Targeted After Generative (TAG) Learning for Causal Inference with Latent Variables

Ben Kelcey
Amota Ataneka
Jean-Baptiste Habarurema

University of Cincinnati

Abstract

Latent variables and causal inference are central to implementing empirical research and advancing substantive theory across the social sciences. Conventional latent variable frameworks such as structural equation modeling (SEM) are constrained by several simplifying assumptions (e.g., linearity, additivity, correct model specification). In this study, we relax key SEM assumptions by building data adaptive estimation methods for causal inference with latent variables. We developed a theory-driven Targeted After Generative (TAG) learning framework that integrates SEM-informed variational autoencoders (SEM VAEs) with targeted machine learning to estimate the average treatment effect. On the generative side, the SEM VAE encodes theory aligned neural networks that learn latent variable distributions through flexible joint modeling of measurement and structural parameters. On the targeted learning side, we construct an efficient influence function for the treatment effect with the SEM-VAE implied posterior distribution of the latent variables to estimate the average effect. Across simulations, the TAG learning approach outperforms alternative approaches and demonstrates low to no bias even in adverse conditions. TAG learning offers a theory-driven flexible approach with methodological promise for causally interpretable data-adaptive analysis with latent variables. More practically, the framework allows estimation of causal effects involving latent variables even in the presence of unknown functional relationships in the measurement and structural models (e.g., data adaptive effect estimation for nonlinear/nonadditive SEMs).

Introduction

Reflective latent variables (e.g., self-efficacy, extraversion, student achievement) are ubiquitous in social science research (e.g., Bainter & Bollen, 2014; Dawnson et al., 2020; Fan et al., 2016; Finn & Wang, 2014; Kelleher et al., 2020; Marsh & Hau, 2007; Nagengast & Trautwein, 2015; Sukserm, 2024). For example, reflective latent variables play a central role in the evaluation of interventions and a foundational role in the empirical development of theory in psychology and across the social sciences (e.g., Bollen, 2002; Connor et al., 2020; Eignor, 2013). Similarly, reflective latent constructs also play essential roles in informing policy and practice (e.g., Raudenbush, 2008). Estimates of treatment effects on latent outcomes such as student achievement often guide study design and high-stakes decisions about resource allocation, instructional strategies, and intervention design (e.g., Carlisle et al., 2020; Cox & Kelcey, 2019; Dwyer et al., 2016). Because reflective latent variables cannot be observed directly, researchers rely on fallible indicators (e.g., survey questions, rater judgements or test items) that reflect the underlying latent traits. Consequently, accurate modeling of the relationships among observed and reflective latent variables is essential for valid inference, theory building and policy development (e.g., Raudenbush, 2008).

Structural equation models (SEMs) have long been the dominant framework for analyzing relationships among reflective latent variables (e.g., Muthén & Muthén, 2009). While SEM provides a principled tool for measurement and structural modeling, conventional applications of SEM rely on strong simplifying assumptions typically including linearity, additivity, and correct model specification (e.g., Klein, 2012). In practice, these assumptions are often violated or at least unsubstantiated because the true functional forms of relationships among variables are complex and/or unknown. Model misspecification (e.g., omission of

nonlinear or higher-order terms when present) can lead to biased estimates and invalid inferences in ways that fundamentally constrain the utility of conventional SEMs in contexts where the functional forms of relationships are unknown (e.g., Rosseel et al., 2025).

Recent advances in machine learning (ML) have introduced flexible, data-adaptive methods that relax or reduce many of the restrictive assumptions inherent in parametric models such as SEM (e.g., Athey et al., 2018, 2019; Bai et al., 2022; Chernozhukov et al., 2017; van der Laan & Rose 2011). For example, several frameworks combine data adaptive algorithms learning to uncover complex, nonlinear, and nonadditive relationships that are causally interpretable and (doubly) robust to model misspecification without ever specifying the functional forms of relationships (e.g., Chernozhukov et al., 2017; van der Laan & Rose 2011).

Despite these advantages, existing ML approaches for causal inference generally assume that all relevant variables are observed directly and measured without error (e.g., Jo et al., 2023). This assumption will often prove problematic in psychological research because core predictors and outcomes are often operationalized as reflective latent constructs that are measured by fallible indicators. Standard applications of ML do not incorporate measurement models or have the mechanisms to account for the uncertainty introduced by reflective latent variables. Much like their simpler parametric counterparts (e.g., linear regression), naively substituting factor scores or averages introduces bias into effect estimates and undermines causal validity. As a result, if social sciences are to capitalize on recent advances, there is a critical need for methods that integrate latent variable modeling with flexible, data-adaptive methods for causal inference.

In this study, we developed a targeted after generative (TAG) learning approach—a framework that pairs causal machine learning with theory-aligned generative modeling for latent variables. The goal of TAG learning is to work toward bringing the benefits of causal machine

learning (e.g., flexibility, double robustness, and valid inference under unknown functional forms) to the setting that characterizes most theory-driven social science research: one where the outcome and key covariates are reflective latent constructs measured by fallible indicators. TAG learning can be conceptually outlined through a simple contrast with conventional SEM when estimating the effect of an observed treatment on a latent outcome controlling for latent covariates. In a standard linear SEM, the measurement model for a latent outcome (and similarly latent covariates) expresses observed indicators as linear functions of the latent variable

$$Y_{ij} = \alpha_j^Y + \lambda_j^Y \eta_i^Y + \varepsilon_{ij}^Y \quad (1)$$

where the observed indicator responses (Y_{ij}) for individual *i* on indicator *j* are a function of indicator-specific intercepts (α_j^Y), the latent outcome (η_i^Y), indicator-specific loading coefficients λ_j^Y and individual-specific error ε_{ij}. In turn, the structural model expresses the latent outcome as a linear and additive function of the treatment (T_i) and latent covariates (η_i^X and η_i^Z):

$$\eta_i^Y = \beta_0 + \delta T_i + \beta_X \eta_i^X + \beta_Z \eta_i^Z + \varepsilon_i \quad (2)$$

Both expressions (measurement and structural) require the correct specification of the functional form (e.g., linear or a specific type of nonlinearity) in advance in order to obtain consistent estimates of the treatment effect (δ) . TAG learning replaces these fixed (linear) relationships with flexible, unknown functions that are estimated through data-adaptive methods. In TAG, the measurement model expresses each indicator as an unspecified function of its latent variable e.g.,

$$Y_{ij} = f(\eta_i^Y) + \varepsilon_{ij}^Y \quad (3)$$

Similarly, the structural model expresses the latent outcome as an unspecified function of treatment and covariates

$$\eta_i^Y = f(T_i, \eta_i^X, \eta_i^Z) + \varepsilon_i \quad (4)$$

In both expressions, $f(\cdot)$ denotes unknown functions that are learned from the data rather than known linear functions. Our goal in this study is to develop data-adaptive methods to estimate the average treatment effect.

TAG learning bridges the gap between SEM, unconstrained deep learning, and causal inference. On the generative side, the SEM-VAE encodes substantive theory into the measurement and structural architecture, producing interpretable, theory-aligned latent variables rather than emergent representations that conflate constructs. On the targeted learning side, TMLE uses posterior summaries from the SEM-VAE to support estimators that, under congeniality conditions outlined in Appendix B, can in principle inherit the doubly robust, semi-parametrically efficient properties of TMLE while accounting for measurement error. TAG is designed for the confirmatory setting that characterizes much of social science research but has been underdeveloped in the ML literature: the relevant constructs, their indicators, and the temporal ordering of variables are established by theory and prior psychometric work before the study begins. TAG leverages and encodes that prior knowledge into the analysis rather than ignore or rediscover it from data. The remainder of the paper reviews existing approaches and their limitations, develops the TAG framework, evaluates it through simulations, presents an empirical illustration, and concludes with implications, limitations, and future directions.

Background

Structural Equation Models. The conventional and dominant approach to latent variable modeling is SEMs. SEMs integrate two components: a measurement model that specifies how observed indicators reflect underlying latent constructs and a structural model that specifies relationships among latent and observed variables. By jointly estimating the measurement and structural models, SEMs address measurement error and provides a principled

framework for inference on latent constructs (e.g., Bollen, 1989; Brown, 2015; Kline, 2016). SEM estimation typically draws on maximum likelihood or related full information estimators under assumptions such as multivariate normality, correct model specification, and large sample sizes. When these assumptions hold, SEM yields consistent and asymptotically efficient estimates of both measurement parameters (e.g., factor loadings, error variances) and structural parameters (e.g., regression coefficients among latent variables).

Despite its strengths, a critical limitation of SEMs is that the consistency of parameter estimates depends on the propriety of the model; that is, consistent parameter estimation requires that the structural and measurement models are correctly specified. This assumption is restrictive when the functional forms of relationships among observed and latent variables are complex or unknown as they often are in real-world applications. For instance, even when considering potential nonlinearity or nonadditivity a model may appropriately include two-way interactions but incorrectly omit three-way interactions, interactions among nonlinear terms, or other nonparametric relationships (e.g., Hainmueller et al., 2019). Conventional SEM is not a data-adaptive method in that lacks the flexibility to detect and incorporate flexible functional forms or interactions without explicit specification. As a result, parameter estimates are highly susceptible to model misspecification and this fundamentally limits the utility and robustness of SEM estimates in real-world settings where functional forms are often unknown and complex (e.g., MacCallum & Austin, 2000; Preacher et al., 2015).

Targeted Machine Learning. A contemporary approach to relaxing model specification requirements is targeted learning, which combines ensemble machine learning with a semiparametric estimation procedure designed for causal inference (van der Laan & Rose, 2011). The super learner—an ensemble method that combines predictions from a library of algorithms

using cross-validation to determine optimal weights—provides data-adaptive approximation of conditional expectations without committing to a specific functional form (van der Laan et al., 2007). Targeted learning leverages the efficient influence function (EIF) for the parameter of interest in a two-step procedure: nuisance functions (outcome regression and propensity score) are first estimated with the super learner, then a targeting step updates these estimates to solve the EIF estimating equation. This targeting step ensures the resulting estimator is consistent, asymptotically normal, and doubly robust (consistent if either the outcome regression or propensity score is correctly specified) and semi-parametrically efficient, preserving valid inference in complex settings (Chernozhukov et al., 2018; van der Laan & Rose, 2011).

Despite these advantages, targeted learning assumes all variables are fully observed and measured without error. Unlike SEM, it provides no mechanism for incorporating a measurement model. As a result, naively substituting factor scores or composite averages for latent variables introduces bias into effect estimates in ways that undermine causal validity (Regier et al., 2014).

Factor Score Methods. A historical approach to incorporate latent variables into subsequent structural models is to substitute factor scores in place of the latent variables (e.g., Bogaert et al., 2026; Devlieger et al., 2016; Devlieger & Rosseel, 2020; Kelcey, 2019). For example, factor scores can be predicted using a simple linear weighted combination of indicators for a linear factor models (see Appendix F). Subsequently these scores are substituted for the latent variables within a regression or the targeted learning framework and used as if they were observed scores. Factor scores are point estimates that fail to capture and adjust for the inherent uncertainty in latent variables. Consequently, using factor scores to estimate treatment effects or structural relationships introduces bias when the reliabilities of the predictors are less than perfect (e.g., Kelcey, 2019; Regier et al., 2014).

Structural After Measurement Approaches. Recent research has sought to overcome the limitations of factor-score based methods by developing structural after measurement (SAM) approaches (e.g., Rosseel et al., 2024). Factor-score methods incur bias because they ignore measurement error in the first stage and this error propagates into structural estimates in the second stage. However, a key insight from the SAM literature is that the bias introduced by ignoring measurement error is a function of the respective measurement models (e.g., Croon, 2002). SAM methods address this bias by developing approaches that avoid such bias altogether or methods that introduce bias correction terms derived from measurement models that adjust structural parameters to account for the uncertainty (e.g., Bogaert et al., 2026; Cox & Kelcey, 2021ab; Kelcey, Cox, & Dong, 2020; Rosseel et al., 2025). Prior research has developed several different approaches and found that they are quite effective across a broad range of SEMs when compared to naïve factor-score approaches (e.g., Kelcey et al., 2020). Similarly, SAM based methods have also outperformed conventional full information estimators in a variety of settings (e.g., Hayes & Usami, 2020). For example, a key consideration in nonlinear SEM is the computational complexity associated with full information estimators (e.g., maximum likelihood); the feasibility of estimation progressively diminishes when we incorporate each additional nonlinear or nonadditive term for latent variables (e.g., Cox & Kelcey, 2021; 2023; Rosseel et al., 2024; 2025)

Although SAM methods work well and address the computational complexity associated with incorporating nonlinear and nonadditive terms for latent variables, the principles underlying SAM approaches suffer from two core limitations in the current context. First, the bias avoiding and bias correction SAM approaches are fundamentally built on the covariance matrix of observed variables. Although covariance is a sufficient statistic in linear SEMs, covariance is not

necessarily a sufficient statistic when considering complex nonlinear, nonadditive or nonparametric relationships. To some extent, the scope of recent SAM developments has incorporated nonlinear and nonadditive relationships (e.g., Cox et al., 2021; Cox et al., 2023; Rosseel et al, 2025). However, in cases where relationships are more complex or not easily represented by parametric forms, the covariance-based adjustments that conventional SAM approaches depend on fail to capture the full structure and scope of relationships in the data. By contrast, modern machine learning algorithms estimate nuisance functions using the entire data distribution, enabling them to flexibly model data without relying solely on second-order moments.

Second, like conventional SEM, even when models incorporate nonlinear or interactive terms, SAM still requires that the measurement and structural models are correctly specified. When functional forms are unknown or complex (e.g., multiple nonlinear relationships), single or multiple misspecifications in the relationships among variables can quickly accrue bias in parameter estimates. As a result, although SAM approaches introduce important flexibility, they are still limited when relationships are unknown or not easily captured by parametric forms.

Plausible Values. Plausible value (PV) approaches, widely used in large-scale assessments such as NAEP and TIMSS, address measurement uncertainty by drawing multiple imputations from the posterior distribution of the latent variable and pooling results across imputations (e.g., Thomas, 2002; von Davier et al., 2007; von Davier & Sinharay, 2013). Unlike factor score methods, PVs retain uncertainty in the latent variable rather than relying on a single point estimate. However, PVs provide consistent structural estimates only when the imputation and analytic models are congenial; that is, when the conditional posterior distribution used for drawing plausible values incorporates the full scope of relationships in the analytic model

(Bartlett & Hughes, 2020; Jewsbury et al., 2024; Meng, 1994). Like conventional SEM and SAM approaches, PV methods therefore remain constrained by correct specification of the measurement and structural models; when either is misspecified, consistency is not preserved.

Variational Autoencoders. Recent developments in machine learning have suggested that variational autoencoders (VAEs) may serve as a central tool across disciplines for data reduction (e.g., Wang et al., 2024). VAEs are a deep learning-based nonlinear dimension reduction method that learns a low-dimensional representation of the data using a probabilistic, smooth, and generative model that compresses inputs through an encoder and then reconstructs those data through a decoder; the low-dimensional representation acts as a nonlinear set of dimensions that preserves the predictive structure in the data (e.g., Kingma & Welling, 2013). By learning a compact distribution, VAEs function as a type of latent variable model that assumes the data arise from a small set of unobserved (latent) factors and the model learns how to both infer those factors from data and how to generate data from those factors.

While variational autoencoders (VAEs) offer significant architectural flexibility, they do not inherently produce interpretable latent variables (Kingma & Welling, 2022). Because the primary goal of a VAE is predictive performance the resulting latent space rarely aligns with interpretable, pertinent or established substantive constructs (Locatello et al., 2019). Rather, VAEs tend to conflate or interleave latent variables in ways that are uninterpretable and untethered to theory (e.g., Bainter & Bollen, 2014; Borsboom, 2008; Finn & Wang, 2014). Similar issues emerge in neighboring areas such as causal representation learning (e.g., Bareinboim, 2025). A major challenge in this literature is that recovering latent causal variables from observational data is generally non-identifiable without additional structural constraints.

More generally, a critical hurdle in this area is the issue of identifiability and factor recovery. Prior research has demonstrated that unsupervised disentanglement is fundamentally impossible without specific inductive biases or supervision (Locatello et al., 2019). That is, without the structural constraints, no unsupervised objective can guarantee the recovery of the true latent variables. This identifiability result has implications beyond statistical theory because interpretability, construct validity, and transparency are not just statistical criteria, they are also foundational and necessary concepts for developing and testing substantive theories in many fields (e.g., Borgstede & Eggert, 2023; Kelcey, McGinn & Hill, 2014).

For these reasons, recent advances in VAEs have moved beyond standard unconstrained architectures to build identification into the model through explicit constraints. For example, linear factor and item response models have been introduced inside the neural architecture to ensure theory-aligned, interpretable latent variables. In these approaches, identification (and interpretability) is enforced through theory-driven architectural constraints such as loading patterns and scale constraints (e.g., Cui et al., 2024; Kilian et al. 2023; Liu et al., 2022; Molenaar et al., 2025; Zhang & Chen, 2024; Zhang et al., 2025).

More conceptually, these developments base the latent variable model on substantive theory and prior empirical evidence so that the learned representation is identifiable and interpretable by construction. An important limitation of these approaches, however, is that they have narrowly focused on measurement models and disentangling latent variables while leaving the causal structure undeveloped. That is, the primary inferential target of these approaches has been on measurement properties and factor recovery to the exclusion of conditional relations among observed and latent variables and treatment effects.

Unconstrained Deep Learning. An extension of the VAE framework to causal inference is the neural causal model (NCM), which parameterizes every structural equation in a causal graph as a neural network and learns the full joint causal data-generating process from observed data (Bareinboim et al., 2022; Xia & Bareinboim, 2021). NCMs attempt to replace fixed parametric structural equations with flexible neural networks and in principle represent any structural causal model without requiring the researcher to specify functional forms in advance.

A core practical limitation of NCMs in social science research is the tradeoff between expressive capacity and sample complexity. In a NCM, each structural equation is parameterized as a multi-layer neural network. As a result, the function class available to each equation is the set of all functions representable by networks of a given depth and width (e.g., effectively all piecewise-linear functions on the input domain for standard ReLU architectures). A key result from statistical learning theory is that the number of observations required before the learned function is close to the true target (i.e., sample complexity of reliably learning from a function class) grows with the effective complexity of that class (Neyshabur et al., 2017). For neural networks, this complexity grows with the number of parameters, which in turn grows with depth and width. Crucially, the NCM must learn the full joint causal data generating process rather than only the specific quantities needed to answer a single targeted query: recovering the complete set of structural equations requires learning every mechanism governing how each variable is generated from its parents and noise, across the full joint distribution of all variables and all potential interventions. The target function class is therefore as large and complex as the structural mechanisms themselves, and the sample complexity required for reliable recovery scales accordingly. For example, past research has established that although NCMs are expressive enough to represent any structural causal model their learnability (i.e., actually

recovering the correct model from observational data) is governed by this complexity and is not generally guaranteed in finite samples (Xia & Bareinboim, 2021).

For typical social science research, the model complexity demanded by theory will often outpace these sample requirements. For example, with sample sizes in the hundreds or low thousands, which represent the practical ceiling of many psychological, educational, and health studies, neural networks with the depth and width required to flexibly represent structural equations are severely underpowered: the function class is large enough to represent the truth but not large enough to be reliably identified from the available data, resulting in high variance, poor generalization, and unstable parameter estimates. This is not a limitation that can be resolved by stronger regularization alone because regularization constrains the function class but does so without the theoretical guarantees that result from inductive bias constraints. The practical consequence is that despite the theoretical expressiveness of NCMs, they are unlikely to deliver reliable structural learning in the sample sizes that characterize most psychological research.

Causal representation learning (CRL) takes this approach further by additionally using unsupervised learning to recover latent causal variables from high-dimensional observational data when neither the latent structure nor the causal relationships among variables are known in advance (e.g., Bareinboim, 2025). CRL methods aim to discover the latent space and its causal organization jointly from data. As with VAEs, CRL faces the identifiability barrier noted above (Locatello et al., 2019). Without constraints, no unsupervised objective can guarantee that the learned latent representations correspond to the true causal variables rather than to some arbitrary transformation of them. This identification failure is not merely a statistical problem of insufficient sample size but rather it is a structural problem: the data simply do not contain enough information to uniquely identify the latent causal structure without prior constraints.

Method

In this study we develop a confirmatory deep learning approach that addresses the gaps identified above by embedding psychometric theory within the measurement architecture, study design within the causal ordering, and substantive theory within the structural architecture. The identification strategy underlying TAG leverages the accumulated theoretical and psychometric knowledge embedded in established measurement instruments—converting prior scientific knowledge about construct structure into identification power—rather than attempting to discover that structure from the data itself. The TAG learning approach offers four conceptual advantages that we explore in this study. First, TAG produces latent variable summaries that are anchored in substantive theory and whose measurement properties are governed by the researcher's prior psychometric knowledge rather than by an unsupervised regularization objective. The result is latent variables that are interpretable by construction, consistent with the theory under investigation, and directly comparable across studies that use the same instruments—properties that unconstrained deep learning approaches cannot guarantee.

Second, TAG replaces fixed linear measurement and structural equations with data-adaptive neural networks that accommodate unknown nonlinear and nonadditive relationships in both the measurement and the structural models. Unlike conventional SEM, SAM, and plausible value approaches that require the researcher to correctly specify the functional forms of these relationships in advance, TAG learns them from data. However, unlike unconstrained approaches, TAG offers data adaptive estimation while preserving interpretable latent structure.

Third, the targeted-learning step is constructed to inherit TMLE's double robustness property under conditions outlined in Appendix B: when the SEM-VAE produces congenial latent summaries, the estimate is consistent if either the outcome regression or the propensity score is consistently estimated, even if the other converges at a slower rate. Whether these conditions hold in practice is an empirical question we examine through simulation rather than a property that is guaranteed by the framework.

Fourth, by combining theory-specified confirmatory constraints in the SEM-VAE with TMLE's semiparametric framework, TAG achieves favorable finite-sample performance relative to unconstrained alternatives. The confirmatory inductive bias (i.e., fixing indicator block assignments and causal ordering as prior knowledge rather than discovering them from data) substantially constrains the function class each SEM-VAE component must search, reducing the effective complexity relative to models that learn the full joint causal data-generating process (e.g., Bartlett et al., 2019; Xia & Bareinboim, 2021). The SEM-VAE's dimensionality reduction from raw indicators to low-dimensional latent posterior means then further simplifies the target function class for the downstream TMLE step, which needs to learn only the outcome regression and propensity score as functions of those summaries rather than as functions of the full indicator space. TMLE's orthogonal estimating equation is designed so that, under the standard product-of-rates conditions, errors in each nuisance function contribute only second-order bias to the ATE estimate when the other nuisance function is adequately estimated. Taken together, these features suggest that TAG's sample requirements may be considerably more favorable than those of unconstrained deep learning models operating on the same data. Although such results are supported by our simulations, we do not formally establish this. TMLE's efficient influence function further potentially provides a basis for uncertainty quantification: the EIF-based standard error captures a type of lower bound on sampling variability and the bootstrap can be used to further capture uncertainty originating from the estimation stages (e.g., Cai et al., 2020).

Causal Identification with Latent Variables

Our analyses focus on the effect of a binary treatment on a latent outcome controlling for latent and observed covariates. We define the inferential target in terms of the potential outcomes framework (Rubin, 1974). Under the standard assumptions of SUTVA, treatment ignorability,

and positivity, the average treatment effect on a directly observed outcome is identified from observed data. When the outcome of interest is a latent variable (i.e., η) measured by fallible indicators, we can adapt the standard definitions by using $\eta_i^Y(t)$ as the potential (latent) outcome under treatment assignment *t* such that the average latent treatment effect ($\delta_i^{\eta^Y}$) is (e.g., Stoetzer et al., 2025)

$$E[\delta_i^{\eta^Y}] = E[\eta_i^Y(T_i = 1) - \eta_i^Y(T_i = 0)] \quad (5)$$

A fundamental problem in causal inference with latent outcomes is that neither of the potential outcomes are observed. As a result, we specify a measurement model that connects the latent outcome to a set of *J* observed items or indicators believed to be reflect of the unobserved latent outcome ($\boldsymbol{Y} = \{Y_1, Y_2, \ldots Y_j, \ldots Y_J\}$). Although we relax measurement model specification subsequently, a common approach is to adopt a linear confirmatory factor model. To identify the average latent treatment effect, $E\left[\delta_i^{\eta^Y}\right]$, we need five primary assumptions:

(1) The stable unit treatment value assumption: $\eta_i^Y = \eta_i^Y(T_i)$ and $Y_i = Y_i(T_i, \eta_i^Y)$ (e.g., latent and manifest potential outcomes are not a function of others treatment assignment).

(2) Treatment Ignorability: $\{\eta_i^Y(T_i), Y_i(T_i, \eta_i^Y) \perp\!\!\!\perp T \mid \boldsymbol{X}, \boldsymbol{\eta^X}$ (with $\boldsymbol{X}$ and $\boldsymbol{\eta^X}$ as observed and latent covariates) for all *t* and η^Y (e.g., covariates capture treatment assignment mechanism). In this context, we further introduce $\boldsymbol{\eta}_i^X$ as (multiple) latent covariates with, e.g., simple linear confirmatory factor models.

(3) Measurement Ignorability: $\{Y_i(T_i, \eta_i^Y) \perp\!\!\!\perp \eta^Y \mid T, \boldsymbol{X}, \boldsymbol{\eta^X}$ for all *t* and η_i^Y (e.g., correct measurement model; similar assumption extends to latent covariates).

(4) Exclusion Restriction: $Y_i(t_i, \eta_i^Y) = Y_i(t_i', \eta_i^Y)$ for all t, t' and η_i^Y (e.g., treatment impacts items via the latent outcome; similar assumption extends to latent covariates).

(5) Positivity or common latent support: for all values of the observed ($\boldsymbol{X}$) and latent covariates ($\boldsymbol{\eta^X}$), there is a positive probability of receiving both treatment and control: $0 < P(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}, \boldsymbol{\eta}_i^{\boldsymbol{X}} = \boldsymbol{\eta}_i^{\boldsymbol{x}}) < 1$ for all $x \in$ support (X). This ensures that meaningful comparisons between treated and control units can be made across the covariate distribution.

Assumptions 1 and 5 are direct extensions of their observed-data counterparts to the latent setting. Assumptions 3 and 4 are specific to the latent variable context and together ensure that the measurement model does not transmit treatment effects through the indicators independently of the latent construct. Appendix A outlines an identification argument that, under these assumptions, the conditional mean function $Q_0(t, \eta^X, \eta^Z, W) = E[\eta^Y \mid T = t, \eta^X, \eta^Z, W]$ and propensity score $g_0(\eta^X, \eta^Z, W) = P(T = 1 \mid \eta^X, \eta^Z, W)$ are identified from the observed data (the functional identification required for TMLE).

Targeted After Generative Learning

To outline the TAG learning approach, consider two latent covariates (e.g., η_i^X, η_i^Z), an observed covariate (W_i), an observed binary treatment (T_i), and a latent outcome (η_i^Y). Let the observed data ($\mathbf{o}_i$) for individual i be denoted by

$$\mathbf{o}_i = (\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i, w_i, t_i) \tag{6}$$

where w_i is an observed covariate for individual i, $t_i \in \{0,1\}$ is an observed binary treatment, and $\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i$ are indicator vectors for latent covariates (η_i^X, η_i^Z) and the latent outcome (η_i^Y) respectively, each assumed to approximately follow a normal distribution. The latent covariate η_i^X is measured by J_X indicators $\mathbf{x}_i = (x_{i1}, \ldots, x_{iJ_X})$, the latent covariate η_i^Z is measured by J_Z indicators $\mathbf{z}_i = (z_{i1}, \ldots, z_{iJ_Z})$, and latent outcome η_i^Y is similarly measured by J_Y indicators $\mathbf{y}_i$.

Generative Learning (SEM-VAE).

In the generative stage, our approach integrates SEM-style specifications (e.g., measurement and structural models) with deep learning to flexibly model indicators as (potentially unknown) functions of the latent variables and their (potentially unknown) structural relationships (Figure 1). Our architecture leverages domain knowledge about the factor structures and their structural connections (similar to SEM) to develop measurement-aligned, structurally-informed VAEs that jointly learn the measurement relationships linking indicators to latent variables and the structural relationships connecting the variables.

A central architectural feature distinguishing the SEM-VAE from standard VAEs is that the latent variable structure is embedded directly into the generative model rather than discovered from data. Construct-specific encoders infer each latent variable from its designated indicator block; construct-specific decoders reconstruct only that block from the corresponding latent variable; and structural networks define explicit conditional mechanisms for treatment assignment and the latent outcome. This confirmatory architecture parallels conventional SEMs in its use of domain knowledge to enforce interpretable, theory-aligned factor structure, but replaces fixed parametric equations with data-adaptive neural networks.

The SEM-VAE architecture consists of three components: encoders, decoders, and structural networks (Figure 1). In the first component, our VAE approach partitions inference pathways into theory-aligned encoders (e.g., inference/recognition networks) to produce variational parameters for the corresponding latent variable. That is, each latent variable is inferred by its own dedicated encoder network using only its designated set of indicators. In our example setting with three latent variables, we use three separate encoder networks with $\mathbf{x}_i$ as the set of indicators to measure η_i^X, $\mathbf{z}_i$ as the set that measures η_i^Z, and $\mathbf{y}_i$ as the set that measures η_i^Y.

For example, we assume the distribution of the latent variable η_i^X is approximately normal and assessed by indicators $\mathbf{x}_i$ such that

$$X_{ij} = \alpha_j^X + f_\theta(\eta_i^X) + \varepsilon_{ij}^X \qquad \varepsilon_{ij}^X \sim N\left(0, \sigma_\theta^2\right),\ \eta_i^X \sim N(0, \sigma_{\eta^X}^2) \tag{7}$$

where $f_\theta(\eta_i^X)$ is an unknown nonlinear function. A VAE approximates the latent posterior density $p\left(\eta_i^X \middle| \mathbf{x}_i\ \right)$ with the variational density

$$q_\phi\left(\eta_i^X \middle| \mathbf{x}_i\ \right) = N(\mu_\phi\left(\mathbf{x}_i\ \right), \sigma_\phi^2(\mathbf{x}_i\ )) \tag{8}$$

where μ_ϕ and σ_ϕ^2 are construct-specific neural networks.

Within each construct-specific encoder, we additionally introduce a shared low-dimensional context representation $\boldsymbol{C}_i$ to provide context-conditioned inference (e.g., Sohn et al., 2015). Our construction of the context representation is similar to that of a global context network that provides each construct encoder with a shared representation of global patterns (Zhang et al., 2025). However, our approach expands the role of the context encoder beyond the latent variable indicators to also include the observed covariates and treatment so that posterior inferences incorporate structural relationships among the latent and observed variables. As a result, our context network learns task-dependent, nonlinear sufficient features from all data that help the construct encoders infer the conditional posterior distributions of the latent variables. In our example, we define a global context representation

$$\boldsymbol{C}_i = f_{context}(\mathbf{x}_i, \mathbf{z}_i, \mathbf{y}_i, w_i, t_i) \tag{9}$$

where $f_{context}$ is a neural network producing a low-dimensional vector $\boldsymbol{C}_i$. In turn, his context representation is used in conjunction with block-specific indicators to infer variational posteriors for each latent variable. For example, the posterior distribution of the latent covariate η_i^X is

$$q_\phi\left(\eta_i^X \middle| \mathbf{x}_i\ , \boldsymbol{C}_i\right) = N(\mu_\phi\left(\mathbf{x}_i\ , \boldsymbol{C}_i\right), \sigma_\phi^2(\mathbf{x}_i\ , \boldsymbol{C}_i)) \tag{10}$$

where μ_ϕ and σ^2_ϕ are construct-specific neural networks.

Although the SEM-VAE context network leverages the full observed data vector, it does not introduce causal leakage or contamination. In causal adjustment analyses, conditioning on post-treatment information can break the identification of causal effects. The SEM-VAE, however, is not performing causal adjustment. It is performing generative imputation of latent variables that are themselves the targets of the subsequent causal analysis. The causal adjustment conditions only on the latent posterior summaries and observed pre-treatment covariates downstream in the targeted learning step and not on the indicators that fed into the encoder.

Two features of the architecture make this separation explicit. First, the latent variables are defined by the measurement and structural model as latent causes of their respective indicator blocks; they are objects of substantive interest specified in advance, not data-driven representations that the encoder constructs. The encoder's task is to approximate the posterior distribution of these prespecified latent variables given the observed data, and posterior inference should condition on all observed information jointly informative about the latent variables. Second, the TMLE step operates on the latent posterior summaries and pretreatment covariates; it does not include the observed indicators or treatment as inputs to the nuisance functions. The downstream causal adjustment set therefore contains no post-treatment information.

Conceptually, this structure mirrors established practice with plausible values, where the conditioning model used to impute latent values intentionally includes all variables that will appear in the downstream analysis (e.g., outcomes, treatment). In the plausible values literature, omitting such variables from the conditioning model introduces bias because it produces imputations that are uncongenial with the analysis model (e.g., Mislevy, 1991). The TAG framework inherits this congeniality requirement: including treatment and outcome information

in the context encoder is not a source of contamination but rather a necessary condition for the SEM-VAE's posterior summaries to be congenial with the structural model that the targeted learning step assumes (see Appendix B). Excluding this information would produce posterior summaries inconsistent with the structural relationships among the latent variables and would itself be a source of bias.

In the second conceptual component of the SEM-VAE, the decoder reconstructs each indicator block (e.g., $\mathbf{x}_i$) using only the construct-specific latent variable (e.g., η_i^X) associated with that block. Similar to other components, decoders can be structured as neural networks (e.g., when the relationships between indicators and a latent variable are unknown and/or nonlinear) or other common measurement models. For example, simple decoders can draw on linear factor models such that reconstruction of the indicators $\mathbf{x}_i$ uses a linear function of η_i^X

$$X_{ij} = \alpha_j^X + \lambda_j^X \eta_i^X + \varepsilon_{ij}^X \tag{11}$$

with identifiability from, for instance, fixing the variances of exogenous and residual variances of endogenous latent variables to one.

In the third component of our approach, we introduce structural networks for the outcome and treatment that specify the conditional distributions as flexible neural network functions, replacing the fixed parametric forms of conventional SEM with data-adaptive counterparts while preserving the assumed causal ordering among the variables. We embed outcome and propensity score models directly into the generative architecture so that the outcome and treatment are modeled as explicit conditional mechanisms. We parameterize the treatment assignment as a Bernoulli likelihood with a logit link, where the log-odds are given by a neural network g_φ with latent and observed covariates

$$T_i \sim Bernoulli(logit^{-1}(g_\varphi(\eta_i^X, \eta_i^Z, W_i))) \tag{12}$$

Similarly, we incorporate a structural outcome network as a conditionally Gaussian mechanism with mean and (log-) variance

$$\eta_i^Y \sim N(\mu_\varphi(T_i, \eta_i^X, \eta_i^Z, W_i), \sigma_\varphi^2(T_i, \eta_i^X, \eta_i^Z, W_i)) \quad (13)$$

with μ_φ and $log\ \sigma_\varphi$ produced by neural networks. Layering structural networks on top of measurement networks requires that the latent variables must simultaneously predict the indicators (through measurement networks) and the treatment and outcome (through structural networks). As a result, the model must learn posteriors that respect both measurement and structural relations.

Latent Outcome Representation

We considered three different approaches for incorporating the latent outcome into the SEM-VAE. In a first approach, we use a joint generative model in which all latent variables (i.e., including the outcome) are embedded inside the SEM-VAE with their own encoder and decoder, and the structural component defines a conditional distribution for the latent outcome given treatment and covariates. The joint approach learns measurement and structural relations simultaneously, conceptually paralleling full-information estimation in conventional SEM. However, embedding the latent outcome inside the SEM-VAE creates a tension between the identification constraints required by the measurement model and the ELBO training objective. Identification constraints on the latent outcome scale (e.g., fixed loading or fixed latent variance) potentially interact with the KL regularization in ways that can cause the optimizer to resolve identification by, for example, misallocating variance across latent variables. As a result, although a joint generative approach is conceptually appealing, it can also be sensitive to hyperparameters choices and susceptible to leakage and requires careful consideration.

For these reasons, we considered a second approach (SEM-VAE with linear CFA outcome proxy) that separated the measurement of the outcome from the full SEM-VAE model. Specifically, in the second approach we (separately) applied a linear factor measurement model for the latent outcome η_i^Y and substituted predicted factor scores ($\hat{\eta}_i^Y$) as an observed outcome proxy in the SEM-VAE. That is, the SEM-VAE was fit with the outcome proxy ($\hat{\eta}_i^Y$) entering as an observed outcome in the structural likelihood (e.g., η_i^Y is replaced by $\hat{\eta}_i^Y$).

Under a linear factor measurement model, outcome factor scores ($\hat{\eta}_i^Y$) can be expressed as a linear combination of indicators (Appendix F) with the conditional expectation as

$$E[\hat{\eta}^Y \,|\eta = E[\mathbf{w}^T(\boldsymbol{\Lambda}\eta + \boldsymbol{\xi})|\, \eta] = \mathbf{w}^T\boldsymbol{\Lambda}\eta + \mathbf{w}^T E[\boldsymbol{\xi}|\eta]) = (\mathbf{w}^T\boldsymbol{\Lambda})\eta \tag{14}$$

with $\mathbf{w}^T$ as the weights and $\boldsymbol{\Lambda}$ as the factor loadings. Although there are many different factor scoring methods (e.g., regression) with alternative properties, two common approaches are regression factor scores and Bartlett factor scores. Prior research has widely demonstrated that Bartlett factor scores have several desirable properties when used as a proxy for latent outcomes (but not necessarily latent covariates) and the inferential goal is the structural effect of treatment on the outcome (e.g., Bartlett, 1937; Croon, 2002; Skrondal & Laake, 2001). In linear measurement models, for example, Bartlett scores ($\hat{\eta}_i^{Y^B}$) provide conditionally unbiased predictions

$$E\left[\hat{\eta}_i^{Y^B}\middle|\eta_i^Y\right] = (\mathbf{w}^T\boldsymbol{\Lambda})\eta_i^Y = (\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda})^{-1}\boldsymbol{\Lambda}'\boldsymbol{\Psi}^{-1}\boldsymbol{\Lambda}\eta_i^Y = [I]\eta_i^Y = \eta_i^Y \tag{15}$$

When the measurement error is independent of the treatment and covariates, the conditional expectation of the outcome based on Bartlett scores is unbiased

$$E\left[\hat{\eta}_i^{Y^B}\middle|\eta_i^Y, T, \boldsymbol{X}\right] = E\left[\hat{\eta}_i^{Y^B}\middle|\eta_i^Y\right] = \eta_i^Y \tag{16}$$

As a result, the use of Bartlett factor scores in place of the latent outcome preserves the measurement interpretation of the outcome without introducing bias (with higher variance) while also simplifying the SEM-VAE because the outcome enters as an observed variable.

A similar analysis of regression factor scores ($\hat{\eta}_i^{Y^R}$) suggests their conditional expectation is scaled by the reliability (ρ_Y) and can be written as a function of Bartlett scores

$$E\left[\hat{\eta}_i^{Y^R}\middle|\eta_i^Y\right] = E\left[\rho_Y\hat{\eta}_i^{\boldsymbol{B}}\middle|\eta_i^Y\right] = \rho_Y\eta_i^Y \tag{17}$$

As a result, the use of regression factor scores in place of the latent outcome can also be used to simplify the SEM-VAE model as long as the effects are scaled by the reliability.

Conceptually, this sequential approach that substitutes outcome factor scores in the SEM-VAE separates measurement of the outcome from the structural components in the SEM-VAE. The outcome proxy is determined solely by the outcome measurement model and is not subject to ELBO/KL regularization or conditional-prior shrinkage that can arise when the outcome is modeled as a latent variable inside the SEM-VAE. As a result, the approach has the potential to ameliorate leakage and outcome variation compression relative to the full SEM-VAE approach.

In a third approach (SEM-VAE with CFA-VAE outcome proxy), we expand this further to replace the linear factor decoder with a more flexible neural network decoder (i.e., CFA-VAE) for the latent outcome. Subsequently, we employ scores of the latent outcome from the posterior distribution of the CFA-VAE as a substitute for the latent outcome in the SEM-VAE. Similar to the second approach, the outcome proxy is not subject to the aforementioned regularization or conditional-prior shrinkage found in the full SEM-VAE, thereby reducing the potential for outcome variance compression and attenuated treatment effects. The utility of this third approach is that it additionally introduces more measurement model flexibility for the outcome because it leverages neural decoders rather than a simple linear measurement model for the decoder (e.g.,

Milano et al., 2024). For example, the approach can handle more complicated settings where an observed indicator Y_{ij} is a nonlinear function of latent outcome η_i^Y

$$Y_{ij} = \alpha_j^Y + f_\theta(\eta_i^Y) + \varepsilon_{ij}^Y \qquad \varepsilon_{ij}^Y \sim N\left(0, \sigma_\theta^2\right),\ \eta_i^Y \sim N(0, \sigma_{\eta^Y}^2) \tag{18}$$

where $f_\theta(\eta_i^Y)$ is a nonlinear function (e.g., Milano et al., 2024). When the indicators are a linear function of the latent variable (i.e., $f_\theta(\eta_i^Y) = \alpha_j^Y + \lambda_j^X \eta_i^X$), the decoder reduces to the conventional linear measurement model outlined in the second approach (see Appendix D for comparing linear and nonlinear measurement models).

Model Training

For each of these approaches, we trained the models by maximizing an ELBO that appropriately combines: (i) likelihood terms for each indicator block given its latent construct, (ii) the treatment likelihood given latent and observed covariates, (iii) the outcome likelihood given latent and observed covariates and the treatment, and (iv) Kullback–Leibler (KL) regularization terms that shrink each latent variable's approximate posterior toward its prior distribution (see Appendix C for outline). We maximize the ELBO with stochastic gradient methods using mini-batches and the reparameterization trick for Gaussian variational posteriors. Each construct encoder outputs the mean and variance for each latent variable and latent samples are drawn via $\eta = \mu + \sigma \odot \epsilon$ with $\epsilon \sim N(0, I)$. These samples jointly drive the measurement decoders (e.g., $p(\mathbf{x}|\eta)$) and the structural heads for the treatment (e.g., $p(t|\eta^X, \eta^Z, w)$) and outcome (e.g., $p(\eta^Y | t, \eta^X, \eta^Z, w)$), with parameters updated by backpropagation through the entire system. We regularize via KL terms to standard-normal priors for each latent variable (β = 1), and we train components jointly for the full SEM-VAE approach and separately for the other approaches.

Latent Variable Posterior Summaries

In turn, we then generate summaries of the posterior distribution for each latent variable based on the SEM-VAE. Although there are many alternative approaches, we adopt a bias avoiding strategy suitable for subsequent targeted learning (e.g., Devlieger et al., 2016; Skrondal & Laake, 2001). Conceptually, we use posterior means or regression-like scores for latent covariates and Bartlett-like scores for the latent outcome. Under a linear measurement model, the posterior mean of the latent variable from a CFA-VAE is the same as the regression factor score when we do not condition on context. When the measurement model is nonlinear, the posterior mean is intractable and we approximated it using a local linearization at η_t^Y by applying a first-order Taylor expansion update around the initial estimate and then a linear-Gaussian Kalman update with the Jacobian of the nonlinear function in place of the linear observation matrix that moves η_t^Y toward its conditional expectation given the indicator responses. The result is a first-order approximation to the posterior mean (and regression-like scores). Similarly, under a linear measurement model for the outcome we use conventional Bartlett scores. When the outcome measurement model is nonlinear, we considered a nonlinear generalization of Bartlett scores using a locally unbiased, Gauss–Newton generalization of Bartlett scores that linearizes f_θ at η_t^Y. The nonlinear Bartlett-like scores provide a first-order approximation for a nonlinear version of an unbiased, minimum measurement-error variance estimator that is similar to Bartlett scores in a linear factor model (Appendix F).

Targeted Learning. SEM VAEs are optimized for reconstruction/prediction losses and are well-suited for density modeling. However, SEM-VAEs are not aligned with a specific causal effect and prediction losses are not necessarily orthogonal to nuisance errors. For example, small errors or regularization in nuisance components (e.g., outcome networks, propensity score networks, encoders) can translate directly into bias when estimating the

treatment effect in spite of a highly predictive model. As a result, optimizing a SEM-VAE model does not necessarily provide an unbiased estimate of the treatment effect.

For this reason, our framework subsequently employs targeted learning with posterior summaries generated by the SEM-VAE. Targeted learning is built to estimate a specific causal effect by constructing an efficient influence function (EIF) for that effect and targeting that score with cross-fitted nuisance estimates. The use of an orthogonal estimating equation makes the first order bias in the estimated effect small even when the nuisance models are imperfect. More generally, the approach yields doubly robust estimators that are asymptotically linear, robust to regularization bias, and provide valid inference even when the nuisance models are learned by flexible ML algorithms.

Our approach uses TMLE with the super learner on summaries of the SEM-VAE posterior distribution (e.g., posterior means, Bartlett-like scores) for each latent variable (e.g., $\hat{\eta}^X, \hat{\eta}^Z$) to estimate the average treatment effect (ATE). TMLE constructs an initial estimate of the conditional outcome

$$Q_0\left(t, \eta^X, \eta^Z, W\right) = E\left[\hat{\eta}^Y \mid T = t, \hat{\eta}^X, \hat{\eta}^Z, W\right] \tag{19}$$

and the propensity score

$$g_0\left(\eta^X, \eta^Z, W\right) = P(T = 1 \mid \hat{\eta}^X, \hat{\eta}^Z, W) \tag{20}$$

using super learner libraries to reduce the risk of misspecification bias. TMLE then implements a one-dimensional fluctuation of Q_0 with the clever covariate

$$H\left(T, \eta^X, \eta^Z, W\right) = \frac{I(t=1)}{g(1|\eta^X, \eta^Z, W)} - \frac{I(t=0)}{g(0|\eta^X, \eta^Z, W)} \tag{21}$$

yielding an updated regression Q^* that solves the efficient influence function (EIF) estimating equation. The ATE is the targeted plug-in

$$ATE = \frac{1}{n}\sum_1^n\{Q^*\ (1, \eta_i^X, \eta_i^Z, W_i) - Q^*\ (0, \eta_i^X, \eta_i^Z, W_i)\} \quad (22)$$

and its variance is estimated by the sample variance of the estimated EIF divided by the sample size.

Under a linear factor model, for example, Bartlett factor scores can be expressed as $\hat{\eta}_i^{Y^B} = \eta_i^Y + \varepsilon_i^Y$ with error (ε_i^Y) that is independent of treatment after conditioning on covariates ($E[\varepsilon_i^Y \mid T, \boldsymbol{X}] = 0$). The EIF of the ATE with TMLE is then

$$D(O) = H(T, \boldsymbol{X})\{\eta^Y - Q(T, \boldsymbol{X})\} + [Q(1, \boldsymbol{X}) - Q(0, \boldsymbol{X}) - \delta^{\eta^Y}] \quad (23)$$

where $Q(t, \boldsymbol{X}) = E[\eta^Y \mid T = t, \boldsymbol{X}]$, $H(T, \boldsymbol{X}) = \frac{I(t=1)}{g(1|\boldsymbol{X})} - \frac{I(t=0)}{g(0|\boldsymbol{X})}$, $g(t|\boldsymbol{X}) = P(T = t|\boldsymbol{X})$, and δ^{η^Y} is the ATE. Because η^Y enters the EIF linearly through the residual ($\eta^Y - Q(T, \boldsymbol{X})$), substituting, for instance, Bartlett scores (i.e., replacing η_i^Y with $\hat{\eta}_i^{Y^B} = \eta_i^Y + \varepsilon_i^Y$) preserves the target and the double robust property for the ATE when the measurement models and latent posterior learning is correct (see Appendix B). More conceptually, additive, mean-zero outcome error perturbs the EIF but does not shift the parameter; that is, using, for example, Bartlett scores inflates variance without inducing bias. By contrast, covariates (e.g., via the $Q(t, \boldsymbol{X})$, $g(t|\boldsymbol{X})$ functions) enter the EIF nonlinearly. As a result, measurement error in covariates (even when mean-zero) can change $Q(t, \boldsymbol{X})$ and $g(t|\boldsymbol{X})$ in ways that introduce bias because it perturbs the nuisance functions themselves rather than just adding a linear term in the latent outcome.

Congeniality of Targeted and Generative Models. A central requirement for valid inference in the TAG framework is congeniality between the generative model used to recover latent variables and the targeted learning procedure used to estimate causal effects. The concept of congeniality originates in the multiple-imputation and plausible values literatures, where it refers to the alignment between how missing or latent quantities are generated and how they are

subsequently analyzed. In the TAG setting this principle takes on a more specific and consequential form: the SEM-VAE must produce latent variable summaries whose induced conditional structure is sufficiently compatible with the nuisance models that targeted learning relies upon. Targeted learning does not operate on the joint distribution directly; it relies on the conditional outcome regression and the propensity score to identify and estimate the causal estimand. If the generative model fails to represent key features of these conditional relationships, targeted learning cannot recover from that failure and bias introduced at the generative stage propagates into the effect estimate and is not corrected by the targeting step. There are three primary considerations for congeniality. First, measurement congeniality requires that the variational posteriors produced by the SEM-VAE encoders are sufficiently close to the true conditional distributions of the latent variables given their indicators. When measurement congeniality holds, the posterior summaries used in TMLE behave as locally unbiased proxies for the true latent quantities: substituting them into the efficient influence function perturbs variance but does not shift the estimand. When it fails, the resulting posterior summaries are systematically distorted, and this distortion propagates directly into the nuisance functions and the effect estimate.

Second, structural congeniality requires the that the SEM-VAE adequately captures the conditional relationships among latent covariates, observed covariates, treatment assignment, and the outcome. Even though targeted learning is robust to model misspecification in the nuisance functions, it is not robust to confounding that remains unmeasured because the generative model failed to appropriately capture the latent structure. Put differently, violations of treatment ignorability at the latent level cannot be repaired downstream.

Third, functional congeniality requires that the outcome proxy substituted into the efficient influence function be conditionally unbiased for the true latent outcome given treatment and covariates. Bartlett factor scores satisfy this condition under a correctly specified linear measurement model; the nonlinear Gauss-Newton Bartlett-like scores from the CFA-VAE satisfy it approximately under nonlinear measurement (Appendix F). When functional congeniality holds, outcome measurement error enters the EIF additively with zero conditional mean, inflating variance but not introducing bias.

In the TAG framework, congeniality is plausibly supported across a broad range of settings by leveraging expressive neural networks in the generative model. With suitably expressive architectures, SEM-VAEs can flexibly approximate both complex measurement relationships and nonlinear, nonadditive structural relationships among latent and observed variables. However, expressiveness does not guarantee congeniality in finite samples. For example, architectural choices in the SEM-VAE implicitly constrain the validity of the causal analysis. Decisions about indicator routing, decoder flexibility, structural network expressiveness, and regularization effectively define an implicit prior over the latent data that targeted learning receives as input. When this implicit prior is well-aligned with the substantive and causal structure of the problem, congeniality is plausible (but still not guaranteed). When the prior is not well-aligned, targeted learning cannot compensate. Appendix B outlines congeniality conditions, provides a bias decomposition when they fail, and states the convergence rates that would be required for double robustness and semiparametric efficiency to hold.

Simulation

We conducted an initial evaluation of the proposed framework using simulated data. We considered a base condition with two latent covariates (e.g., η_i^X) and an observed covariate (W_i)

that were sampled from a multivariate standard normal distribution with all covariances set to $\Sigma_{ij} = 0.3$, an observed binary treatment (T_i), and a latent outcome (η_i^Y). We extended this to additional conditions that considered five latent covariates and ten observed covariates as well as more complex data generating processes (Appendix E). For the base conditions, data were generated using linear or nonlinear factor models with the structural model for the treatment assignment and outcome following nonlinear (quadratic) and nonadditive (two-way interactions between all latent variables) relationships with the latent covariates and additive relationships for the observed covariates

$$\eta_i^Y = \beta_0 + \delta T_i + \sum_{j=1}^{J} \beta_{X_j} \eta_i^{X_j} - \sum_{j=1}^{J} \beta_{X_j^{(2)}} (\eta_i^{X_j})^2 - \sum_{1 \le j \le l \le J} \beta_{X_j X_l} \eta_i^{X_j} \eta_i^{X_l} + \sum_{k=1}^{K} \beta_{W_k} W_{ik} + \varepsilon_i \quad (24)$$

$$logit(T_i) = \gamma_0 + \sum_{j=1}^{J} \gamma_{X_j} \eta_i^{X_j} + \sum_{j=1}^{J} \gamma_{X_j^{(2)}} (\eta_i^{X_j})^2 + \sum_{1 \le j \le l \le J} \gamma_{X_j X_l} \eta_i^{X_j} \eta_i^{X_l} + \sum_{k=1}^{K} \gamma_{W_k} W_{ik}$$

The outcome error variance followed a normal distribution centered at zero and variance of one minus the variance of the structural components. The coefficients were selected so that ignoring nonlinear and nonadditive terms introduced a downward bias. Across simulations, we varied the sample size, the number of indicators per latent variable, the factor loadings, the reliabilities of the latent variables, the nonlinearity of the measurement models, the approach for incorporating the outcome (see section above), the machine learning algorithms used in the super learner, the number of latent and observed covariates, the number of latent variable interactions and quadratic terms, and the underlying data generating process (see Appendix E).

Our SEM VAE is implemented in R using the torch package as a custom module with amortized encoders for the latent variables and a shared context network (see Appendix E for details; code implementing the analyses is available at BLINDED). We report the bias in the ATE estimator, its standard deviation and its root mean-square error (RMSE) based on 100

replications for each condition. To benchmark the performance, we compared the performance of TAG learning with several alternative approaches that are each unworkable or ineffective in practice. Our first comparison was with an idealized model—a nonlinear SEM (NLSEM). We considered a correctly specified model (e.g., expression 24). However, this approach is unworkable in practice for two reasons. First, our analysis focuses on estimating treatment effects when the correct model specification is unknown and this approach requires and makes use of the known true specification. Second, estimation feasibility for NLSEM progressively diminishes with each additional nonlinear or nonadditive term for latent variables. We limit the scope of our comparisons between NLSEM and TAG learning to quadratic and two-way interaction terms for latent variables to ensure estimation feasibility for NLSEM. In practice, however, more complex nonlinear (e.g., cubic, nonparametric) and nonadditive (e.g., three-way interactions) relationships are plausible. While TAG learning expands easily to accommodate such complex relationships, the computational demand for estimation for SEM is prohibitive and unworkable in practice. Our second comparison illustrates this principle by further contrasting TAG learning with a linear SEM (LSEM) that incorrectly omits nonlinear and nonadditive terms.

Our third comparison was with another idealized model—a nonlinear regression with the latent variable values produced by generative learning (nonlinear regression after generative; NLRAG). This approach is unworkable in practice because it also assumes the correct structural model specification when it would be unknown in practice. However, comparisons between TAG and NLRAG can be used to separate how much of the observed bias in TAG learning is attributable to the generation learning phase (identical for TAG and NLRAG) and how much owes to the targeted learning phase. When NLRAG estimates are nearly unbiased but TAG learning is biased, the implication is that the generative learning is working well but that bias

enters through targeted learning's data adaptive estimation of the nuisance functions. Our fourth comparison illustrates misspecification effects by additionally contrasting TAG learning with a linear regression with the latent variable values produced by generative learning (linear regression after generative; LRAG) that incorrectly omits nonlinear and nonadditive).

The fifth comparison captures a viable alternative but one that ignores measurement error—targeted learning with Bartlett factor scores for the latent outcome and regression factor scores for the latent covariates (TMLE-FS). Comparisons between TAG learning and TMLE-FS outline the bias in targeted learning introduced by ignoring measurement error.

The sixth comparison considers nonlinear regression with Bartlett factor scores for the latent outcome and regression factor scores for the latent covariates (NLR-FS). The approach is unworkable because it presumes knowledge of the correct structural model while also ignoring measurement error. Similarly, the seventh comparison considers a linear regression with those factor scores (LR-FS) that incorrectly omits nonlinear and nonadditive terms. These comparators serve as baselines to outline how much bias reduction is provided by TAG learning and also reflect historical practice (e.g., Castro Torres & Akbaritabar, 2024).

Table 1 and Figures 2-3 summarize the simulation results (see Appendix E for full results including for null effect conditions). The results consistently demonstrated that TAG learning provides nearly unbiased effect estimates, even under adverse conditions (e.g., low reliabilities). Across the base simulation conditions, TAG and NLRAG tracked closely, providing empirical support that the full-data SEM-VAE approach introduced negligible within-sample bias in the current settings (e.g., Table 1; Figure 2; Appendix Table 2 and 4). The collective results from all alternative methods were unsurprising and reiterated a well-established warning—ignoring/misspecifying the measurement and/or structural models introduces bias into effect

estimates. For instance, with linear measurement models, 2000 observations, three indicators per latent variable, and low to moderate reliability (roughly 0.6), TAG centered closely on the true effect, yielding essentially zero bias (Figure 2; Table 1). In contrast, approaches that either ignored measurement error (e.g., TMLE-FS) or misspecified the structural model (linear SEM) exhibited sizable bias. Further, when measurement models were nonlinear, TAG learning remained nearly unbiased while NLSEM and other methods were substantially biased (Figure 3; Appendix Tables 4 and 5). Even in the idealized but impossible-in-practice nonlinear SEM approach that exploited additional knowledge regarding the correct structural form (that was unavailable to TAG learning), treatment effect estimates were biased when the true measurement models were nonlinear (but treated as linear).

At larger sample sizes or high reliabilities, TAG learning broadly demonstrated little bias (Table 1, Figures 2-3; Appendix Tables 2-8). At smaller sample sizes and lower reliabilities, consistent with reduced information for both measurement and structural learning, TAG displayed small amounts of bias but still substantially outperformed workable alternatives. In higher dimensional conditions, the pattern was broadly similar but unbiased estimation required larger sample sizes or higher reliabilities (e.g., Figure 3; Appendix Tables 6 and 7). For example, with ten observed covariates and five latent covariates (reliabilities between 0.6 and 0.7) that entered nonlinearly and nonadditively, TAG learning exhibited a small negative bias (-0.03) in samples of 10,000. When the sample dropped to 5000, the bias rose to -0.06 but was reduced to -0.04 when reliabilities increased to 0.7 and 0.8. With reliabilities around 0.6 and a sample of 2000, TAG learning had a bias of -0.15; with reliabilities increased to around 0.8, bias was -0.08.

Comparisons among the three different approaches for incorporating the latent outcome suggested that while all provide nearly unbiased estimation with moderate sample sizes, the TAG

estimator based on the full SEM-VAE tended to be much more dispersed (Table 1, Appendix Tables 2-8). This pattern is consistent with the identification/regularization tensions noted earlier—when the latent outcome is modeled inside the SEM-VAE, identification constraints interact with the ELBO/KL regularization and conditional priors in ways that misallocate or compress variance (e.g., shift variability across latent variance, decoder residuals, or structural components). The Bartlett-proxy approach severs this coupling to improve precision without introducing bias. Finally, this study was not preregistered. Large language models (Microsoft, 2026) were used to edit the manuscript and R code (available at BLINDED and in Appendix G).

Illustration

We illustrate TAG learning using data from an open-access online personality study (n = 4,270) that combined Duckworth and colleagues' (2007) Grit Scale with the International Personality Item Pool (IPIP) Big Five personality inventory (see Appendix G for details). Grit—defined by Duckworth et al. (2007) as perseverance of effort and consistency of interest over long-term goals—serves as the latent outcome, operationalized by three Grit Scale indicators (reliability = 0.48). Two Big Five personality dimensions serve as latent covariates: Agreeableness (reliability = 0.46) and Conscientiousness (reliability = 0.54), each measured by three indicators. Our analysis examines the effect of completing a college or graduate degree on grit while controlling for these two personality covariates. We emphasize that this example is intended solely as a convenient and accessible methodological illustration rather than a substantive empirical investigation, and the resulting estimates should not be interpreted as evidence of a causal effect of educational attainment on grit for numerous reasons (Appendix G).

We estimated TAG learning using linear measurement models for all latent variables and the SEM-VAE with linear CFA outcome proxy (Approach 2). Both the SEM-VAE and the

targeted learning step were cross-fit using K = 5 folds, with SEM-VAE training in each fold governed by Adam optimization and early stopping based on a 10% within-training-fold validation split. The super learner ensemble for TMLE nuisance function estimation matched the algorithms used in the base simulations.

EIF-based standard errors produced by TMLE capture sampling uncertainty from the nuisance function estimation and targeting step but does not account for the additional uncertainty introduced by latent variables and estimation of the SEM-VAE parameters. To propagate uncertainty into effect estimates, we used nonparametric case-based bootstrapping (e.g., Cai et al., 2020; Efron & Tibshirani, 1994). We sampled a new dataset of size n with replacement from the original sample, stratified by treatment assignment. The complete TAG pipeline was then re-estimated on each bootstrapped sample. Standard errors and percentile-based 95% confidence intervals were constructed using 100 bootstrap replications.

The results are summarized in Table 9. Differences among estimators are consistent with the possibility that linear models and conventional factor-score approaches fail to capture features of the structural relationships, although the illustration is not designed to isolate whether these differences arise from nonlinearity, measurement error, model regularization, or residual confounding. The results additionally suggest that the naïve EIF-based standard errors underestimate effect uncertainty.

Discussion

Recent research has demonstrated the power of ML to uncover structure in complex data and deliver accurate predictions. Yet the predominance of questions in social science research do not rely on prediction problems alone (e.g. Brand et al., 2023). Core lines of inquiry in the social sciences involve building durable theories that integrate substantive, measurement and causal

theories. Standard ML approaches rarely engage with such complexities. To capitalize on recent advances, we need approaches that leverage ML as a fortifying tool—rather than a predictive substitute—that integrates ML with substantive, measurement and causal theories. TAG learning contributes to this gap by pairing theory aligned generative models for latent variables with targeted, semi-parametric estimators for causal effects. TAG learning harnesses the algorithmic power of ML to construct theory-grounded data-driven scientific analysis.

The results of our analyses demonstrated that TAG learning consistently produced nearly unbiased estimates across a wide range of conditions. A core consideration with ML methods in the social sciences is sample size. The results suggested that each of the SEM-VAE approaches considered were feasible with finite samples with the CFA-VAE outcome proxy approaches (Approaches 2 and 3) producing lower dispersion than the full SEM-VAE (Approach 1). The result is consistent with the argument that decoupling outcome measurement from the joint ELBO objective prevents identification constraints from interacting destructively with KL regularization. More generally, the combination of theory-driven measurement constraints, amortized variational inference, and targeted estimation appears to reduce the effective sample size requirement relative to unconstrained deep learning approaches—a practical advantage in the social sciences where samples in the thousands often represent a ceiling rather than a floor.

Limitations. Several limitations qualify our findings. We examined a limited set of structural and measurement models and more generally a limited set of conditions in our simulations; although TAG learning demonstrated strong performance in these conditions, more comprehensive investigations are needed to outline the conditions under which the approach performs well. For example, we partitioned indicators into mutually exclusive blocks; future applications should examine multidimensionality and sensitivity to cross loadings and method

effects. For example, although the core SEM-VAE routes each indicator block to a construct specific encoder, instruments often contain cross cutting measurement artifacts (e.g., method effects, correlated errors, or subtle cross loadings) that spill across constructs. Prior architectures have suggested nuisance encoders to address measurement artifacts (e.g., common error spanning indicators) (e.g., Zhang et al., 2025). Conceptually, the nuisance latent variable absorbs global measurement variation, allowing construct specific latent variables to remain clean and factor aligned. In turn, cleaner latent variables stabilize nuisance models in targeted learning by reducing residual confounding caused by mis-measured latent covariates.

Similarly, our simulations did not assess the coverage or finite-sample validity of bootstrapped confidence intervals. Although the theoretical justification for bootstrapping with multi-stage estimators is well-established, its behavior under the full range of simulation conditions (e.g., small samples, low reliabilities, high-dimensional covariate spaces) remains unexamined (e.g., Efron & Tibshirani, 1994). Characterizing the empirical coverage of bootstrap CIs across these conditions, and identifying when the EIF-based standard error is a sufficient approximation, is an important direction for future work.

We also assumed measurement invariance and normal distributions for latent variables (e.g., Lyu et al., 2025; Ren et al., 2025; Wang et al., 2023); both can be relaxed (e.g., group specific parameters or non Gaussian priors). We also relied on a domain knowledge to determine measurement and structural models; conclusions therefore inherit any misspecification of that knowledge. For example, if a latent covariate was a post-treatment variable (e.g., mediator) conditioning on that variable would undermine unbiased effect estimation.

Further, the results suggested that the sufficiency of a sample size was a function of multiple variables including model complexity and latent variable reliabilities—more complex

nuisance functions and less reliable factors generally required larger samples for data adaptive approaches to effectively reduce bias to zero. Our analyses suggested that bias can arise from either side of the pipeline (generative or targeted learning). For example, when the sample size was small relative to model complexity (e.g., high dimensional latent covariates) and latent variables were unreliable, bias largely entered through the generative phase (e.g., Table 7). In contrast, in lower dimensional settings, bias largely entered through targeted learning. These results potentially suggest that as analyses become more complex, constructing and tuning each phase (e.g., richer super learner libraries with highly adaptive learners and fold wise targeting) takes on increasing importance. Future research needs to more carefully examine the conditions under which bias arises from the generative and/or targeted learning phases, data requirements that support nearly unbiased estimation for different model complexity levels, and architectural modifications and hyper parameter tuning that improves estimation.

Conclusion. There are substantial opportunities to advance the social sciences through ML. However, to realize these opportunities in the social sciences, we must build ML methods that integrate and respect substantive, measurement, and causal theory. TAG illustrates a path toward respecting measurement theory, model structural complexity, and obtain valid causal inferences on latent outcomes. Taken together, the simulations suggest the methodological promise of pairing appropriate treatment of measurement error with data-adaptive modeling of complex structure. Establishing the general theoretical properties of the resulting estimator, and characterizing its behavior across a broader range of conditions, are important directions for future work.

References

Bai, H., Liu, X., Bai, F., Chen, Y., & Pandohie, R. (2022). Machine Learning Method for High-Dimensional Education Data. Journal of Methods and Measurement in the Social Sciences, 13(1), 1-14.

Bainter, S. A., & Bollen, K. A. (2014). Interpretational confounding or confounded interpretations of causal indicators?. Measurement: Interdisciplinary Research & Perspectives, 12(4), 125-140.

Bareinboim, E., Correa, J. D., Ibeling, D., & Icard, T. (2022). On Pearl's hierarchy and the foundations of causal inference. In Probabilistic and Causal Inference: The Works of Judea Pearl (pp. 507–556). ACM.

Bartlett, J. W., & Hughes, R. A. (2020). Bootstrap inference for multiple imputation under uncongeniality and misspecification. Statistical methods in medical research, 29(12), 3533-3546.

Bogaert, J., Loh, W. W., & Rosseel, Y. (2026). Consistent Factor Score Regression: A Better Alternative for Uncorrected Factor Score Regression? Educational and Psychological Measurement, 0(0).

Bollen, K. A. (2002). Latent variables in psychology and the social sciences. Annual review of psychology, 53(1), 605-634.

Borsboom, D. (2008). Latent variable theory.

Borgstede, M., & Eggert, F. (2023). Squaring the circle: From latent variables to theory-based measurement. Theory & Psychology, 33(1), 118-137.

Brand, J. E., Zhou, X., & Xie, Y. (2023). Recent developments in causal inference and machine learning. Annual Review of Sociology, 49, 81–110.

Cai, W. & van der Laan, M. (2020). Nonparametric bootstrap inference for the targeted highly adaptive least absolute shrinkage and selection operator (LASSO) estimator. The International Journal of Biostatistics, vol. 16, no. 2, pp. 20170070.

Casella, M., Milano, N., Esposito, R., & Marocco, D. (2024, October). Modeling Nonlinear Relationships in Psychometric Data Using Variational Autoencoders: Insights from Simulated Data. In 2024 IEEE International Conference on Metrology for eXtended Reality, Artificial Intelligence and Neural Engineering (MetroXRAINE) (pp. 1040-1044). IEEE.

Castro Torres, A. F., & Akbaritabar, A. (2024). The use of linear models in quantitative research. Quantitative Science Studies, 5(2), 426-446.

Cheung, A. C., & Slavin, R. E. (2016). Effects of Success for All on reading achievement: A secondary analysis using data from the Study of Instructional Improvement (SII). AERA Open, 2(4), 2332858416674902.

Cho, A. E., Wang, C., Zhang, X., & Xu, G. (2021). Gaussian variational estimation for multidimensional item response theory. British Journal of Mathematical and Statistical Psychology, 74, 52-85.

Connor, C., Kelcey, B., Sparapani, N., Petscher, Y., Siegal, S., Adams, A., Hwang, J., & Carlisle, J. (2020). Predicting Second and Third Graders' Reading Comprehension Gains: Observing Students' and Classmates Talk During Literacy Instruction using COLT. Scientific Studies of Reading, 24, 5, 411-433.

Cox, K. & Kelcey, B. (2019). Optimal Design of Cluster-Randomized and Multisite Studies Using Fallible Outcome Measures. Evaluation Review, 43, 3-4, 189-225.

Cui, C., Wang, C., & Xu, G. (2024). Variational Estimation for Multidimensional Generalized Partial Credit Model. Psychometrika.
Wang, X., H. Chen, S. Tang, Z. Wu and W. Zhu, Disentangled Representation Learning, in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 46, no. 12, pp. 9677-9696.
Dawson, A. Z., Walker, R. J., Gregory, C., & Egede, L. E. (2020). Examination of the Association Between Latent Variables for Social Determinants of Health and Blood Pressure Control in Immigrants using Structural Equation Modeling. Journal of the National Medical Association, 112(2), 186–197.
Devlieger, I., Mayer, A., & Rosseel, Y. (2016). Hypothesis testing using factor score regression: A comparison of four methods. Educational and psychological measurement, 76(5), 741-770.
Devlieger, I., & Rosseel, Y. (2020). Multilevel factor score regression. Multivariate Behavioral Research, 55(4), 600-624.
Dwyer, J., Kelcey, B., Berebitsky, D., & Carlisle, J. (2016). A Study of Teachers' Discourse Moves That Support Text-based Discussions. Elementary School Journal 117, 285-309.
Efron, B., & Tibshirani, R. J. (1994). An introduction to the bootstrap. Chapman and Hall/CRC.
Eignor, D. R. (2013). The standards for educational and psychological testing.
Fan, Y., Chen, J., Shirkey, G., John, R., Wu, S. R., Park, H., & Shao, C. (2016). Applications of structural equation modeling (SEM) in ecological studies: an updated review. Ecological processes, 5(1), 19.
Finn, A., & Wang, L. (2014). Formative vs. reflective measures: Facets of variation. Journal of Business Research, 67(1), 2821-2826.
Hainmueller, J., Mummolo, J., & Xu, Y. (2019). How Much Should We Trust Estimates from Multiplicative Interaction Models? Simple Tools to Improve Empirical Practice. Political Analysis, 27(2), 163–192.
Hayes, T., & Usami, S. (2020). Factor score regression in the presence of correlated unique factors. Educational and Psychological Measurement, 80(1), 5-40.
Ichimura, H., & Newey, W. K. (2022). The influence function of semiparametric estimators. Quantitative Economics, 13(1), -.
Jewsbury, P.A., Jia, Y. & Gonzalez, E.J. Considerations for the use of plausible values in large-scale assessments. Large-scale Assess Educ 12, 24 (2024).
Jo, B., Hastie, T. J., Li, Z., Youngstrom, E. A., Findling, R. L., & Horwitz, S. M. (2023). Reorienting Latent Variable Modeling for Supervised Learning. Multivariate Behavioral Research, 58(6), 1057–1071.
Kelleher, M., Kinnear, B., Sall, D., Schumacher, D., Schauer, D., Warm, E., & Kelcey, B. (2020). A Reliability Analysis of Entrustment-Derived Workplace-Based Assessments. Academic Medicine, 95, 616-622.
Kelcey, B., McGinn, D., & Hill, H. (2014). Approximate Measurement Invariance in Cross-classified Rater-Mediated Assessments. Frontiers in Psychology, 5, 1-13.
Kelcey, B., Bai, F., Xie, Y., & Cox, K. (2020). Micro-Macro and Macro-Micro Effect Estimation in Small Scale Latent Variable Models with Croon's Method. Testing, Psychometrics & Methodology in Applied Psychology, 27, 3, 477-500.
Khorramdel, L., von Davier, M., Gonzalez, E., Yamamoto, K. (2020). Plausible Values: Principles of Item Response Theory and Multiple Imputations. In: Maehler, D., Rammstedt, B. (eds) Large-Scale Cognitive Assessment . Methodology of Educational Measurement and Assessment. Springer, Cham.

Kilian, P., Leyhr, D., Urban, C. J., Höner, O., & Kelava, A. (2023). A deep learning factor analysis model based on importance-weighted variational inference and normalizing flow priors: Evaluation within a set of multidimensional performance assessments in youth elite soccer players. Statistical Analysis and Data Mining: The ASA Data Science Journal, 16(5), 474-487.

Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep conditional generative models. Advances in neural information processing systems, 28: 3483–3491, 2015.

Kline, R. B. (2012). Assumptions in structural equation modeling. Handbook of structural equation modeling, 111, 125.

Liu, T., Wang, C., & Xu, G. (2022). Estimating three-and four-parameter MIRT models with importance-weighted sampling enhanced variational auto-encoder. Frontiers in Psychology, 13, 935419.

Lyu, W., Wang, C., & Xu, G. (2025). Detecting differential item functioning across multiple groups using group pairwise penalty. Psychometrika, 1-28.

Marsh, H. W., & Hau, K. T. (2007). Applications of latent-variable models in educational psychology: The need for methodological-substantive synergies. Contemporary educational psychology, 32(1), 151-170.

Meng, X. L. (1994). Multiple-imputation inferences with uncongenial sources of input. Statistical science, 538-558.

Microsoft (2026). Copilot (GPT-4) [Large Language Model]. https://copilot.microsoft.com/

Milano, N., Casella, M., Esposito, R., & Marocco, D. (2024). Exploring the potential of variational autoencoders for modeling nonlinear relationships in psychological data. Behavioral Sciences, 14(7), 527.

Moccia, C., Moirano, G., Popovic, M., Pizzi, C., Fariselli, P., Richiardi, L., … Maule, M. (2024). Machine learning in causal inference for epidemiology. European Journal of Epidemiology, 39, 1097–1108. https://doi.org/10.1007/s10654-024-01173-x

Molenaar, D., Grasman, R. P., & Cúri, M. (2025). Autoencoders for Amortized Joint Maximum Likelihood Estimation of Confirmatory Item Factor Models. Multivariate Behavioral Research, 1-21.

Muthén, B., & Muthén, B. O. (2009). Statistical analysis with latent variables (Vol. 123, No. 6, pp. 55-112). New York: Wiley.

Nagengast, B., & Trautwein, U. (2015). The prospects and limitations of latent variable models in educational psychology. In Handbook of educational psychology (pp. 55-72). Routledge.

Neyshabur, B., Bhojanapalli, S., McAllester, D., & Srebro, N. (2017). Exploring generalization in deep learning. Advances in neural information processing systems, 30.

Ramspek, C. L., Steyerberg, E. W., Riley, R. D., Rosendaal, F. R., Dekkers, O. M., Dekker, F. W., & van Diepen, M. (2021). Prediction or causality? A scoping review of their conflation within current observational research. European journal of epidemiology, 36(9), 889-898.

Raudenbush, S. W. (2008). Advancing educational policy by advancing research on instruction. American Educational Research Journal, 45(1), 206-230.

Raza, D. (2024). "Why Machine Learning Falls Short for Causal Estimation," Towards Data Science/Medium

Regier, M., Moodie, E. & Platt, R. (2014). The Effect of Error-in-Confounders on the Estimation of the Causal Parameter When Using Marginal Structural Models and Inverse Probability-

of-Treatment Weights: A Simulation Study. The International Journal of Biostatistics, 10(1), 1-15.
Ren, H., Lyu, W., Wang, C., & Xu, G. (2025). A Novel Method for Detecting Intersectional DIF: Multilevel Random Item Effects Model with Regularized Gaussian Variational Estimation. Psychometrika, 1-25.
Rowan, B., & Correnti, R. (2009). Studying reading instruction with teacher logs: Lessons from the study of instructional improvement. Educational Researcher, 38(2), 120-131.
Royston, P. (2004). Multiple imputation of missing values. The Stata Journal, 4(3), 227-241.
Sankaran, K. (2023). Bootstrap Confidence Regions for Learned Feature Embeddings. Journal of Computational and Graphical Statistics, 32(4), 1337-1347.
Shin, M., Cho, H., Min, H. S., & Lim, S. (2021). Neural bootstrapper. Advances in Neural Information Processing Systems, 34, 16596-16609.
Shmueli, G. (2010). To explain or to predict?. Statistical science, 289-310.
Stoetzer, L. F., Zhou, X., & Steenbergen, M. (2025). Causal inference with latent outcomes. American Journal of Political Science, 69(2), 624-640.
Sukserm, P. (2024). Understanding Latent Variables in EFL Contexts. Shanlax International Journal of Education, 12(4), 60-69.
Tamano, H., Hino, H., & Mochihashi, D. (2025). Misspecifying non-compensatory as compensatory IRT: analysis of estimated skills and variance. Behaviormetrika, 52(2), 393-412.
Thomas N. (2002). The role of secondary covariates when estimating latent trait population distributions. Psychometrika, 67, 33-48
van der Laan, M. J., Polley, E. C., & Hubbard, A. E. (2007). Super learner. Statistical applications in genetics and molecular biology, 6, Article25.
van der Laan, M. J., & Rose, S. (2011). Targeted Learning: Causal Inference for Observational and Experimental Data. Springer.
Van Buuren, S., & Van Buuren, S. (2012). Flexible imputation of missing data (Vol. 10, p. b1182). Boca Raton, FL: CRC press.
Veldkamp, K., Grasman, R., & Molenaar, D. (2025). Handling missing data in variational autoencoder based item response theory. British Journal of Mathematical and Statistical Psychology, 78(1), 378-397.
von Davier M., Sinharay S., Oranje A., Beaton A. E. (2007). The statistical procedures used in National Assessment of Educational Progress: Recent developments and future directions. In Rao C. R., Sinharay S. (Eds.), Handbook of statistics (pp. 1039-1055). Amsterdam, Netherlands: North-Holland.
von Davier M., Sinharay S. (2013). Analytics in international large-scale assessments: Item response theory and population models. In Rutkowski L., von Davier M., Rutkowski D. (Eds.), Handbook of international large-scale assessment: Background, technical issues, and methods of data analysis (pp. 155-174). Boca Raton, FL: CRC Press.
Wang, C., Zhu, R., & Xu, G. (2023). Using lasso and adaptive lasso to identify DIF in multidimensional 2PL models. Multivariate Behavioral Research, 58(2), 387-407.
Xia, K., & Bareinboim, E. (2021). The causal-neural connection: Expressiveness, learnability, and inference. Advances in Neural Information Processing Systems, 34
Zhong, Z., Guo, H. & Qian, K. (2024). Deciphering the impact of machine learning on education: Insights from a bibliometric analysis using bibliometrix R-package. Educ Inf Technol 29, 21995–22022.

Figure 1: Conceptual mapping of a (a) SEM-VAE encoders and decoders for each the latent variable and of a (b) SEM-VAE that uses CFA-VAE scores as an observed proxy for the latent outcome

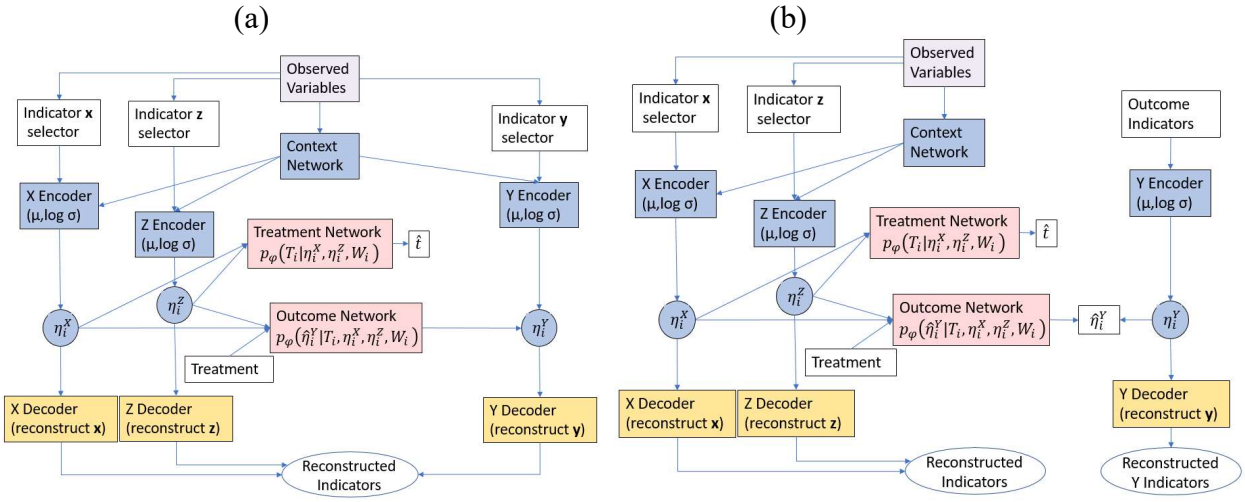

Figure 2 Simulation results outlining bias as a function of reliability, sample size, and method under base conditions for each of the three approaches for handling the outcome in the SEM-VAE

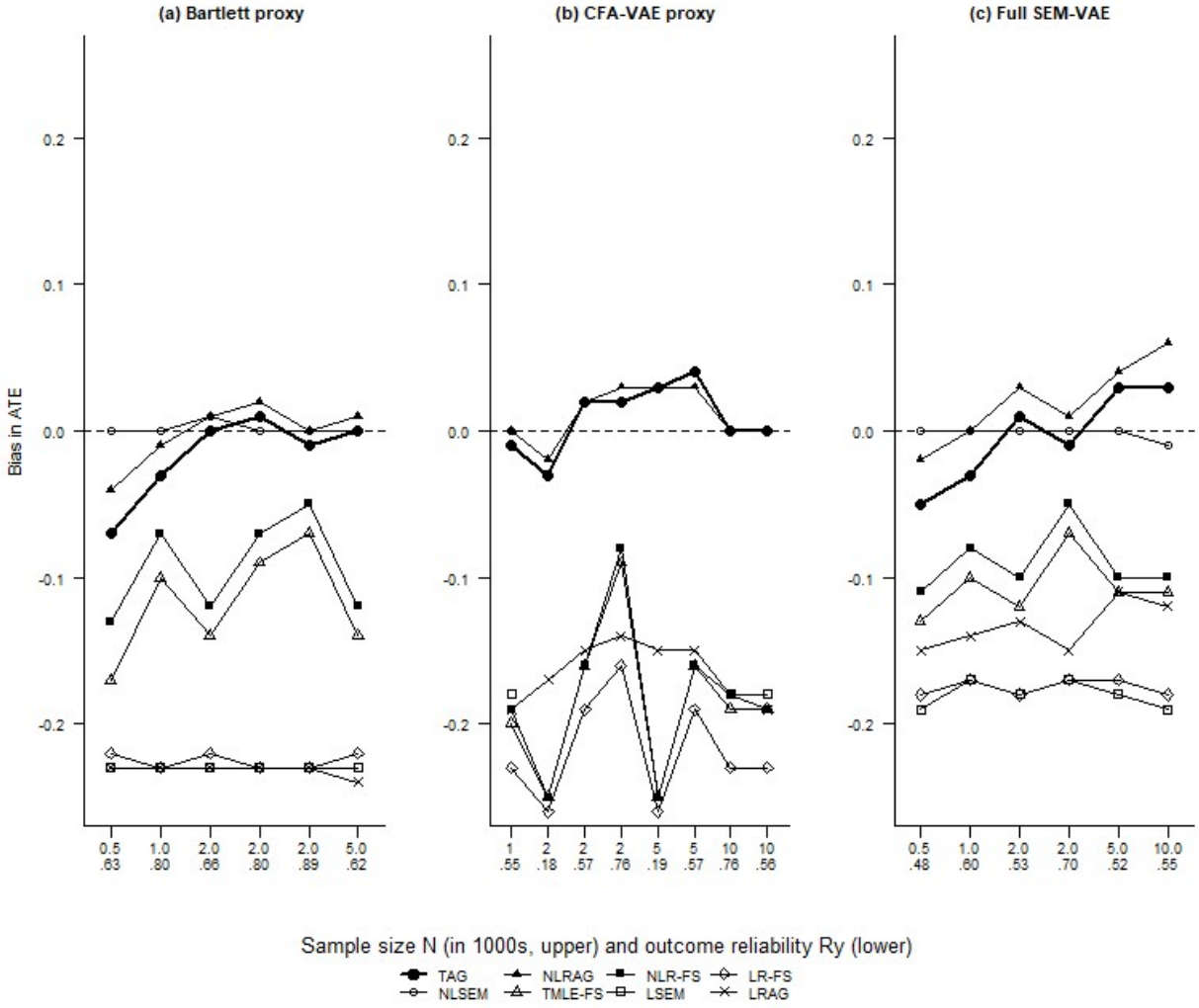

Figure 3 Simulation results outlining bias as a function of reliability, sample size, and method under nonlinear measurement models or linear measurement with high dimensional covariates set

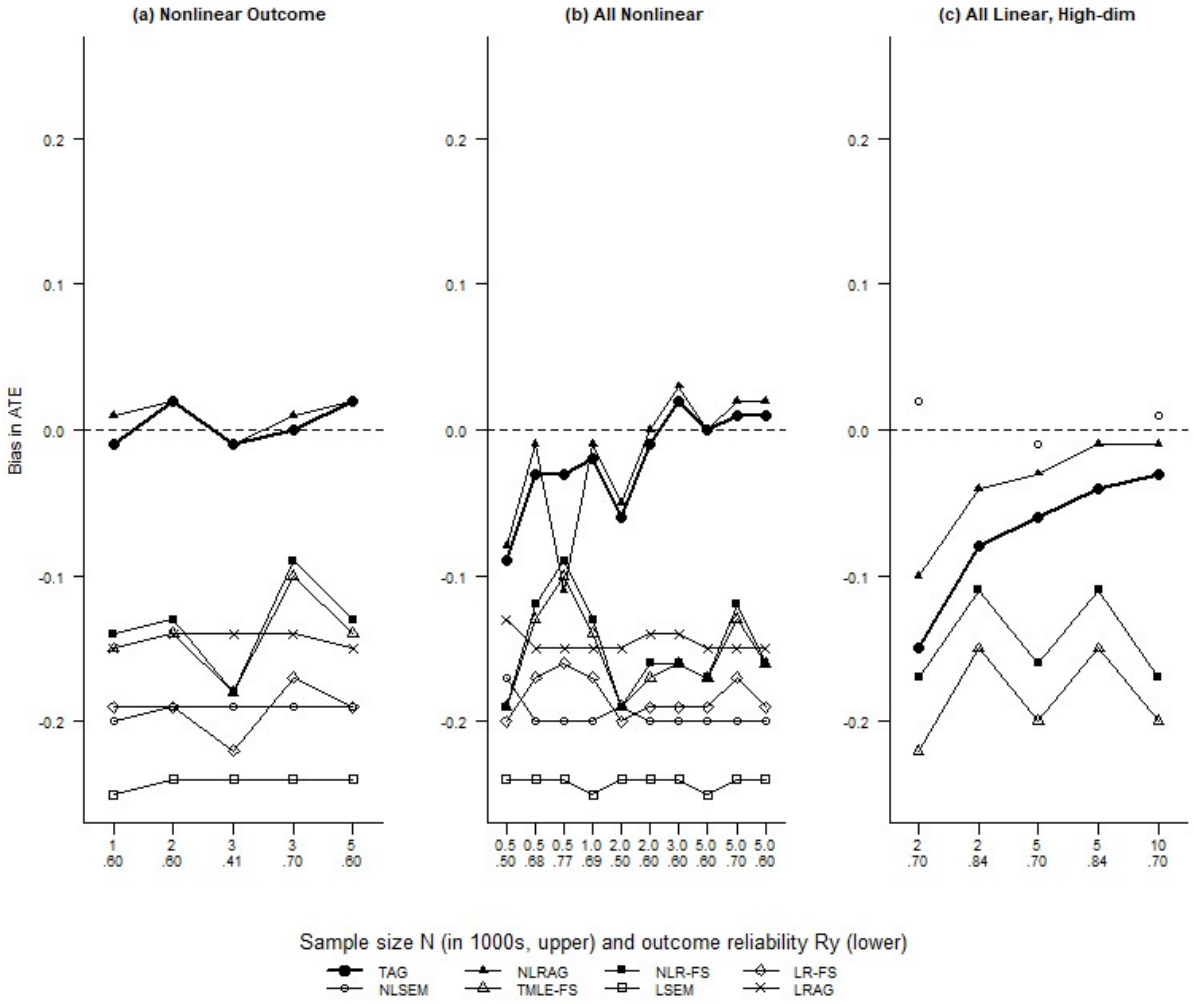

Table 1 *Summary of simulation results under linear measurement models, structural models that follow expression (24) with 2 latent covariates and 1 observed covariate, and Bartlett scores as an outcome proxy in the SEM-VAE.*

N	R_y	R_x	R_z	Items	Estimator	Bias	SD	RMSE
500	0.63	0.57	0.58	3	NLSEM	0.00	0.17	0.17
					TAG	-0.07	0.14	0.16
					TMLE-FS	-0.17	0.13	0.21
					LSEM	-0.23	0.13	0.26
					LR-FS	-0.22	0.13	0.26
					NLR-FS	-0.13	0.13	0.18
					LRAG	-0.23	0.13	0.26
					NLRAG	-0.04	0.14	0.15
1000	0.8	0.79	0.8	10	NLSEM	0.00	0.08	0.08
					TAG	-0.03	0.08	0.09
					TMLE-FS	-0.10	0.08	0.13
					LSEM	-0.23	0.08	0.25
					LR-FS	-0.23	0.08	0.24
					NLR-FS	-0.07	0.08	0.11
					LRAG	-0.23	0.08	0.25
					NLRAG	-0.01	0.08	0.08
2000	0.66	0.57	0.56	3	NLSEM	0.01	0.07	0.07
					TAG	0.00	0.08	0.08
					TMLE-FS	-0.14	0.06	0.16
					LSEM	-0.23	0.06	0.24
					LR-FS	-0.22	0.06	0.23
					NLR-FS	-0.12	0.06	0.14
					LRAG	-0.23	0.07	0.24
					NLRAG	0.01	0.07	0.07
2000	0.8	0.79	0.8	10	NLSEM	0.00	0.06	0.06
					TAG	0.01	0.06	0.06
					TMLE-FS	-0.09	0.06	0.11
					LSEM	-0.23	0.06	0.24
					LR-FS	-0.23	0.06	0.24
					NLR-FS	-0.07	0.06	0.09
					LRAG	-0.23	0.06	0.24
					NLRAG	0.02	0.06	0.06
2000	0.89	0.88	0.88	20	NLSEM	0.00	0.06	0.06
					TAG	-0.01	0.06	0.06
					TMLE-FS	-0.07	0.06	0.09
					LSEM	-0.23	0.06	0.24
					LR-FS	-0.23	0.06	0.24
					NLR-FS	-0.05	0.06	0.07
					LRAG	-0.23	0.05	0.24
					NLRAG	0.00	0.06	0.06

5000	0.62	0.6	0.59	3	NLSEM	0.00	0.05	0.05
					TAG	0.00	0.05	0.05
					TMLE-FS	-0.14	0.04	0.15
					LSEM	-0.23	0.04	0.23
					LR-FS	-0.22	0.04	0.23
					NLR-FS	-0.12	0.04	0.13
					LRAG	-0.24	0.04	0.24
					NLRAG	0.01	0.05	0.05

Note. Super learner uses GLM, mean, GAM, and MARS

Table 2 *Illustration example results*

Estimator	Est	SE	CI
TAG (naïve uncertainty)	0.11	0.04	[0.03, 0.19]
TAG (bootstrap uncertainty)		0.05	[0.03, 0.22]
TMLE-FS	0.23	0.04	[0.15, 0.31]
LSEM	0.20	0.07	[0.06, 0.34]
LR-FS	0.24	0.04	[0.16, 0.32]