X)$.

More generally, Schölkopf et al. (2012) contend that $P(X)$ is "independent" of f if reverse causality does not exist (i.e., $Y \not\to X$). However, this independence may no longer hold when $Y \to X$. This notion of "independence" can be formalized mathematically in at least three distinct ways: (1) through Kolmogorov complexity (Janzing & Schölkopf, 2010), (2) by examining the uncorrelatedness between p and the derivative of f (Janzing & Schölkopf, 2015), and (3) through the uncorrelatedness between p and the logarithm of the derivative of f (Daniusis et al., 2010). Whereas the mathematical formulation of Schölkopf et al.'s (2012) insight necessarily varies with this underlying notion of independence—e.g., see Janzing and Schölkopf (2015, Lemma 1) for one variation—the two observations discussed before hold true regardless of this specific mathematical formulation.

3.2. Mathematical analysis of cross-lagged effects

To mathematically illustrate the link between causality testing and semi-supervised learning, we use the CLPM (Hamaker et al., 2015) as an example of the underlying data-generating process. This analysis aims to emphasize two key observations. First, in the absence of reverse causality (i.e., $Y \not\to X$), no information about the coefficient representing the $X \to Y$ relationship can be inferred from a dataset containing only X. Second, when reverse causality exists (i.e., $Y \to X$), a standard least squares estimate for the coefficient representing $Y \to X$ can be obtained even without Y in the dataset.

In its simplest form, CLPM assumes a model of structural equations:

$$(x_{it} - \kappa_i) = \alpha(x_{i,t-1} - \kappa_i) + \beta(y_{i,t-1} - \omega_i) + u_{it}, \tag{1}$$
$$(y_{it} - \omega_i) = \delta(y_{i,t-1} - \omega_i) + \gamma(x_{i,t-1} - \kappa_i) + v_{it}, \tag{2}$$

where x_{it} and y_{it} are the values of X and Y at time t, respectively; κ_i and ω_i represent the sample-specific random intercepts for X and Y, respectively; and u_{it} and v_{it} are i.i.d. random impulses. With this model, the cross-lagged parameter β captures the (within-person) reverse causal effect $Y \to X$.

As the sample-specific random intercepts κ_i and ω_i are extraneous to our ensuing analysis, we assume that $\kappa_i = \omega_i = 0$ for all $i \in [1, n]$ (where n is the sample size). For the ease of understanding, we also assume that all coefficients α, β, δ, and γ remain unchanged between time $t-1$ and $t+1$. These simplifying assumptions yield the following structural equations:

$$x_{it} = \alpha x_{i,t-1} + \beta y_{i,t-1} + u_{it}, \tag{3}$$
$$y_{it} = \delta y_{i,t-1} + \gamma x_{i,t-1} + v_{it}. \tag{4}$$

The crux of connecting semi-supervised learning with causality testing hinges on the following inquiry: Can any information about the regression coefficients in Equation 4 (i.e., δ and γ for predicting Y)

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

be inferred from observations of X at time $t-1$, t, and $t+1$ (i.e., $x_{1,t-1},\ldots,x_{n,t-1}$; $x_{1t},\ldots,x_{nt}$; $x_{1,t+1},\ldots,x_{n,t+1}$), but no observation of Y, assuming the other coefficients (i.e., α and β) as given?

When $\beta = 0$, this is obviously impossible because Equation 3 would only contain an autoregressive term $\alpha x_{i,t-1}$ and the remainder u_{it}, meaning that we would have no information about Y from any of the inputs X, α, and β.

When $\beta \neq 0$, however, knowledge of α and β would allow us to make a probabilistic inference of $y_{i,t-1}$ from Equation 3 based on x_{it} and $x_{i,t-1}$ and, similarly, y_{it} based on $x_{i,t+1}$ and x_{it}. Substituting these inferred values of $y_{i,t-1}$ and y_{it} into Equation 4 would then let us make a probabilistic inference of the regression coefficients δ and γ. To elucidate the reasoning behind this procedure, we rewrite Equation 3 as

$$x_{i,t+1} = \alpha x_{it} + \beta(\delta y_{i,t-1} + \gamma x_{i,t-1} + v_{it}) + u_{i,t+1}, \tag{5}$$
$$= \alpha x_{it} + \delta(\beta y_{i,t-1}) + \gamma\beta x_{i,t-1} + \beta v_{it} + u_{i,t+1}, \tag{6}$$
$$= \alpha x_{it} + \delta(x_{it} - \alpha x_{i,t-1} - u_{it}) + \gamma\beta x_{i,t-1} + \beta v_{it} + u_{i,t+1}, \tag{7}$$

which can be further simplified to

$$(x_{i,t+1} - \alpha x_{i,t}) = \delta(x_{it} - \alpha x_{i,t-1}) + \gamma\beta x_{i,t-1} + \varepsilon_i, \tag{8}$$

where $\varepsilon_i = \beta v_{it} + u_{i,t+1} - \delta u_{it}$ is a remainder term that aggregates various random impulses and satisfies $E(\varepsilon_i) = E(\varepsilon_i x_{it}) = E(\varepsilon_i x_{i,t-1}) = 0$. Equation 8 can be used to obtain a standard least squares estimate for δ and γ, i.e.,

$$\begin{bmatrix}\hat{\delta}\\ \hat{\gamma}\end{bmatrix} = \left(\left[\mathbf{x}_t - \alpha\mathbf{x}_{t-1}, \beta\mathbf{x}_{t-1}\right]^\top \left[\mathbf{x}_t - \alpha\mathbf{x}_{t-1}, \beta\mathbf{x}_{t-1}\right]\right)^{-1} \left[\mathbf{x}_t - \alpha\mathbf{x}_{t-1}, \beta\mathbf{x}_{t-1}\right]^\top (\mathbf{x}_{t+1} - \alpha\mathbf{x}_t), \tag{9}$$

where $\mathbf{x}_j$ represents the vector of X at time j. The variances of $\hat{\delta}$ and $\hat{\gamma}$ are

$$\mathrm{Var}(\hat{\delta}) = \sigma^2\left((\mathbf{x}_t - \alpha\mathbf{x}_{t-1})^\top(\mathbf{x}_t - \alpha\mathbf{x}_{t-1}) - \left((\mathbf{x}_t - \alpha\mathbf{x}_{t-1})^\top\mathbf{x}_{t-1}\right)^2\left(\mathbf{x}_{t-1}^\top\mathbf{x}_{t-1}\right)^{-1}\right)^{-1}, \tag{10}$$
$$\mathrm{Var}(\hat{\gamma}) = \sigma^2\left(\beta^2\mathbf{x}_{t-1}^\top\mathbf{x}_{t-1} - \left((\mathbf{x}_t - \alpha\mathbf{x}_{t-1})^\top\mathbf{x}_{t-1}\right)^2\left((\mathbf{x}_t - \alpha\mathbf{x}_{t-1})^\top(\mathbf{x}_t - \alpha\mathbf{x}_{t-1})\right)^{-1}\right)^{-1}, \tag{11}$$

where σ^2 is the variance of ε_i. Note from Equation 11 that the variance of $\hat{\gamma}$ tends to infinity when β approaches 0. This substantiates our prior assertion that estimating the coefficient for $X \to Y$ (i.e., γ) from X is only feasible when the coefficient for $Y \to X$ (i.e., β) is significant.

Given that δ and γ can be inferred (at least asymptotically) using just three inputs—α, β, and X—without any information on Y, a natural question is: which among the three inputs carries information about δ and γ? We can rule out α and β because they are free parameters (independent of δ and γ) per CLPM. This leaves the only possibility to be that the distribution of X carries certain information about the $X \to Y$ relationship (i.e., δ and γ). Importantly, this information carriage is valid if and only if reverse causality $Y \to X$ is present (i.e., $\beta \neq 0$).

At this juncture, the connection between semi-supervised learning and reverse causality testing becomes evident. Recall that a semi-supervised learning algorithm aims to learn the $X \to Y$ relationship $f(\cdot)$—i.e., γ in the case of CLPM—using a small labeled set of $\langle X, Y\rangle$ pairs and a large unlabeled set consisting solely of X. Clearly, the only information revealed by the unlabeled set is the distribution of X. When reverse causality does not exist (i.e., $\beta = 0$), the distribution of X does not carry any information about γ. In this case, the unlabeled set is useless for improving the prediction of Y, making the deployment of a semi-supervised learning algorithm futile. However, when reverse causality exists (i.e., $\beta \neq 0$), the distribution of X—and therefore the unlabeled set—does carry information about γ (as shown in Equation 9), making semi-supervised learning potentially fruitful.

For a more specific example, consider the aforementioned self-training method for semi-supervised learning. When the X–Y relationship is unidirectional flowing from X to Y (i.e., $\beta = 0$), expanding the labeled set with pseudo-labels will not lead to a more accurate prediction of Y for the simple reason that

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

even an infinitely large unlabeled set, which perfectly reveals $P(X)$, still contains no information about the $X \rightarrow Y$ relationship. Put simply, we can test reverse causality by comparing the predictive accuracy pre and post self-training, with an increase in predictive accuracy suggesting the existence of reverse causality.

3.3. Generalized test

Whereas we developed the connection between semi-supervised learning and reverse causality testing through the structural model of CLPM, the connection indeed persists regardless of the underlying data-generating process. To understand why, consider an extension of Equations 3 and 4 to data-generating processes beyond CLPM. When the X–Y relationship is unidirectional flowing from X to Y, we have

$$x_{it} = f_t(x_{i,t-1}, u_{it}), \tag{12}$$

$$y_{it} = g_t(x_{i,t-1}, y_{i,t-1}, v_{it}), \tag{13}$$

where f_t and g_t can be any arbitrary stochastic function.

Two key insights emerge from the equations. First, the $X \rightarrow Y$ relationship is wholly captured by g_t. Second, alterations in g_t has no influence on the value (and distribution) of X. Combining these two insights, it is evident that the distribution of X does not carry any information about the $X \rightarrow Y$ relationship unless there exists a reciprocal $Y \rightarrow X$ relationship. This underscores that the nexus between semi-supervised learning and reverse causality testing remains intact, independent of assumptions about the data-generating process.

4. Reverse causality testing using semi-supervised learning

In this section, we develop our novel method for reverse causality testing. As established in the last section, a semi-supervised learning algorithm can effectively leverage the large unlabeled (i.e., X-only) data to improve predictive accuracy for Y only when reverse causality $Y \rightarrow X$ is at play. Building upon this insight, our method revolves around assessing the predictive accuracy of semi-supervised learning. In the passages that follow, we outline our methodological design in two steps. First, we describe the input and output of semi-supervised learning, explaining in detail how the output of semi-supervised learning is used for reverse causality testing. Then, we delve into the algorithmic design of semi-supervised learning.

4.1. Input and output of semi-supervised learning

The objective of our method is to detect the presence of reverse causality, i.e., $Y \rightarrow X$, given a longitudinal panel dataset $D = \{\langle x_{i1}, y_{i1}, \ldots, x_{im}, y_{im}\rangle | i \in [1,n]\}$, where n is the sample size and m is the number of waves. Our method requires at least three waves (i.e., $m \geq 3$) for identification. It imposes no assumptions about the lag, provided that the lag is not so extensive that y_{i1} exerts no causal effect on x_{im}, as this would render reverse causality unidentifiable within the given dataset. Our method also makes no assumption about the functional form of the relationships between X and Y. The sole assumption underlying our method is the working condition of causal learning, which states that $P(X)$ is independent[4] of f if $X \rightarrow Y$ fully describes the relationship between X and Y (Schölkopf et al., 2012). Also note that, if the purpose is to test the existence of $X \rightarrow Y$ instead, the only revision required is to swap X and Y in the input data.

[4] As discussed earlier, see Daniusis et al. (2010) and Janzing & Schölkopf (2010, 2015) for formal definitions.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

4.1.1. Input

Recall from the literature review that a semi-supervised learning algorithm takes as input two datasets, a small labeled set D_{lb}, which consists of n_1 data points with paired X–Y values, and a large unlabeled set D_{ul}, which consists of n_2 data points ($n_2 \gg n_1$) with X values only. To properly specify D_{lb} and D_{ul} based on the input data D, there are three issues to be addressed.

First, since semi-supervised learning algorithms generally require the prediction target Y to be a scalar (i.e., a single variable), we need to select (the Y value of) one wave as the prediction target for semi-supervised learning. Mathematically, any wave except the last one would work because, as discussed earlier, if $Y \to X$ exists with a lagged effect, then the distribution of X in the tth wave carries information about Y in all previous waves (i.e., 1 to $t-1$). In other words, semi-supervised learning could effectively boost the predictive accuracy toward Y for all but the last wave. Since we are interested in minimizing the number of waves required for identification, a proper choice is to select Y in the first wave, i.e., $\{y_{i1}\}$, as the prediction target because otherwise data in the first wave would become useless for identification (as Y in a latter wave cannot have a causal effect on X in the first wave).

Second, we need to determine the variable composition of X. When $Y \to X$ exists, the distribution of X from the second wave onward is informative for predicting the value of Y from the first wave. We therefore include $\langle x_{i2}, \ldots, x_{im} \rangle$ as the predictor vector X.

Third, we also need to determine n_1 and n_2, the sample sizes for D_{lb} and D_{ul}, respectively. Recall that our purpose is to test whether semi-supervised learning can effectively leverage the unlabeled set D_{ul} to enhance predictive accuracy. Clearly, assuming $Y \to X$ exists, the smaller n_1 and the larger n_2 is, the more likely we would be able to detect the effectiveness of semi-supervised learning. From this perspective, a natural choice is to make n_1 the minimum labeled-sample size required by the semi-supervised learning algorithm, and to include all other data in the unlabeled set (i.e., $n_2 = n - n_1$). For example, the aforementioned self-training algorithm generally requires $n_1 \geq m$ to avoid an initial degenerate solution. Hence, we generate the labeled set D_{lb} and the unlabeled set D_{ul} by first randomly permuting the order of all data points in D, before setting

$$D_{\text{lb}} = \{\langle x_{i2}, \ldots, x_{im}, y_{i1} \rangle | i \in [1, m]\}, \text{ and} \tag{14}$$

$$D_{\text{ul}} = \{\langle x_{i2}, \ldots, x_{im} \rangle | i \in [m+1, n]\}. \tag{15}$$

4.1.2. Output

Our method for reverse causality testing focuses on the accuracy of f_{ssl}, the output of semi-supervised learning. To determine whether the unlabeled set D_{ul} is useful for improving the predictive accuracy of f_{ssl} (which is only possible when reverse causality exists), we compare f_{ssl} against a baseline model f_0, which is the initial model of semi-supervised learning generated from the small labeled set D_{lb} only *before* accessing the unlabeled set D_{ul}. A smaller error of f_{ssl} would serve as evidence against the null hypothesis of no reverse causality.

Recall from earlier discussions that D_{lb} and D_{ub} are drawn uniformly at random from the input data D. Since D_{lb} and D_{ub} are random samples, comparing the predictive errors of the machine learning models trained on them—i.e., f_0 and f_{ssl}, respectively—requires a statistical significance test. Specifically, we need to determine whether any observed reduction in predictive error of f_{ssl} relative to f_0 is statistically significant. To assess this, we repeatedly draw i.i.d. samples of D_{lb} from D, derive the corresponding D_{ub}, and compare the predictive error of f_{ssl} against f_0 for each sample. The statistical significance of the reduction in predictive error can then be evaluated based on these comparisons. There are three key issues worth discussing in this design: (1) how to measure the predictive error of a model, (2) how to assess the statistical significance of the reduction in predictive error, and (3) how to determine the number of resamples necessary for the test.

First, in terms of measuring the predictive error of a machine learning model, a prevalent practice in machine learning is to separate the testing dataset from the training data due to concerns of overfitting (Bishop & Nasrabadi, 2006). It is important to note that overfitting is *not* a concern in our case because

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

the prediction target y_{i1} for the vast majority of data (i.e., $i \in [m+1, n]$) is *hidden* from semi-supervised learning. As such, we can assess the predictive error of f_{ssl} (and the baseline f_0) directly over the input dataset D. Further, our method has no specific requirement on the metric for predictive error, as a reduction of any error metric indicates the existence of reverse causality. For example, when y_{i1} is binary, we could use the total number of prediction errors as the metric. The cross-entropy metric (Bishop & Nasrabadi, 2006) can be used for categorical y_{i1}. When y_{i1} is continuous, the error metric could be the mean squared error, mean absolute error, etc.

Second, for comparing the predictive errors of two machine learning models, Demšar (2006) reviewed three types of statistical tests used in the literature: (1) a parametric test, like the paired t-test (Dietterich, 1998), (2) a nonparametric test that makes assumptions about the distribution of difference in predictive accuracy, like the Wilcoxon signed-rank test (Santafe et al., 2015), and (3) a nonparametric test that only requires predictive accuracy to be comparable (i.e., at least on an ordinal scale), like the binomial test, also known as the sign test (Salzberg, 1997).

Problems with parametric tests such as the paired t-test have been well documented in the literature (Dietterich, 1998). As summarized by Demšar (2006), these tests suffer from sensitivity to outliers and the issue of commensurability between different runs, making it possible for the total failure of semi-supervised learning on a single sample of D_{lb} to dominate the test result. Similarly, key assumptions of the Wilcoxon signed-rank test do not hold in our context. For instance, it assumes that the difference in predictive accuracy between the two models is *symmetrically* distributed around a central value, yet there is no evidence to support this symmetry in our setting. In fact, semi-supervised learning is likely to amplify the skewness in the labeled set, suggesting that the difference could follow a heavy-tailed rather than symmetric distribution, violating the assumptions of the Wilcoxon signed-rank test. This leaves the (exact) binomial test (i.e., sign test) as a suitable alternative, as it requires only that each pair of predictive accuracy scores be comparable (Salzberg, 1997) and is inherently robust to outliers (Demšar, 2006). In our case, we employ the binomial test (with $\pi_0 = 0.5$) to compare the number of runs (i.e., D_{lb} samples) where f_{ssl} outperforms f_0 against the number of runs where the reverse is true.

Finally, two key factors influence the determination of the number of resamples required for the test. First, increasing the number of resamples enhances the statistical power of the test, which is especially critical given the well-documented limitation of the binomial test in terms of low statistical power (Demšar, 2006; Rainio et al., 2024; Salzberg, 1997). Second, a larger number of resamples incurs higher computational costs, as the semi-supervised learning algorithm must be executed for each sampled D_{lb} and D_{ul}. This computational burden can be particularly significant for resource-intensive algorithms, such as those based on deep learning (Goodfellow et al., 2016). To balance these considerations, we recommend following the default setting proposed by Demšar (2006) and performing 1,000 resamples. As demonstrated in the simulation studies, this setting offers a practical trade-off between statistical power and computational efficiency.

4.2. Design of semi-supervised learning algorithm

Numerous algorithms have been proposed for semi-supervised learning (Van Engelen & Hoos, 2020). We chose to implement the self-training paradigm discussed earlier. This choice is driven by two primary considerations. First, self-training is highly versatile, as it can be seamlessly integrated with a wide range of supervised learning algorithms as its base learner (Sohn et al., 2020). Second, the mathematical analysis of self-training is one of the few that have been thoroughly developed in the literature (Amini & Gallinari, 2002; Grandvalet & Bengio, 2004), providing a clear connection between the Bayes risk of its predictions and the foundational working condition of semi-supervised learning that has been discussed earlier in the article.

In the passages that follow, we describe the mathematical foundation and algorithmic design of self-training. Note that, our method represents a novel use of self-training for the purpose of reverse causality testing, and we did not make any change to its canonical algorithmic design. For the ease of discussion, we start with a simple setting where the prediction target y_{i1} is binary (i.e., $y_{i1} \in \{0, 1\}$)

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

and the underlying learning algorithm is logistic regression. At the end of this section, we describe a generalization to continuous variables and any regression algorithm.

4.2.1. Mathematical foundation of self-training

Any algorithm aiming to learn a prediction model that estimates a binary y_{i1} based on a predictor vector $\mathbf{x}_i = \langle x_{i2}, \ldots, x_{im} \rangle$ can be viewed as learning a function[5] $f(\mathbf{x}_i)$ that approximates $\Pr\{y_{i1} = 1 | \mathbf{x}_i\}$. For example, logistic regression specifies $f(\mathbf{x}_i)$ as

$$f(\mathbf{x}_i) = \text{sigmoid}(\beta_1 + \beta_2 \cdot x_{i2} + \cdots + \beta_m \cdot x_{im}), \tag{16}$$

where $\boldsymbol{\beta} = \langle \beta_1, \ldots, \beta_m \rangle$ are the coefficients to be estimated from training data.

Semi-supervised learning in general, and self-training in particular, starts by estimating $\boldsymbol{\beta}$ from the small labeled set D_{lb}. Specifically, it does so by finding $\boldsymbol{\beta}$ that maximizes the following log-likelihood function:

$$\ell(f) = \log P(\mathbf{Y}_{\text{lb}} | f, \mathbf{X}_{\text{lb}}) \tag{17}$$

$$= \sum_{i=1}^{m} (y_{i1} \cdot \log(f(\mathbf{x}_i)) + (1 - y_{i1}) \cdot \log(1 - f(\mathbf{x}_i))), \tag{18}$$

where $\mathbf{X}_{\text{lb}}$ and $\mathbf{Y}_{\text{lb}}$ represent the predictor (i.e., $x_{i2}, \ldots, x_{im}$) and prediction target (i.e., y_{i1}) portions of D_{lb}, respectively. In the Bayesian framework, these initial coefficient estimates form the maximum a posteriori (MAP) estimate under the uniform prior. As shown by Seeger (2000), this MAP estimate does *not* change if we merely add the unlabeled set D_{ul} into the observed data, because semi-supervised learning only works under the condition that the distribution of $\mathbf{x}_i$ reveals information about y_{i1}. In other words, in order to proceed beyond this initial step and generate the MAP estimate for semi-supervised learning, we need to encode its working condition into the prior distribution used to calculate the MAP estimate.

In the Bayesian framework, a common method for deriving the prior distribution from a given constraint (e.g., the working condition for semi-supervised learning) is the principle of maximum entropy (Jaynes, 1957), which sets the prior distribution as the one that satisfies the given constraint while having the maximum information entropy (Cover & Thomas, 2006). Grandvalet & Bengio (2004) followed this principle to prove that, for a semi-supervised learning algorithm that uses both D_{lb} and D_{ul}, the MAP estimate for $\boldsymbol{\beta}$ is the maximizer of the following criterion $C(f)$,

$$C(f) = \ell(f) - \lambda \cdot \sum_{i=m+1}^{n} (-f(\mathbf{x}_i) \cdot \log(f(\mathbf{x}_i)) - (1 - f(\mathbf{x}_i)) \cdot \log(1 - f(\mathbf{x}_i))), \tag{19}$$

where λ ($\lambda > 0$) is the Lagrange multiplier that, roughly speaking, captures the amount of information about y_{i1} that can be revealed by the distribution of $\mathbf{x}_i$. Amini & Gallinari, 2002 analyzed self-training with logistic regression as the underlying learning algorithm, and proved that the resulting coefficient estimates maximizes $C(f)$ when $\lambda = 1$. In other words, semi-supervised learning in general, and self-training in particular, can be viewed as approximating the MAP estimate of f under the working condition of semi-supervised learning. Specifically, it does so by finding f that maximizes $C(f)$ in Equation 19.

4.2.2. Algorithmic design of self-training

The algorithmic design of self-training can be readily derived from Equation 19. Note that the second term in the equation is proportional to the sum of

$$H(f(\mathbf{x}_i)) = -f(\mathbf{x}_i) \cdot \log(f(\mathbf{x}_i)) - (1 - f(\mathbf{x}_i)) \cdot \log(1 - f(\mathbf{x}_i)) \tag{20}$$

[5] For simplicity in notation, this section uses f to denote the machine learning estimate, replacing the earlier notation $\hat{f}$. This causes no ambiguity, as the ground-truth value of f is not discussed in this section.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

for all data points in the unlabeled set D_{ul}. Recall that $f(\mathbf{x}_i)$ is the prediction from semi-supervised learning for $\Pr\{y_{i1} = 1|\mathbf{x}_i\}$. This makes $H(f(\mathbf{x}_i))$ in Equation 20 the entropy (Cover & Thomas, 2006) of the predicted distribution of y_{i1} given $\mathbf{x}_i$, meaning that it captures the amount of *uncertainty* in machine learning predictions. For example, we have $H(f(\mathbf{x}_i)) = 0$, its minimum possible value, when semi-supervised learning is fully confident of its prediction (i.e., $f(\mathbf{x}_i) = 0$ or 1). In contrast, when $f(\mathbf{x}_i) = 0.5$ (i.e., maximum uncertainty), $H(f(\mathbf{x}_i))$ reaches its maximum value[6] of 1.

With this understanding, the goal of self-training—i.e., the maximization of $C(f)$—pertaining to the unlabeled set is equivalent with minimizing prediction uncertainty (i.e., $H(f(\mathbf{x}_i))$) for the unlabeled data. Self-training does so in an iterative manner. As described earlier, the initial iteration estimates $\boldsymbol{\beta}$ from the labeled set D_{lb} by applying the underlying learning algorithm, which in this case is the standard logistic regression. The estimated $\boldsymbol{\beta}$ is then used to compute $f(\mathbf{x}_i)$, and thereby $H(f(\mathbf{x}_i))$, for all data points in the unlabeled set D_{ul}. The first iteration concludes by adding into the labeled set D_{lb} all unlabeled data points with prediction uncertainty below a pre-determined threshold h, using their predicted labels as if they were real. In other words, the labeled and unlabeled sets are updated as

$$D_{\text{lb}} := D_{\text{lb}} \cup \{\langle \mathbf{x}_i, \mathbb{1}_{f(\mathbf{x}_i)\geq 0.5}\rangle | \mathbf{x}_i \in D_{\text{ul}}, H(f(\mathbf{x}_i)) \leq h\}, \tag{21}$$

$$D_{\text{ul}} := \{\mathbf{x}_i | \mathbf{x}_i \in D_{\text{ul}}, H(f(\mathbf{x}_i)) > h\}, \tag{22}$$

where $\mathbb{1}_{f(\mathbf{x}_i)\geq 0.5}$ is the predicted label for $\mathbf{x}_i$—i.e., an indicator function that returns 1 if $f(\mathbf{x}_i) \geq 0.5$ and 0 otherwise. The updated D_{lb} and D_{ul} are then entered as input to the next iteration. The iterative process ends when no new X–Y pair is added to D_{lb} after an iteration. As can be seen from the description, self-training minimizes prediction uncertainty among unlabeled data using the idea of *pseudo-labels*, i.e., by promoting those with minimal prediction uncertainty to the labeled set, using their predicted labels as if they were real. These promoted data points, in turn, reduce prediction uncertainty for the remaining unlabeled data (Amini & Gallinari, 2002), pushing the coefficient estimates closer to their MAP values that maximize $C(f)$.

4.2.3. Generalization to continuous variables

Whereas the above description of self-training is based on a binary y_{i1} and logistic regression being the underlying learning algorithm, the same iterative process can be adapted to support continuous y_{i1} and any underlying learning algorithm. This adaption requires addressing two issues. One is the design of an appropriate uncertainty measure (i.e., $H(f(\mathbf{x}_i))$ in the binary case), and the other is the generation of pseudo-label (i.e., $\mathbb{1}_{f(\mathbf{x}_i)\geq 0.5}$ in the binary case). The reason why these two issues arise for continuous variables is because, unlike in the binary case where $f(\mathbf{x}_i)$, as an estimate of $\Pr\{y_{i1} = 1|\mathbf{x}_i\}$, inherently captures uncertainty about the predicted y_{i1}, a point-estimate for a continuous y_{i1} contains no such uncertainty information. Therefore, instead of deriving both the uncertainty measure and the pseudo-label from $f(\mathbf{x}_i)$ itself, we may have to resort to other information in order to do so in the continuous case.

A well-known method for addressing both issues in semi-supervised learning is co-training (Blum & Mitchell, 1998). With this method, instead of generating a single prediction of $f(\mathbf{x}_i) \approx y_{i1}$, we generate two predictions, f_1 and f_2, based on two (slightly) different subsets of variables in $\mathbf{x}_i$. Then, the difference between f_1 and f_2 (i.e., $|f_1 - f_2|$) is a natural measure of uncertainty, while the mean of the two (i.e., $(f_1 + f_2)/2$) can be used as pseudo-label for expanding the labeled set.

More specifically, recall that our method includes in $\mathbf{x}_i$ a total of $m - 1$ variables $\langle x_{i2}, \ldots, x_{im}\rangle$, where $m \geq 3$ because we require a minimum of three waves for identification. A natural choice is to associate f_1 with the first $m - 2$ variables $\mathbf{x}_i^{\text{F}} = \langle x_{i2}, \ldots, x_{i,m-1}\rangle$, and f_2 with the last $m - 2$ variables $\mathbf{x}_i^{\text{L}} = \langle x_{i3}, \ldots, x_{im}\rangle$. This way, we allow a divergence of two predictions (to allow the uncertainty estimate) while minimizing the number of predictors withheld from either. With this design, the updates of labeled and unlabeled

[6] Without loss of generality, we follow the convention in computer science to assume a base of 2 for all log operations, so as to measure entropy in binary bits.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Table 1. Pseudocode for reverse causality testing with continuous variables

Algorithm 1: Reverse Causality Testing Using Self-Training

Data: A longitudinal dataset $D = \{\langle x_{i1}, y_{i1}, \ldots, x_{im}, y_{im}\rangle | i \in [1,n]\}$
Input: Uncertainty threshold h; Number of iterations r; A base supervised learner Γ that outputs $\hat{f} : X \to Y$ from an input labeled dataset
Result: p-value from binomial exact test (with $\pi_0 = 0.5$)

$c \leftarrow 0$;
for *r iterations* **do**
  Randomly permute the order of all n data points in D;
  $D_{\text{lb}} \leftarrow \{\langle x_{i2}, \ldots, x_{im}, y_{i1}\rangle | i \in [1,m]\}$;
  $D_{\text{ul}} \leftarrow \{\langle x_{i2}, \ldots, x_{im}\rangle | i \in [m+1,n]\}$;
  $f_0 \leftarrow \Gamma(D_{\text{lb}})$;
  repeat
    Remove the $(m-1)$-th column (i.e., $x_{\cdot m}$) from D_{lb} to form $D^1_{\text{lb}} = \{\langle \mathbf{x}^{\text{F}}_i, y_{i1}\rangle\}$;
    Remove the first column (i.e., $x_{\cdot 2}$) from D_{lb} to form $D^2_{\text{lb}} = \{\langle \mathbf{x}^{\text{L}}_i, y_{i1}\rangle\}$;
    $f_1 \leftarrow \Gamma(D^1_{\text{lb}})$; $f_2 \leftarrow \Gamma(D^2_{\text{lb}})$;
    $D_{\text{lb}} \leftarrow D_{\text{lb}} \cup \{\langle \mathbf{x}_i, (f_1(\mathbf{x}^{\text{F}}_i) + f_2(\mathbf{x}^{\text{L}}_i))/2\rangle | \mathbf{x}_i \in D_{\text{ul}}, |f_1(\mathbf{x}^{\text{F}}_i) - f_2(\mathbf{x}^{\text{L}}_i)| \le h\}$;
    $D_{\text{ul}} \leftarrow \{\mathbf{x}_i | \mathbf{x}_i \in D_{\text{ul}}, |f_1(\mathbf{x}^{\text{F}}_i) - f_2(\mathbf{x}^{\text{L}}_i)| > h\}$;
  until D_{lb} *and* D_{ub} *remain unchanged*;
  $f_{\text{ssl}} \leftarrow \Gamma(D_{\text{lb}})$;
  if f_{ssl} *has a smaller mean squared error over* D *than* f_0 **then**
    $c \leftarrow c + 1$;
  end
end
return $\sum_{i=c}^{r} \binom{r}{i} \cdot 0.5^i \cdot 0.5^{r-i}$

set in each iteration become

$$D_{\text{lb}} := D_{\text{lb}} \cup \{\langle \mathbf{x}_i, (f_1(\mathbf{x}^{\text{F}}_i) + f_2(\mathbf{x}^{\text{L}}_i))/2\rangle | \mathbf{x}_i \in D_{\text{ul}}, |f_1(\mathbf{x}^{\text{F}}_i) - f_2(\mathbf{x}^{\text{L}}_i)| \le h\}, \tag{23}$$

$$D_{\text{ul}} := \{\mathbf{x}_i | \mathbf{x}_i \in D_{\text{ul}}, |f_1(\mathbf{x}^{\text{F}}_i) - f_2(\mathbf{x}^{\text{L}}_i)| > h\}, \tag{24}$$

whereas everything else in the iterative process follows directly from the binary case. Clearly, this design for continuous y_{i1} is compatible with any underlying learning algorithm for generating f_1 and f_2, e.g., linear regression (Stine, 1985), support vector machines (SVMs; De Brabanter et al., 2010), neural networks (Heskes, 1996), etc. The pseudocode of our algorithm for continuous variables is available in Table 1.

4.3. Transparency and openness

The complete code implementation of our algorithm is publicly available at https://github.com/calearn/revc (Python) and https://github.com/calearn/revc_m (MATLAB).

5. Simulation studies

We conducted two simulation studies. The main study evaluates the statistical power of our method in identifying reverse causality, and a followup study assesses the Type I error rate of our method in the absence of reverse causality, because existing methods such as the random-intercept CLPM (RI-CLPM; Hamaker et al., 2015) are known to generate Type I errors.

5.1. Data-generating process

In the main simulation study, we aimed to delineate the primary factors influencing the statistical power of our method. To achieve this, we employed a straightforward structural model-Equations 1

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

and 2—as the data-generating process, establishing a clear ground truth for the causal direction. To ensure a comprehensive analysis, we created multiple levels for three key parameters in the data-generating process: the total number of observations N and the cross-lagged parameters β (i.e., $Y \to X$) and γ (i.e., $X \to Y$). For the sample size N, we created four levels: 100, 250, 500, and 1,000. For the cross-lagged parameters β and γ, we created six levels for each: 0, 0.1, 0.2, 0.3, 0.4, and 0.5. In total, the design for the main simulation study consists of 4 (N) × 6 (β) × 6 (γ) = 144 unique conditions.

We adopted a standard parameter setup (e.g., Hamaker et al., 2015) to configure the other fixed parameters for the data-generating process in the main simulation study. Specifically, we followed Hamaker et al. (2015) to set the autoregressive parameters for both variables to $\alpha = \delta = 0.5$. All random impulses u_{it} and v_{it} were generated from a Gaussian distribution $\mathcal{N}(0, 0.5^2)$, while the sample-specific random intercepts were fixed at $\kappa_i = \omega_i = 0$ for all $i \in [1, N]$. To generate the initial (Wave 1) values of x_{i1} and y_{i1}, we deliberately used different distributions[7] to emphasize that our method does not rely on specific distributional assumptions for the input data. For each $i \in [1, N]$, x_{i1} was sampled uniformly at random from the interval $[-3, 3]$, while y_{i1} was drawn from a Gaussian distribution $\mathcal{N}(0, 1)$.

In the followup simulation study, we followed Lucas (2023), which demonstrates the spurious cross-lagged effects generated by CLPM, in adopting the widely used Stable Trait Autoregressive Trait and State (STARTS) model (Kenny & Zautra, 1995) as the underlying data-generating process. We also followed Lucas (2023) in the parameter setup, specifically by setting (for both X and Y) the stability parameter as 0.5, the random intercept variance as 1, and variance of autoregressive component as 1. We also added a measurement error of variance 0.3 to both X and Y. Since the focus of the followup study is on Type I errors, the cross-lagged parameter for reverse causality ($Y \to X$) was always set to zero, while the parameter for $X \to Y$ (represented as γ for consistency with the main study) was varied from 0.1 to 0.5. Like in the main study, we created four levels, 100, 250, 500, and 1,000, for the sample size N. In total, the design for the followup simulation study consists of 4 (N) × 5 (γ) = 20 unique conditions.

5.2. Algorithmic implementations

Recall from earlier discussions that our method applies to both discrete and continuous Y, and can be used with any supervised learning algorithm as its base learner. To demonstrate the versatility of our method, we implemented two variants of it. The first was designed for continuous Y and implemented in MATLAB R2023b (with statistics and machine learning toolbox). We selected a simple design, i.e., OLS regression, as the underlying learning algorithm. We set the number of resamples for binomial test to be 1,000, leading to a computational overhead of about 20 seconds per simulation run on a laptop computer with Apple M2 CPU and 8GB RAM. For the number of waves taken as input, we set $m = 3$ (i.e., the first three waves), the minimum value that satisfies the identification requirement of our method while providing a conservative estimate of its statistical power. For the uncertainty threshold h (i.e., the maximum difference in prediction between the two models for an unlabeled data point to be added to the labeled set, see Equation 24), after testing a wide range of threshold values, we found that the output of our method is insensitive to the threshold setting as long as it is neither too large—so as to allow all unlabeled data to enter the labeled set at once—nor too small to permit any unlabeled data point into the labeled set. Due to this finding, we set the threshold to a constant of $h = 0.1$, which is about 10% of the standard deviation of the label (i.e., Y).

Since the OLS regression algorithm relies on a linear model and our data-generating process is also linear, concerns may arise regarding a potential unfair advantage due to their inherent consistency. To address these concerns, we implemented a variant of our method using a nonlinear learning model. Specifically, for the underlying base learner in self-training, we employed the SVM (Cortes & Vapnik, 1995) algorithm with a polynomial kernel of degree 3 (i.e., a cubic kernel SVM).

Given the significantly higher computational overhead of SVM, we implemented this variant in Python using the scikit-learn (Pedregosa et al., 2011) machine learning library. We adhered to the

[7] We also tested alternative distributions (e.g., Gaussian mixtures) and observed no qualitative differences.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

default settings of scikit-learn's LIBSVM implementation (Chang & Lin, 2011) for all SVM parameters, including the kernel configuration (i.e., in function `sklearn.svm.SVC`). For the uncertainty threshold in the self-training algorithm, we used the built-in options for the `criterion` parameter in `sklearn.semi_supervised.SelfTrainingClassifier`. Apart from the use of SVM as the base learner, all other design elements were identical to the OLS implementation, except for an additional input data normalization process required for the SVM implementation. Specifically, Y was converted to binary values (using median dichotomization), and X was normalized to the range $[-1, 1]$ (through min-max scaling) to account for SVM's sensitivity to input feature scaling (Chang & Lin, 2011; Tax & Duin, 2004).

To accommodate the computational demands of the nonlinear SVM algorithm and the large number of simulation runs required for the main study, this implementation was executed on Amazon Web Services (AWS) Batch using Elastic Container Service (ECS) clusters configured with 256 virtual CPUs, provisioned and scaled automatically using AWS Fargate. With the same number of resamples (1,000) as the OLS implementation, the SVM variant completed each simulation run in approximately 30 seconds on the AWS cluster.

5.3. Simulation results

For the main study, Table 2 presents the statistical power achieved by our method under a significance level of $\alpha = .05$ for all simulation settings with $\beta > 0$. For settings where $\beta = 0$ (highlighted in gray), the table reports the Type I error rates. Each cell in the table is based on 1,000 independent runs of our method. Note that results for the nonlinear SVM-based implementation are included only for $N = 1{,}000$, as the SVM-based implementation required more than 500 samples to consistently converge to reliable predictions.

We can draw three key observations from the table. First, regarding Type I error rates (highlighted rows with $\beta = 0$), our method consistently remains below 0.075 across all simulation conditions, aligning with Bradley's (1978) liberal robustness criterion. This demonstrates the robustness of our method against finding spurious reverse causal effects.

Second, note that the top six rows of Table 2, along with the highlighted rows, represent simulation conditions where X and Y exhibit a unidirectional relationship. The statistical power achieved by our method in these scenarios highlights its effectiveness in identifying the direction of a unidirectional effect—a critical use-case for our method when competing theories stipulate different causal directions between X and Y. For example, Rows 3–6 show that, when $\beta \geq 0.2$, the OLS implementation of our method achieves statistical power exceeding 0.97 when $N \geq 500$, while the SVM implementation achieves the same when $N = 1{,}000$.

Third, the statistical power of our method is influenced by three key factors. One factor is N, the input sample size. Larger sample sizes significantly improve statistical power. For example, when $\beta = 0.3$ and $\gamma = 0$, the OLS implementation achieves a statistical power of only 0.44 with $N = 100$ but exceeds 0.99 when $N \geq 500$. Another factor is β, the cross-lagged parameter representing the strength of the reverse causal effect. When $\gamma = 0$ (i.e., unidirectional effect), larger β values correspond to higher statistical power. For instance, with $N = 250$ and $\gamma = 0$, the OLS implementation achieves a power of 0.19 at $\beta = 0.1$, increasing to 0.77 at $\beta = 0.3$, and exceeding 0.99 for $\beta \geq 0.4$. The SVM implementation exhibits a similar trend. The final factor is γ, the strength of the causal effect $X \rightarrow Y$. The impact of γ on statistical power varies depending on other parameters. For example, when $\beta = 0.5$ and $N = 1{,}000$, a stronger forward effect of $\gamma = 0.5$ can obscure reverse causality, reducing power from 1.00 at $\gamma = 0$ to 0.86 at $\gamma = 0.5$. Conversely, when γ and β are smaller yet closer in magnitude (e.g., $\gamma = 0.2$, $\beta = 0.2$, $N = 250$), γ could potentially enhance power, increasing it from 0.30 at $\gamma = 0$ to 0.67 at $\gamma = 0.2$. This increase can be attributed to the technical design of the two-step self-training algorithm in our method. Specifically, in situations where the causal effect $X \rightarrow Y$ is nonexistent (i.e., $\gamma = 0$), the initial step of self-training, which learns from the labeled set only, is destined to yield highly inaccurate predictions. In such scenarios, particularly with smaller sample sizes, it becomes challenging for the subsequent step (of learning from

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Table 2. Type I error rate and statistical power of our method in the main simulation study

		OLS				SVM
β	γ	$N = 100$	$N = 250$	$N = 500$	$N = 1000$	$N = 1000$
0.0	0.0	0.005	0.003	0.037	0.000	0.017
0.1	0.0	0.006	0.188	0.335	0.448	0.518
0.2	0.0	0.091	0.301	0.976	0.972	1.000
0.3	0.0	0.440	0.774	0.998	1.000	1.000
0.4	0.0	0.593	0.999	0.999	1.000	1.000
0.5	0.0	0.820	1.000	1.000	1.000	1.000
0.0	0.1	0.001	0.058	0.002	0.000	0.032
0.1	0.1	0.000	0.018	0.557	0.421	0.542
0.2	0.1	0.092	0.636	0.800	0.719	1.000
0.3	0.1	0.362	0.589	0.988	0.986	1.000
0.4	0.1	0.503	0.994	0.998	1.000	1.000
0.5	0.1	0.856	0.908	1.000	0.999	0.990
0.0	0.2	0.073	0.000	0.013	0.000	0.018
0.1	0.2	0.012	0.073	0.469	0.886	0.546
0.2	0.2	0.202	0.669	0.763	0.982	1.000
0.3	0.2	0.373	0.793	0.833	0.997	1.000
0.4	0.2	0.618	0.990	0.995	1.000	0.986
0.5	0.2	0.731	0.943	0.935	0.992	0.841
0.0	0.3	0.050	0.000	0.071	0.007	0.036
0.1	0.3	0.004	0.179	0.702	0.575	0.487
0.2	0.3	0.157	0.311	0.586	0.816	1.000
0.3	0.3	0.306	0.816	0.999	0.961	0.997
0.4	0.3	0.373	0.855	0.979	0.998	0.909
0.5	0.3	0.309	0.879	0.914	0.975	0.510
0.0	0.4	0.000	0.004	0.001	0.037	0.048
0.1	0.4	0.008	0.052	0.323	0.475	0.456
0.2	0.4	0.277	0.320	0.688	0.933	1.000
0.3	0.4	0.446	0.615	0.896	0.939	0.993
0.4	0.4	0.401	0.726	0.882	0.863	0.725
0.5	0.4	0.475	0.575	0.838	0.792	0.370
0.0	0.5	0.058	0.057	0.040	0.001	0.037
0.1	0.5	0.003	0.018	0.144	0.343	0.504
0.2	0.5	0.255	0.429	0.648	0.740	0.997
0.3	0.5	0.240	0.501	0.931	0.930	0.966
0.4	0.5	0.243	0.486	0.704	0.879	0.668
0.5	0.5	0.590	0.347	0.840	0.856	0.432

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Table 3. Type I error rate of RI-CLPM and our method under STARTS model

		$\gamma = 0.1$	$\gamma = 0.2$	$\gamma = 0.3$	$\gamma = 0.4$	$\gamma = 0.5$
RI-CLPM	$N = 100$	0.055	0.077	0.099	0.139	0.186
	$N = 250$	0.066	0.096	0.174	0.262	0.352
	$N = 500$	0.073	0.148	0.277	0.445	0.614
	$N = 1000$	0.106	0.254	0.478	0.713	0.883
Our Method	$N = 100$	0.001	0.002	0.012	0.006	0.002
	$N = 250$	0.001	0.000	0.004	0.004	0.013
	$N = 500$	0.022	0.055	0.038	0.022	0.018
	$N = 1000$	0.003	0.015	0.012	0.007	0.016

the unlabeled set) to significantly enhance predictive accuracy even in the presence of reverse causality. Therefore, a larger γ value, which improves the accuracy of the initial step, also tends to boost the statistical power.

Interestingly, the negative impact of γ on the statistical power of our method appears more pronounced in the SVM implementation compared to the OLS implementation. For example, with $N = 1{,}000$, the OLS implementation maintains a statistical power of at least 0.70 for $\beta \geq 0.2$, regardless of γ. In contrast, the SVM implementation's power drops from 1.00 at $\gamma = 0$ to 0.43 at $\gamma = 0.5$ for $\beta = 0.5$. We attribute this discrepancy to SVM's sensitivity to feature scaling (Tax & Duin, 2004) and vulnerability to outliers (Debruyne, 2009), both of which are exacerbated at larger γ.

Table 3 summarizes the results of the follow-up simulation study comparing the Type I error rates of RI-CLPM and the OLS implementation of our method.[8] Following Mulder & Hamaker (2021), RI-CLPM was implemented using default settings in Mplus.[9] The results reveal that RI-CLPM frequently identifies spurious reverse causal relationships, with Type I error rates reaching 0.88 when $N = 1{,}000$ and $\gamma = 0.5$. In general, RI-CLPM's Type I error rate increases with larger N and higher γ. In contrast, our method maintains a Type I error rate below 0.05 across all conditions, again aligning with Bradley's (1978) liberal robustness criterion and demonstrating an advantage of our method over CLPM and its variants, which are known to find spurious cross-lagged effects (Lucas, 2023).

6. A case study

We applied our method over a real-world panel dataset to demonstrate its value for reverse causality testing in practice. Specifically, we examined the relationship between work-family conflict (WFC)—i.e., "a form of inter-role conflict in which the role pressures from the work and family domains are mutually incompatible in some respect" (Greenhaus & Beutell, 1985, p. 77)—and job satisfaction (JAS)—i.e., the "overall evaluative judgement one has about one's job" (Judge et al., 2017, p. 357). We chose this focal relationship for two main reasons. First, like many constructs in industrial-organizational psychology, it is often practically difficult, if not ethically dubious, to manipulate WFC or JAS in a randomized controlled trial. Second, whereas a robust correlation between WFC and JAS has been widely recognized (Allen et al., 2020; Amstad et al., 2011), there are ongoing debates about the causal direction between the two constructs. Some posit a unidirectional effect of WFC negatively affecting JAS (e.g., Allen et al., 2020; Amstad et al., 2011; Kossek & Ozeki, 1998). Others contend that the causal influence flows in the

[8] We also ran the SVM implementation and found qualitatively identical results.

[9] We also implemented RI-CLPM in R with the `lavaan` package, replicating the setup from Mulder & Hamaker (2021). Results were consistent with those reported in Table 3.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

opposite direction, as higher JAS leads to greater work-life balance and, correspondingly, lower WFC (Landolfi et al., 2022). Yet others suggest the existence of reciprocal effects, with WFC and JAS affecting each other over time (e.g., Demerouti et al., 2004). Given these varied viewpoints, we sought to apply our method to empirically test the existence of causal effect in either direction.

6.1. Data

We drew from the Swiss Household Panel (SHP Group, 2023; Voorpostel et al., 2023) to examine the WFC–JAS relationship. The SHP began in 1999 with a nationally representative sample of 5,074 Swiss Households, introducing supplementary samples in 2004 (addition of 2,537 households), 2013 (addition of 3,989 households), and 2020 (addition of 4,380 households; Voorpostel et al., 2023). All household members aged 14 and above are surveyed annually via telephone and written surveys. We mirrored the timeframe used in prior work on CLPM and analyzed data collected in Waves 6 through 9 (i.e., the annual surveys between 2004 and 2007; Ozkok et al., 2022). We excluded participants who did not work or who did not complete a single survey in our timeframe, resulting in N = 7,748. Participants (51.55% women, 48.45% men) were on average 39.28 (SD = 14.50) years old. On average, participants had 12.98 (SD = 3.20) years of education.

In the case study, WFC was measured by a single item on a 11-point Likert scale (0 = not at all, 10 = very strongly). The item read "How strongly does your work interfere with your private activities and family obligations more than you would want this to be?" JAS was measured by a six-item measure created for the SHP on a 11-point Likert scale (0 = not at all satisfied, 10 = completely satisfied). The lead-in to the items was "Can you indicate your degree of satisfaction for each of the following points?" Sample items include "your job in general" and "the amount of your work." The internal consistency for the four waves was .78, .78, .79, and .79, respectively.

6.2. Results

In terms of algorithmic implementations, we used the same implementations (OLS and SVM) as the simulation study, and set the number of resamples to 10,000 and 100, respectively. Note that we intentionally set the number of resamples for the OLS implementation to be much higher than required in order to support a detailed analysis described later. With either implementation, we first tested the presence of WFC → JAS by setting the self-training algorithm to predict WFC from JAS, before testing the presence of JAS → WFC by setting the self-training algorithm to predict JAS from WFC. For the OLS implementation, self-training was effective for predicting WFC from JAS in 5,743 out of 10,000 resamples (one-tailed binomial exact test with $\pi_0 = 0.5$: $p < 10^{-6}$), and was effective for predicting JAS from WFC in 5,006 out of 10,000 resamples (one-tailed binomial exact test with $\pi_0 = 0.5$: $p = 0.4562$). For the nonlinear SVM implementation, the results were consistent with the OLS implementation. Specifically, for the prediction of WFC from JAS, self-training was effective in 67 out of 100 iterations, leading to a p-value of $p < 0.0005$. On the flip side, when predicting JAS from WFC, self-training was effective only in 37 out of 100 runs,[10] leading to a p-value of $p = 0.9967$. In sum, the results suggest that work-family conflict influences JAS, but not vice versa.

To further inspect *why* semi-supervised learning succeed in predicting WFC but not JAS, we examined how its effectiveness depends on the small labeled set, which is the only information source about Y that semi-supervised learning receives. Recall from the design of self-training that, if the distribution of X reveals signals about the X–Y relationship, we would expect self-training to be *more* effective when the labeled set is a more representative sample of the full dataset. This is because, when the labeled set is a severely biased sample, say featuring only a single value of Y, then self-training would stand no chance in becoming effective, as it would not even know what the other values of Y might be.

[10]Note that this means the use of unlabeled set actually *increased* predictive error in the majority (63%) of runs, likely due to the aforementioned sensitivity of SVM to outliers in the (uninformative) unlabeled set (Debruyne, 2009).

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

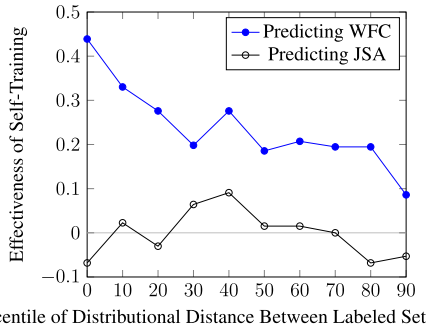

Figure 2. Relationship between effectiveness of self-training and representativeness of labeled set. *Note*: WFC, work-family conflict; JSA, job satisfaction. *Source*: Swiss Household Panel (SHP).

To inspect how the effectiveness of semi-supervised learning varies with the bias of labeled sample, we leveraged the 10,000 resample runs executed for the OLS implementation. Specifically, we sorted all runs according to the distributional distance (as measured by Kolmogorov–Smirnov test statistic; Daniel, 1990) between the labeled set and the full dataset, before stratifying them into 10 equi-sized bins (i.e., each containing 1,000 runs) according to the sorted order and calculating the average effectiveness of semi-supervised learning for each bin as measured by the fraction of runs that improve predictive accuracy minus the fraction of those that reduce accuracy. As discussed before, if self-training were ineffective, we would expect the average effectiveness measure to hover around zero for all ten bins. In contrast, if self-training were effective, we would expect the average effectiveness to be high when the distributional distance is small, and gradually decrease when the distributional distance becomes larger.

As can be seen from Figure 2, when self-training is used to predict WFC (i.e., to test the presence of WFC → JAS), there is a clear, negative, correlation between its effectiveness and the distributional distance. In contrast, such correlation disappears in the prediction of JSA (i.e., testing of JAS → WFC), as the effectiveness of self-training hovers around zero regardless of the distributional distance. This clearly shows that, whereas the distribution of JSA reveals valuable signals for predicting WFC—meaning that WFC may have a causal effect on JAS—the distribution of WFC carries no information for predicting JAS, suggesting that the causal direction is unlikely to flow from JAS to WFC.

7. General discussion

In this section, we first discuss the research implications of our new method. We then review the limitations of our method and the future research needed to address them.

7.1. Research implications

For testing reverse causality, the main research implication of our semi-supervised learning based method is its ability to identify causal directions without presuming distributional or functional features of the data-generating process. Our method achieves this by leveraging recent advances in machine learning that directly link the causal direction with the effectiveness of semi-supervised learning. A unique feature of our method is that it is an umbrella algorithm independent of the underlying learning algorithm, which can be any algorithm for supervised learning. As such, researchers can freely choose a learning algorithm that fits the type and scale of their data before using our method for causal identification.

More broadly, the main research implication of our work is its transfer of insights of causal learning (Peters et al., 2017), in particular Schölkopf et al.'s (2012) seminal finding that links causal direction with the effectiveness of semi-supervised learning, into the methodological arsenal of causal inference

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

in psychology. Whereas most existing work in causal learning—perhaps owing to the disciplinary focus of machine learning—leverages this link to explain (Schölkopf et al., 2012) or improve (Kügelgen et al., 2019) the performance of machine learning algorithms, our contribution is to demonstrate that the exact same link can be used to develop a concrete method for identifying reverse causality, addressing long-standing challenges in the analysis of longitudinal data in psychology and other disciplines.

7.2. Limitations and future directions

7.2.1. Capability of semi-supervised learning

The statistical power of our method depends on the capability of the semi-supervised learning algorithm in approximating the functional relationship between X and Y. On the one hand, this allows our method to generalize beyond linearity to allow nonlinear X–Y relationships. On the other hand, it also brings about the limitation that, if the semi-supervised learning algorithm being used cannot effectively approximate a nonlinear relationship between X and Y, our method may generate Type II errors.

In the main simulation study in our current work, we employed a linear model (i.e., CLPM) as the underlying data-generating process. This choice allowed us to concentrate on the overarching design of our method rather than on fine-tuning the underlying learning algorithm, given that even a simple algorithm (like OLS) would likely suffice for a linear relationship. Nonetheless, although the theoretical foundations of our method extend readily to more complex, nonlinear relationships between X and Y, we did not evaluate the performance of our method in the presence of such relationships. As a natural next step, future research could provide a comprehensive empirical assessment of the statistical power of our method across a wider range of data-generating processes, encompassing both linear and nonlinear scenarios beyond those considered in this study. Additionally, exploring various underlying learning algorithms and other semi-supervised learning designs will be valuable for capturing the inherent complexity of nonlinear relationships between X and Y.

In the longer term, recent advancements in machine learning offer two key insights. First, there are semi-supervised learning algorithms that offer *universal approximation* (Goodfellow et al., 2016), meaning that they are theoretically capable of approximating any arbitrary function in a Euclidean space. Examples include the use of our method (i.e., self-training) with deep neural networks (Hornik et al., 1989) or variational Gaussian processes (Tran et al., 2016) as the underlying learning algorithm. Second, unfortunately, one would have to restrict the type of relationship between X and Y in order to translate such theoretical feasibility into any practical guarantee. To understand why, consider a stylized example where Y is the encrypted value (i.e., ciphertext) of X based on a secret key. Theoretically, it is feasible for a machine learning algorithm to eventually learn how to predict Y from X. In practice, doing so constitutes a brute-force ciphertext-only attack against the encryption algorithm, which is commonly believed to be practically infeasible (Goldreich, 2004).

7.2.2. Leveraging other advances in machine learning

A central premise of our method is the link between causal direction and the predictive accuracy of machine learning algorithms. When the causal direction flows only from X to Y (i.e., $X \rightarrow Y$), the distribution of X would reveal no information about the X–Y relationship, rendering semi-supervised learning ineffective.

Following the same logic, causal direction could also affect the effectiveness of machine learning algorithms besides semi-supervised learning. For example, a common generalizability issue facing supervised learning, covariate shift (Sugiyama et al., 2007), arises when a machine learning model degrades in predictive accuracy because of the distributional differences in predictor variables (i.e., X) between the training dataset and the dataset actually in need of prediction. As discussed before, it would be impossible for covariate shift to arise if the causal direction flows only from X to Y because, in this case, a change of X's distribution should have no bearing on the X–Y relationship. In machine learning, Kügelgen et al. (2019) leveraged this property to address the degradation of predictive accuracy caused by covariate shift. Similarly, future research could examine the use of covariate shift or, more

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

broadly, the generalizability of supervised learning models to help identify the causal relationship between X and Y.

7.2.3. Integration of our method with existing methods

A key feature of our method is robustness to model misspecification, as it imposes no restrictions on the functional form of the data-generating process (e.g., linearity or the addition of independent noise). This flexibility minimizes the risk of spurious discoveries, such as the incorrect inference of reverse causality, caused by inappropriate model specifications or dependency structures. However, a notable limitation of our method lies in its relatively weak statistical power. For instance, our simulations demonstrate that it may struggle to detect reverse causality when the sample size is small. As such, our method may be better suited to larger datasets (e.g., with $N > 500$).

These characteristics make our method a valuable complement to existing approaches, such as additive noise models and DDA, which can be sensitive to model misspecifications (e.g., Schultheiss & Bühlmann, 2024; Thoemmes, 2015). For example, while existing methods may not be ideally suited for exploratory investigations into causal directions between variables (Wiedermann & von Eye, 2015), our method can serve this purpose effectively. Researchers could first employ our approach as an exploratory tool to identify promising relationships. Subsequently, they may leverage existing methods to conduct confirmatory analyses of those relationships supported by robust theoretical formulations but not identifiable through our approach (e.g., in cases where neither $X \to Y$ nor $Y \to X$ is detected).

Our method can also be seamlessly integrated with panel models such as CLPM and its variants. To illustrate, applying our method to infer the causal direction between X and Y could result in one of four possible outcomes: (1) the detection of $Y \to X$ but not $X \to Y$, (2) the detection of $X \to Y$ but not $Y \to X$, (3) the detection of both $X \to Y$ and $Y \to X$, suggesting reciprocal relationships or the potential existence of an unobserved confounder, and (4) the detection of neither $X \to Y$ nor $Y \to X$. With the first two outcomes, our model informs the model specification for CLPM, as researchers could choose to remove the cross-lagged effect inconsistent with the outcome of our method. An important future direction for research is to study whether causal learning could be used not only for identifying the causal direction (as in our work) but also for estimating the temporal lag of an effect, which would further improve the model specification for CLPM. The third outcome could prompt researchers to consider panel models that allow for certain types of latent confounders—e.g., the RI-CLPM (Hamaker et al., 2015), which models time-invariant confounders, or Latent Curve Models with Structured Residuals (LCM-CR; Curran et al., 2014), which models time-varying confounders with patterns such as autoregressive structures. Finally, as discussed earlier, researchers facing the fourth outcome could consider the use of existing confirmatory methods, e.g., additive noise models and DDA, to identify the causal direction.

8. Conclusion

In conclusion, our work proposes a novel method that integrates machine learning and reverse causality testing. Through mathematic analysis, simulation studies, and a case illustration, we demonstrates the effectiveness of this method. We hope this approach inspires future research to more strongly embrace the advancements in computer science and machine learning to enrich the methodological toolkit of psychological sciences.

Acknowledgements. This study has been realized using data collected by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The SHP project is supported by the Swiss National Science Foundation.

Funding statement. NZ and HX were supported in part by the US National Science Foundation under Grants 1851637 and 2040807, and by gifts from Amazon Science and Meta. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the sponsors.

Competing interests. Apart from the affiliations and sponsors disclosed above, the authors declare no competing interests.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

References

Aggarwal, C. C., Hinneburg, A., & Keim, D. A. (2001). On the surprising behavior of distance metrics in high dimensional space. In *Proceedings of the international conference on database theory*. Springer, 420–434.

Allen, T. D., French, K. A., Dumani, S., & Shockley, K. M. (2020). A cross-national meta-analytic examination of predictors and outcomes associated with work–family conflict. *Journal of Applied Psychology*, *105*(6), 539–576.

Amini, M.-R., & Gallinari, P. (2002). Semi-supervised logistic regression. In *Proceedings of the 15th European conference on artificial intelligence*, 390–394.

Amstad, F. T., Meier, L. L., Fasel, U., Elfering, A., & Semmer, N. K. (2011). A meta-analysis of work–family conflict and various outcomes with a special emphasis on cross-domain versus matching-domain relations. *Journal of Occupational Health Psychology*, *16*(2), 151–169.

Balkundi, P., Kilduff, M., & Harrison, D. A. (2011). Centrality and charisma: Comparing how leader networks and attributions affect team performance. *Journal of Applied Psychology*, *96*(6), 1209–1222.

Bauer, S., Schölkopf, B., & Peters, J. (2016). The arrow of time in multivariate time series. In *International conference on machine learning*. PMLR, 2043–2051.

Belkin, M., Niyogi, P., & Sindhwani, V. (2006). Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *Journal of Machine Learning Research*, *7*(11), 2399–2434.

Bishop, C. M., & Nasrabadi, N. M. (2006). *Pattern recognition and machine learning*. Springer.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *Proceedings of the 11th annual conference on computational learning theory*, 92–100.

Bollen, K. A. (1989). *Structural equations with latent variables*. John Wiley & Sons.

Bradley, J. V. (1978). Robustness? *British Journal of Mathematical and Statistical Psychology*, *31*(2), 144–152.

Brass, D. J. (1984). Being in the right place: A structural analysis of individual influence in an organization. *Administrative Science Quarterly*, *29*(4), 518–539.

Buja, A., Brown, L., Berk, R., George, E., Pitkin, E., Traskin, M., Zhang, K., & Zhao, L. (2019). Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, *34*(4), 523–544.

Burt, R. (1992). *Structural holes: The social structure of competition*. Harvard University Press.

Busenbark, J. R., Yoon, H., Gamache, D. L., & Withers, M. C. (2022). Omitted variable bias: Examining management research with the impact threshold of a confounding variable (ITCV). *Journal of Management*, *48*(1), 17–48.

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 1–27.

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *82*(1), 39–67.

Cortes, C., & Vapnik, V. (1995). Support vector networks. *Machine Learning*, *20*, 273–297.

Cover, T. M.,& Thomas, J. A. (2006). *Elements of information theory*. Wiley-Interscience.

Curran, P. J., Howard, A. L., Bainter, S. A., Lane, S. T., & McGinley, J. S. (2014). The separation of between-person and within-person components of individual change over time: A latent curve model with structured residuals. *Journal of Consulting and Clinical Psychology*, *82*(5), 879.

Daniel, W. W. (1990). Kolmogorov–Smirnov one-sample test. In *Applied nonparametric statistics* (pp. 319–330). PWS-Kent.

Daniusis, P., Janzing, D., Mooij, J., Zscheischler, J., Steudel, B., Zhang, K., & Schölkopf, B. (2010). Inferring deterministic causal relations. In *26th conference on uncertainty in artificial intelligence (UAI 2010)*. AUAI Press, 143–150.

De Brabanter, K., De Brabanter, J., Suykens, J. A., & De Moor, B. (2010). Approximate confidence and prediction intervals for least squares support vector regression. *IEEE Transactions on Neural Networks*, *22*(1), 110–120.

Debruyne, M. (2009). An outlier map for support vector machine classification. *The Annals of Applied Statistics*, *3*(4), 1566–1580.

Demerouti, E., Bakker, A. B., & Bulters, A. J. (2004). The loss spiral of work pressure, work–home interference and exhaustion: Reciprocal relations in a three-wave study. *Journal of Vocational Behavior*, *64*(1), 131–149.

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, *7*, 1–30.

Dietterich, T. G. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural Computation*, *10*(7), 1895–1923.

Geiger, P., Zhang, K., Schoelkopf, B., Gong, M., & Janzing, D. (2015). Causal inference by identification of vector autoregressive processes with hidden components. *Proceedings of Machine Learning Research*, *37*(1), 1917–1925.

Goldreich, O. (2004). *Foundations of cryptography* (Vol. *2*). Cambridge University Press.

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. MIT press.

Grandvalet, Y., & Bengio, Y. (2004). Semi-supervised learning by entropy minimization. In *Advances in neural information processing systems* (Vol. *17*, pp. 529–536).

Greenhaus, J. H., & Beutell, N. J. (1985). Sources of conflict between work and family roles. *Academy of Management Review*, *10*(1), 76–88.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Halaby, C. N. (2004). Panel models in sociological research: Theory into practice. *Annual Review of Sociology*, *30*, 507–544.

Hamaker, E. L. (2024). The within-between dispute in cross-lagged panel research and how to move forward. *Psychological Methods*, [Advanced Online Publication]

Hamaker, E. L., Kuiper, R. M., & Grasman, R. P. (2015). A critique of the cross-lagged panel model. *Psychological Methods*, *20*(1), 102–116.

Harring, J. R., McNeish, D. M., & Hancock, G. R. (2017). Using phantom variables in structural equation modeling to assess model sensitivity to external misspecification. *Psychological Methods*, *22*(4), 616–631.

Heskes, T. (1996). Practical confidence and prediction intervals. In M.C. Mozer, M. Jordan and T. Petsche (Eds.). *Advances in neural information processing systems* (Vol. *9*, pp. 176–182). Neural Information Processing Systems Foundation, Inc.

Hornik, K., Stinchcombe, M., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, *2*(5), 359–366.

Hoyer, P., Janzing, D., Mooij, J. M., Peters, J., & Schölkopf, B. (2008). Nonlinear causal discovery with additive noise models. In D. Koller and D. Schuurmans and Y. Bengio and L. Bottou (Eds.). *Advances in neural information processing systems* (Vol. *21*, pp. 689–696). Neural Information Processing Systems Foundation, Inc.

Hyvärinen, A., Zhang, K., Shimizu, S., & Hoyer, P. O. (2010). Estimation of a structural vector autoregression model using non-gaussianity. *Journal of Machine Learning Research*, *11*(5), 1709–1731.

Janzing, D., & Schölkopf, B. (2010). Causal inference using the algorithmic Markov condition. *IEEE Transactions on Information Theory*, *56*(10), 5168–5194.

Janzing, D., & Schölkopf, B. (2015). Semi-supervised interpolation in an anticausal learning scenario. *The Journal of Machine Learning Research*, *16*(1), 1923–1948.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical Review*, *106*(4), 620–630.

Judge, T. A., Weiss, H. M., Kammeyer-Mueller, J. D., & Hulin, C. L. (2017). Job attitudes, job satisfaction, and job affect: A century of continuity and of change. *Journal of Applied Psychology*, *102*(3), 356–374.

Kenny, D. A., & Zautra, A. (1995). The trait-state-error model for multiwave data. *Journal of Consulting and Clinical Psychology*, *63*(1), 52.

Kossek, E. E., & Ozeki, C. (1998). Work–family conflict, policies, and the job–life satisfaction relationship: A review and directions for organizational behavior–human resources research. *Journal of Applied Psychology*, *83*(2), 139–149.

Kügelgen, J., Mey, A., & Loog, M. (2019). Semi-generative modelling: Covariate-shift adaptation with cause and effect features. *Proceedings of Machine Learning Research*, *89*, 1361–1369.

Landolfi, A., Brondino, M., Molino, M., & Presti, A. L. (2022). Dont worry, be happy! positive affect at work, greater balance at home a daily diary study on work-family balance. *European Review of Applied Psychology*, *72*(1), 100715.

Leszczensky, L., & Wolbring, T. (2022). How to deal with reverse causality using panel data? recommendations for researchers based on a simulation study. *Sociological Methods & Research*, *51*(2), 837–865.

Li, X., & Wiedermann, W. (2020). Conditional direction dependence analysis: Evaluating the causal direction of effects in linear models with interaction terms. *Multivariate Behavioral Research*, *55*(5), 786–810.

Linnik, Y. V. (1957). On the decomposition of the convolution of gaussian and poissonian laws. *Theory of Probability & Its Applications*, *2*(1), 31–57.

Lucas, R. E. (2023). Why the cross-lagged panel model is almost never the right choice. *Advances in Methods and Practices in Psychological Science*, *6*(1), 1–22.

Lukacs, E. (1970). *Characteristic functions*. Griffin.

Mauro, R. (1990). Understanding LOVE (left out variables error): A method for estimating the effects of omitted variables. *Psychological Bulletin*, *108*(2), 314–329.

Mulder, J. D., & Hamaker, E. L. (2021). Three extensions of the random intercept cross-lagged panel model. *Structural Equation Modeling: A Multidisciplinary Journal*, *28*(4), 638–648.

Niyogi, P. (2013). Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, *14*(5), 1229–1250.

Orth, U., Clark, D. A., Donnellan, M. B., & Robins, R. W. (2021). Testing prospective effects in longitudinal research: Comparing seven competing cross-lagged models. *Journal of Personality and Social Psychology*, *120*(4), 1013–1034.

Ozkok, O., Vaulont, M. J., Zyphur, M. J., Zhang, Z., Preacher, K. J., Koval, P., & Zheng, Y. (2022). Interaction effects in cross-lagged panel models: Sem with latent interactions applied to work-family conflict, job satisfaction and gender. *Organizational Research Methods*, *25*(4), 673–715.

Pastor, J.-C., Meindl, J. R., & Mayo, M. C. (2002). A network effects model of charisma attributions. *Academy of Management Journal*, *45*(2), 410–420.

Pearl, J. (1998). Graphs, causality, and structural equation models. *Sociological Methods & Research*, *27*(2), 226–284.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference: Foundations and learning algorithms*. The MIT Press.
Peters, J., Mooij, J. M., Janzing, D., & Schölkopf, B. (2014). Causal discovery with continuous additive noise models. *Journal of Machine Learning Research*, *15*(1), 2009–2053.
Pornprasertmanit, S.,& Little, T. D. (2012). Determining directional dependency in causal associations. *International Journal of Behavioral Development*, *36*(4), 313–322.
Rainio, O., Teuho, J., & Klén, R. (2024). Evaluation metrics and statistical tests for machine learning. *Scientific Reports*, *14*(1), 6086.
Rogosa, D. (1980). A critique of cross-lagged correlation. *Psychological Bulletin*, *88*(2), 245–258.
Rosenström, T. H., Czajkowski, N. O., Solbakken, O. A., & Saarni, S. E. (2023). Direction of dependence analysis for pre-post assessments using non-Gaussian methods: A tutorial. *Psychotherapy Research*, *33*(8), 1058–1075.
Salzberg, S. L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, *1*, 317–328.
Santafe, G., Inza, I., & Lozano, J. A. (2015). Dealing with the evaluation of supervised classification algorithms. *Artificial Intelligence Review*, *44*, 467–508.
Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., & Mooij, J. (2012). On causal and anticausal learning. In *Proceedings of the 29th international conference on machine learning*, 459–466. Omnipress (Madison, WI).
Schultheiss, C., & Bühlmann, P. (2024). Assessing the overall and partial causal well-specification of nonlinear additive noise models. *Journal of Machine Learning Research*, *25*(159), 1–41.
Seeger, M. (2000). *Learning with labeled and unlabeled data*. (tech. rep.) University of Edinburgh.
Shils, E. (1965). Charisma, order and status. *American Sociological Review*, *30*(2), 199–213.
Shimizu, S., Hoyer, P. O., Hyvärinen, A., Kerminen, A., & Jordan, M. (2006). A linear non-Gaussian acyclic model for causal discovery. *Journal of Machine Learning Research*, *7*(10), 2003–2030.
SHP Group (2023), Living in Switzerland Waves 1-23 + Covid 19 data (6.0.0) [Dataset]. FORS data service. https://doi.org/10.48573/642z-p311
Shojaie, A., & Fox, E. B. (2022). Granger causality: A review and recent advances. *Annual Review of Statistics and its Application*, *9*, 289–319.
Sohn, K., Berthelot, D., Carlini, N., Zhang, Z., Zhang, H., Raffel, C. A., Cubuk, E. D., Kurakin, A.,& Li, C.-L. (2020). Fixmatch: Simplifying semi-supervised learning with consistency and confidence. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan and H. Lin (Eds.). *Advances in neural information processing systems* (Vol. *33*, pp. 596–608). Neural Information Processing Systems Foundation, Inc.
Sterner, P., Goretzko, D., & Pargent, F. (2025). Everything has its price: Foundations of cost-sensitive learning and its application in psychology. *Psychological Methods*, *30*(1), 112–127.
Stine, R. A. (1985). Bootstrap prediction intervals for regression. *Journal of the American Statistical Association*, *80*(392), 1026–1031.
Sugiyama, M., Krauledat, M., & Müller, K.-R. (2007). Covariate shift adaptation by importance weighted cross validation. *Journal of Machine Learning Research*, *8*(5), 985–1005.
Tax, D., & Duin, R. (2004). Feature scaling in support vector data descriptions. *Pattern Recognition Letters*, *25*(11), 1161–1169.
Thoemmes, F. (2015). Empirical evaluation of directional-dependence tests. *International Journal of Behavioral Development*. *39*(6), 560–569.
Tran, D., Ranganath, R., & Blei, D. M. (2016). The variational Gaussian process. In *4th international conference on learning representations*.
Vaisey, S., & Miles, A. (2017). What you can—and can't—Do with three-wave panel data. *Sociological Methods & Research*, *46*(1), 44–67.
Van Engelen, J. E., & Hoos, H. H. (2020). A survey on semi-supervised learning. *Machine Learning*, *109*(2), 373–440.
von Eye, A., & DeShon, R. P. (2012). Directional dependence in developmental research. *International Journal of Behavioral Development*, *36*(4), 303–312.
Voorpostel, M., Tillmann, R., Lebert, F., Kuhn, U., Lipps, O., Ryser, V.-A., Antal, E., Dasoki, N., & Wernli, B. (2023). *Swiss household panel user guide (1999-2021)*. (23rd ed.) FORS.
Wiedermann, W., & Li, X. (2018). Direction dependence analysis: A framework to test the direction of effects in linear models with an implementation in SPSS. *Behavior Research Methods*, *50*, 1581–1601.
Wiedermann, W., & von Eye, A. (2015). Direction-dependence analysis: A confirmatory approach for testing directional theories. *International Journal of Behavioral Development*, *39*(6), 570–580.
Wilcox, K. T., Jacobucci, R., Zhang, Z., & Ammerman, B. A. (2023). Supervised latent dirichlet allocation with covariates: A Bayesian structural and measurement model of text and covariates. *Psychological Methods*, *28*(5), 1178–1206.
Wüthrich, K.,& Zhu, Y. (2023). Omitted variable bias of lasso-based inference methods: A finite sample analysis. *Review of Economics and Statistics*, *105*(4), 982–997.
Xie, Q., Luong, M.-T., Hovy, E., & Le, Q. V. (2020). Self-training with noisy student improves imagenet classification. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. IEEE, 10684–10695.

https://doi.org/10.1017/psy.2025.13 Published online by Cambridge University Press

Zhang, K., & Hyvärinen, A. (2009). On the identifiability of the post-nonlinear causal model. In *Proceedings of the 25th conference on uncertainty in artificial intelligence (UAI)*. AUAI Press, 647–655.

Zimmer, F., & Debelak, R. (2023). Simulation-based design optimization for statistical power: Utilizing machine learning. *Psychological Methods*. Advance online publication. https://doi.org/10.1037/met0000611

Cite this article: Zhang, N., Xu, H., Vaulont, M. J. and Zhang, Z., (2025). Testing of Reverse Causality Using Semi-Supervised Machine Learning. *Psychometrika*, 1–25. https://doi.org/10.1017/psy.2025.13