Causal Inference with Ordinal Outcomes: Two Approaches Using Density Estimation Techniques

Chanhyuk Park

Abstract

Ordinal outcomes, such as Likert-type survey responses, are ubiquitous in political science, yet their implications for causal inference are often overlooked. This paper shows that, even under standard causal assumptions, treatment effects on an underlying continuous attitude can be identified from ordinal outcomes only up to scale. This fundamental limitation undermines common empirical practices: treating ordinal responses as cardinal and applying OLS yields estimates whose magnitudes and signs depend arbitrarily on chosen numeric labels, while parametric ordinal models, such as ordered logit and probit, rely on rigid distributional assumptions that induce bias under misspecification. This paper proposes the *Normalized Latent Treatment Effect (NLTE)* as a causal estimand that ensures comparability across models and samples by standardizing the latent scale. To estimate the NLTE, this paper introduces two flexible estimators that recover latent treatment effects without specifying a parametric error distribution: a semiparametric kernel density estimator and a normalizing-flow-based maximum likelihood estimator. Monte Carlo simulations demonstrate that these estimators remain consistent across diverse latent error distributions where standard models exhibit substantial bias. Replications of Tomz and Weeks (2020) and Mattingly et al. (2025) demonstrate that standard approaches can lead to misleading substantive conclusions.

1. Introduction

With the growing emphasis on causal inference, political science has increasingly focused on identifying the effects of policies, institutions, and information. Experimental and quasi-experimental designs, such as survey experiments, difference-in-differences, and regression discontinuity designs, have become standard empirical strategies across subfields. Survey experiments are especially prominent because random assignment promises clean identification; indeed, from 2021 to 2024, roughly 20% of articles in the top three general political science journals (*American Political Science Review, American Journal of Political Science, and Journal of Politics*) included a survey experiment as part of their empirical strategy.

Much of this research studies outcomes that are inherently latent and subjective, such as attitudes toward redistribution (Alt and Iversen, 2017; Magni, 2021), approval of political leaders (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009), or foreign policy preferences (Tomz and Weeks, 2020). These concepts are typically assumed to lie on a unidimensional latent continuum but are observed through ordinal response categories. Likert-type items, ranging from *strongly disagree* to *strongly agree*, are the most common example. Although such scales are intended to capture gradations of an underlying attitude, they encode only order, not the distances between categories.

Despite this limitation, applied work commonly treats ordinal responses in ways that implicitly assume cardinal information. One of the most frequent practices is *cardinalization*, which assigns numeric scores (e.g., 1 to 5) to categories and applies OLS. Because these numeric labels are arbitrary, the resulting estimates depend on the chosen scoring scheme and are difficult to interpret as effects on the latent outcome (Schröder and Yitzhaki, 2017; Bond and Lang, 2018; Bloem, 2022). Different plausible labelings can change the magnitude—and in some cases, even the sign—of estimated effects. Another common strategy is to collapse multi-category outcomes into binary indicators (e.g., support vs. oppose). Binarization discards information about intermediate categories, yielding inefficient estimators and potentially masking substantively important shifts in the response distribution.

A third class of approaches uses parametric ordinal regression models, such as ordered logit and ordered probit. These methods respect the ordered nature of the data and invoke a latent-variable representation. However, such models rest on strong assumptions about the distribution of the latent error term (standard logistic distribution for logit and standard normal distribution for probit). When these assumptions are violated, due to the reasons such as polarization (bimodality), social desirability (skewness) in the underlying preferences, or unobserved confounders, estimators converge to biased parameters. The resulting errors can be substantial, leading to incorrect inferences about treatment intensity or even sign reversals in finite samples

(Manski, 1988; Johnston, McDonald and Quist, 2020).

This paper contributes to the study of causal inference with ordinal outcomes regarding both identification and estimation. First, we clarify the limits of identification in this setting. Even under standard causal assumptions (SUTVA, ignorability, and positivity), treatment effects on the continuous latent outcome are identified only up to scale. Because the observed ordinal responses are invariant to positive monotonic transformations, the absolute scale of the latent average treatment effect cannot be pinned down without additional assumptions. To address this, we propose the *Normalized Latent Treatment Effect (NLTE)* as a formal causal estimand. By standardizing the latent scale (e.g., via error-variance normalization), the NLTE ensures that treatment effects are comparable across models and samples, providing a transparent metric for latent treatment intensity.

Second, we introduce two flexible estimators that recover the NLTE without imposing a parametric form on the error distribution. The first is a semiparametric kernel density estimation (KDE)–based quasi-maximum likelihood estimator that nonparametrically estimates the error distribution from the data. The second employs normalizing flows (NF), a class of invertible transformations, to model the error density within a maximum likelihood framework. Both approaches allow the latent distribution to take its true shape, while remaining robust to the misspecification that plagues parametric models.

We evaluate these methods using Monte Carlo simulations and two empirical applications. The simulations demonstrate that while OLS is sensitive to arbitrary labels and parametric models are biased under non-normal errors, the KDE and NF-based estimators remain consistent across a wide range of distributions. The empirical applications revisit Tomz and Weeks (2020) and Mattingly et al. (2025), both of which rely on cardinalized or binarized OLS to study international cues. In our reanalysis of Tomz and Weeks (2020) on human rights and military force, we find that standard approaches significantly understate the latent treatment intensity. In the replication of Mattingly et al. (2025) on authoritarian propaganda, the semiparametric estimators reveal that the perceived superiority of Chinese state cues over American cues is largely a methodological artifact of OLS's inability to handle floor effects. In both cases, the proposed estimators provide a substantively different evidence of the persuasive power of experimental treatments.

The remainder of the paper is organized as follows. Section 2 formalizes the causal inference problem with ordinal outcomes and establishes the identification result. Section 3 reviews existing approaches and introduces the two density-estimation-based estimators. Section 4 presents Monte Carlo evidence on finite-sample performance. Section 5 applies the methods to the Tomz and Weeks (2020) survey experiment. Section 7 concludes with implications for applied work and directions for future research.

2. Causal Inference with Ordinal Outcomes

Many causal questions in political science concern attitudes and preferences that are inherently latent, such as support for redistribution (Alt and Iversen, 2017; Magni, 2021), immigration (Mayda and Rodrik, 2005), trust in institutions, or foreign policy orientations (Scheve and Slaughter, 2001; Tomz and Weeks, 2020). These constructs are typically assumed to lie on a unidimensional latent continuum, yet are almost always observed through ordinal response categories: Likert-type items (*strongly disagree* to *strongly agree*), ordered approval levels, or frequency scales. Researchers then apply standard causal designs such as survey or field experiments, difference-in-differences, regression discontinuity to these ordinal outcomes.

In practice, applied work commonly handles ordinal outcomes in one of three ways. First, many studies assign numeric scores (for example, 1 to 5) to categories and then apply difference-in-means or OLS, implicitly treating the outcome as cardinal. Second, it is common to collapse multi-category responses into binary indicators (for example, support vs. oppose) before estimation. Third, analysts often fit ordered logit or ordered probit models, or item response theory (IRT) models when multiple items are available. Each of these strategies embeds strong, and often implicit, assumptions about how the observed categories relate to the underlying latent attitude. As we show in this section, ordinal responses provide only limited information about the latent outcome, so both identification and interpretation of causal effects must be approached with care.

2.1. Latent-Variable Framework and Causal Estimand

We formalize the setting in a standard threshold crossing latent-variable framework. Let D_i denote a (possibly vector-valued) treatment indicator, and let X_i be a vector of observed covariates. The latent continuous outcome Y_i^* represents respondent i's true attitude or preference (for example, support for using military force). We posit the partially linear latent outcome model

$$Y_i^* = f(X_i, \beta) + D_i^\top \tau + \varepsilon_i,$$

where $f(X_i, \beta)$ is a known function of X_i indexed by parameters β, τ is the vector of treatment effects on the latent scale, and ε_i is an unobserved scalar error term with cumulative distribution function F.

The observed ordinal outcome $Y_i \in \{0, 1, \ldots, J\}$ arises from applying a *reporting function* g that maps the latent attitude Y_i^* into ordered response categories. We begin with a common-threshold specification.

Assumption 1 (Common Reporting Function). *There exist threshold parameters* $-\infty = \alpha_{-1} < \alpha_0 < \cdots <$

$\alpha_J = \infty$ *such that*

$$Y_i = g(Y_i^*) = j \quad \Longleftrightarrow \quad \alpha_{j-1} < Y_i^* \leq \alpha_j, \qquad j = 0, 1, \ldots, J.$$

All individuals share the same threshold vector $\alpha = (\alpha_0, \ldots, \alpha_{J-1})$.

This assumption abstracts from interpersonal differences in how respondents interpret response categories. We relax it later by allowing thresholds to depend on observed characteristics X_i, analogously to generalized ordered logit/probit and partial proportional odds models (Williams, 2006).[1]

To discuss causal inference side, we adopt the potential outcomes framework (Rubin, 1974). Let $Y_i^*(d)$ denote the potential latent outcome under treatment status $D_i = d$, with d ranging over the support of D_i (for simplicity, think of $d \in \{0, 1\}$ in the binary case). The corresponding observed latent outcome satisfies $Y_i^* = Y_i^*(D_i)$, and the observed ordinal outcome is $Y_i = g(Y_i^*(D_i))$.

In usual case, the target estimand is the average treatment effect on the latent scale,

$$\text{ATE}^* = \mathbb{E}\big[Y_i^*(1) - Y_i^*(0)\big],$$

where the expectation is taken over the population of interest. When D_i is a vector of treatments, we are typically interested in the components of τ, which capture marginal effects of each treatment on Y_i^*.

We impose the usual causal assumptions on the latent outcome.

Assumption 2 (SUTVA). *The stable unit treatment value assumption holds on the latent scale:* $Y_i^* = Y_i^*(D_i)$, *and there is no interference between units.*

Assumption 3 (Ignorability). *Conditional on* X_i, *treatment is as good as randomly assigned with respect to the latent potential outcomes:*

$$\big(Y_i^*(0),\, Y_i^*(1)\big) \perp\!\!\!\perp D_i, |, X_i.$$

Assumption 4 (Positivity). *For all covariate values with positive probability, treatment assignment is nondegenerate:*

$$0 < \mathbb{P}(D_i = 1 \mid X_i) < 1 \quad \textit{with probability one.}$$

If Y_i^* were observable directly, these assumptions would suffice to identify ATE^* using standard methods.

[1] Additional design-based approaches, such as anchored vignettes (King et al., 2004; King and Wand, 2007), can also be used to address heterogeneity in reporting functions.

Under the assumptions above and for a binary treatment, we have

$$\text{ATE}^* = \mathbb{E}\big[\, Y_i^*(1) - \, Y_i^*(0)\big] = \mathbb{E}\big[\, Y_i^* \mid D_i = 1, X_i\big] - \mathbb{E}\big[\, Y_i^* \mid D_i = 0, X_i\big],$$

and ATE^* is point-identified. In randomized experiments, a simple difference-in-means estimator and OLS can estimate the estimand. In observational studies, regression adjustment, inverse probability weighting, or doubly robust estimators can be used.

The central difficulty in our setting is that Y_i^* is not observed, but we only see its ordinal representation Y_i. When we observe only the ordinal outcome Y_i, the joint distribution of $(\, Y_i, X_i, D_i)$ does not uniquely pin down the scale or location of the latent outcome Y_i^*. In other words, we do not have any good information on whether the latent outcome Y_i^* spans from 0 to 1 or -100 to 100. The key reason is that the ordinal model is invariant to positive affine transformations of the latent variable and thresholds.

For given parameters (α, β, τ, F), the conditional probability of observing category j is

$$\begin{aligned}
\mathbb{P}(\, Y_i = j \mid X_i, D_i) &= \mathbb{P}\big(\alpha_{j-1} < \, Y_i^* \le \alpha_j \big| X_i, D_i\big) \\
&= \mathbb{P}\big(\alpha_{j-1} < f(X_i, \beta) + D_i^\top \tau + \varepsilon_i \le \alpha_j \big| X_i, D_i\big) \\
&= F\big(\alpha_j - f(X_i, \beta) - D_i^\top \tau\big) - F\big(\alpha_{j-1} - f(X_i, \beta) - D_i^\top \tau\big).
\end{aligned}$$

Now consider any constants $a \in \mathbb{R}$ and $c > 0$, and define the transformed latent variable and thresholds

$$\tilde{Y}_i^* = a + c\, Y_i^*, \qquad \tilde{\alpha}_j = a + c\alpha_j.$$

Let $\tilde{\varepsilon}_i = c\varepsilon_i$ and $\tilde{F}$ be the CDF of $\tilde{\varepsilon}_i$. Then

$$\tilde{Y}_i^* = a + c\, f(X_i, \beta) + c D_i^\top \tau + \tilde{\varepsilon}_i = f\big(X_i, \tilde{\beta}\big) + D_i^\top \tilde{\tau} + \tilde{\varepsilon}_i,$$

where, for a linear f, $\tilde{\beta} = c\beta$ and $\tilde{\tau} = c\tau$. The corresponding conditional probabilities are

$$\begin{aligned}
\mathbb{P}(\, Y_i = j \mid X_i, D_i) &= \mathbb{P}\big(\tilde{\alpha}_{j-1} < \, \tilde{Y}_i^* \le \tilde{\alpha}_j \big| X_i, D_i\big) \\
&= \tilde{F}\big(\tilde{\alpha}_j - f(X_i, \tilde{\beta}) - D_i^\top \tilde{\tau}\big) - \tilde{F}\big(\tilde{\alpha}_{j-1} - f(X_i, \tilde{\beta}) - D_i^\top \tilde{\tau}\big),
\end{aligned}$$

which coincide with the original probabilities because the transformation simply rescales both the index and the thresholds. Thus the observed distribution of $(\, Y_i, X_i, D_i)$ cannot distinguish between (α, β, τ) and $(a + c\alpha, c\beta, c\tau)$.

THEOREM 1 (Scale and Location Non-Identification). *Any parameter vector* (α, β, τ) *and its positive affine transformation* $(a + c\alpha, c\beta, c\tau)$ *with* $c > 0$ *induce the same distribution for the observed ordinal outcome* Y_i *conditional on* (X_i, D_i)*. Consequently, the absolute scale and location of the latent outcome* Y_i^**, and hence of ATE***, are not identified from ordinal data alone.*

This result implies that only *scale-invariant* features of the model are empirically meaningful. These include the sign of treatment effects and the ordering of coefficients by magnitude, but not their absolute size on the latent scale. Any comparison of coefficients, either across models or across samples, therefore requires an explicit normalization.

2.2. Transformations of Ordinal Outcomes: Cardinalization and Binarization

Researchers often attempt to circumvent the limitations of ordinal data by transforming the outcome before estimation. Two common transformations are cardinalization and binarization. We first discuss them conceptually and then use a simple example to illustrate how they can fail.

By *cardinalization* we mean assigning numeric labels $l_0 < \cdots < l_J$ to the ordinal categories and treating the resulting variable $\tilde{Y}_i = l_{Y_i}$ as if it were a correct representation for the latent outcome Y_i^*. Researchers then estimate treatment effects using difference-in-means, OLS, or related methods. A typical choice is $(1, 2, 3, 4, 5)$ for a five-point Likert item, but any strictly increasing set of labels would be admissible from the perspective of order.

Two problems arise with cardinalization of the ordinal outcomes. First, because the labels are arbitrary, the magnitude of the estimated effect depends on the particular labeling scheme. Replacing $(1, 2, 3, 4, 5)$ with $(2, 4, 6, 8, 10)$ doubles the estimated difference-in-means without changing the underlying data. More generally, different plausible labelings can generate substantively different estimates and, in some cases, even flip the sign of the estimated effect (Schröder and Yitzhaki, 2017; Bond and Lang, 2018; Bloem, 2022). Second, there is no guarantee that any given set of labels corresponds to the true, unknown transformation g from the latent attitude to the observed categories. By putting specific numerical labels, the analyst is effectively imposing a particular form of g without justification.

Binarization is an coarser transformation, collapsing several ordinal categories into a single binary indicator (for example, treating *strongly agree* and *agree* as 1 and all other responses as 0). Although this approach has some gains from interpretation of the coefficients (Breen, Karlson and Holm, 2018), it discards all information about movement within the merged categories and about distinctions among intermediate categories, reducing statistical efficiency and potentially obscuring substantively meaningful shifts in attitudes. Some articles in clinical studies suggest that the use of ordinal outcomes can significantly increase the power of the study and

thus reduce the required sample size by 2 to 3 times smaller compared to the binary outcomes (Armstrong and Margaret, 1989; D'Amico et al., 2020). Moreover, binarization cannot remedy the identification problem discussed above. It merely defines a different, coarser estimand based on coarser outcomes that may be quite distant from the underlying latent ATE.

To make these issues concrete, consider a simple finite-sample example with five respondents, labeled A through E. Suppose we know each respondent's latent potential outcomes $Y_i^*(0)$ and $Y_i^*(1)$ as in Table 1. These can be thought of as their true attitudes toward a policy under control and treatment, respectively.

Respondent	$Y_i^*(0)$	$Y_i^*(1)$
A	1.37	0.09
B	-0.56	1.71
C	0.36	0.11
D	0.63	2.22
E	0.40	0.14

TABLE 1. Latent potential outcomes in the illustrative example.

The true latent ATE based on Tabel 1 is

$$\begin{aligned} \text{ATE}^* &= \mathbb{E}\big[\, Y_i^*(1) - \, Y_i^*(0)\big] \\ &= \frac{0.09 + 1.71 + 0.11 + 2.22 + 0.14}{5} \\ &= \frac{1.37 - 0.56 + 0.36 + 0.63 + 0.40}{5} \approx 0.41. \end{aligned}$$

This suggests that on average, the treatment increases the latent attitude by about 0.41 units.

Now suppose that we do *not* observe $Y_i^*(d)$, but only the ordinal responses generated by applying a reporting function g. Consider two plausible reporting functions, g_a and g_b, which differ in how they map latent attitudes into five ordered categories: *Strongly Disagree*, *Disagree*, *Neither*, *Agree*, and *Strongly Agree*. Table 2 shows the resulting observed potential outcomes under each g.

	g_a		g_b	
Respondent	$Y_{i,g_a}(0)$	$Y_{i,g_a}(1)$	$Y_{i,g_b}(0)$	$Y_{i,g_b}(1)$
A	Agree	Agree	Agree	Disagree
B	Neither	Agree	Disagree	Agree
C	Agree	Agree	Neither	Disagree
D	Agree	Strongly Agree	Neither	Agree
E	Agree	Agree	Neither	Disagree

TABLE 2. Observed ordinal potential outcomes under two reporting functions g_a and g_b.

Both g_a and g_b are monotone reporting functions: higher latent values correspond to "more favorable" categories. However, they differ in how they partition the latent scale. For instance, under g_b the thresholds

for middle categories may be shifted relative to g_a, so that the same latent change produces different observed category movements.

Suppose we now analyze these data using the standard cardinalization that assigns numeric scores $1, 2, 3, 4, 5$ to *Strongly Disagree*, *Disagree*, *Neither*, *Agree*, and *Strongly Agree*, respectively. For each reporting function, we can compute the average cardinalized outcome under control and treatment and their difference. Under g_a,

$$\widehat{\text{ATE}}_{\text{card},g_a} = \frac{4+3+4+4+4}{5} - \frac{4+4+4+5+4}{5} = \frac{19}{5} - \frac{21}{5} = -0.4,$$

based on Table 2. Reversing the order (control minus treatment) gives $+0.4$, which happens to be close in magnitude to the true latent ATE of 0.41. In this case, the usual 1–5 coding is "lucky": it yields a cardinalized ATE that approximates the latent ATE.

Under g_b, applying the *same* 1–5 coding to the ordinal categories yields

$$\widehat{\text{ATE}}_{\text{card},g_b} = \frac{4+2+3+3+3}{5} - \frac{2+4+2+4+2}{5} = \frac{15}{5} - \frac{14}{5} = 0.2,$$

or, with the opposite ordering of potential outcomes, -0.2. Either way, the cardinalized ATE under g_b is far from the true latent ATE of 0.41 and can even have the wrong sign depending on whether we subtract control from treatment or vice versa. Since we do not observe Y_i^* or know the true g, we cannot tell whether we are in the "lucky" case (g_a) or the "unlucky" case (g_b). This illustrates that even the standard 1–5 labeling can be highly misleading, depending on the unknown reporting function.

If we consider binarizing the outcome by defining

$$Y^{\text{bin}} = \begin{cases} 1, & \text{if } Y \in \text{Agree}, \text{Strongly Agree}, \\ 0, & \text{otherwise.} \end{cases}$$

Table 3 reports the binarized potential outcomes and the implied average treatment effects under each reporting function.

	g_a		g_b	
Respondent	$Y^{\text{bin}}i, g_a(0)$	$Y^{\text{bin}}i, g_a(1)$	$Y^{\text{bin}}i, g_b(0)$	$Y^{\text{bin}}i, g_b(1)$
A	1	1	1	0
B	0	1	0	1
C	1	1	0	0
D	1	1	0	1
E	1	1	0	0

TABLE 3. Binarized potential outcomes (1 if *Agree* or *Strongly Agree*, 0 otherwise).

Under g_a, the binarized means are

$$\mathbb{E}[Y^{\text{bin}} g_a(0)] = \frac{4}{5} = 0.8, \qquad \mathbb{E}[Y^{\text{bin}} g_a(1)] = \frac{5}{5} = 1.0,$$

so the binarized ATE is 0.2. Under g_b,

$$\mathbb{E}[Y^{\text{bin}} g_b(0)] = \frac{1}{5} = 0.2, \qquad \mathbb{E}[Y^{\text{bin}} g_b(1)] = \frac{2}{5} = 0.4,$$

again yielding a binarized ATE of 0.2. In both cases, binarization detects a positive effect but compresses the information in the five categories into a single bit, substantially understating the magnitude of change relative to the latent ATE of 0.41 and failing to reflect differences between g_a and g_b. Also, the results depend on how we collapse the ordinal outcome. If we were to choose binarize the outcomes with 1 with *Stongly Agree* and 0 otherwise, the resulting estimates become different from what we have.

2.3. Normalization of Latent Treatment Effect

The scale and location non-identification result implies that the absolute magnitude of Y_i^* and ATE* is not determined by the data. Any attempt to interpret the size of estimated coefficients or to compare them across models and samples requires a formal normalization scheme. In this section, we define our primary metric, the **Normalized Latent Treatment Effect (NLTE)**, and discuss two alternative strategies: threshold anchoring and predicted probabilities.

2.3.1. The Normalized Latent Treatment Effect (NLTE)

In this paper, we adopt an *error-variance normalization*. Specifically, we fix the variance of the latent error term to one,

$$\text{Var}(\varepsilon_i) = 1.$$

Under this convention, the latent outcome Y_i^* is measured in units where the residual (unexplained) variation has standard deviation one, and each component of τ can be interpreted as the change in Y_i^* (in these units) associated with a one-unit change in the corresponding treatment, holding other covariates fixed. This can be formally defined as below:

$$\tau^{NLTE} = \frac{\mathbb{E}\left[Y^{*(1)} - Y^{*}(0)\right]}{\sqrt{Var(\varepsilon)}} \tag{1}$$

This normalization aligns naturally with existing practice in ordered probit models, which impose

$\varepsilon_i \sim \mathcal{N}(0,1)$ by construction. For ordered logit models, where ε_i is assumed to follow a standard logistic distribution with variance $\pi^2/3$, we rescale all slope coefficients by the factor $\sqrt{3}/\pi$ to place them on the same unit-error-variance scale. That is, if $\hat{\tau}^{\text{ologit}}$ denotes the treatment coefficient from an ordered logit model, we report

$$\hat{\tau}^{\text{NLTE}} = \hat{\tau}^{\text{ologit}} \times \frac{\sqrt{3}}{\pi},$$

and analogously for other slope coefficients. For our density-estimation-based models, which we will introduce in the next section, we estimate the error distribution $\hat{F}$, compute its implied variance

$$\hat{\sigma}_\varepsilon^2 = \text{Var}_{\hat{F}}(\varepsilon_i),$$

and then divide all coefficients by $\hat{\sigma}\varepsilon$:

$$\hat{\tau}^{\text{NLTE}} = \frac{\hat{\tau}}{\hat{\sigma}_\varepsilon},$$

This convention has two advantages. First, it enables *direct comparison across models within the same sample*. Ordered probit, ordered logit (after rescaling), other ordinal regression models all produce treatment coefficients on the same latent scale once normalized. They measure how many units of Y_i^* (where one unit corresponds to one standard deviation of the latent error) the treatment shifts the latent outcome. Differences in estimates can then be attributed to modeling choices (e.g. parametric vs. semiparametric error distributions) rather than arbitrary scaling. Second, it facilitates *comparison across samples*, to the extent that different studies measure the same underlying construct with comparable instruments. If every model is normalized so that $\text{Var}(\varepsilon_i) = 1$, then treatment coefficients from distinct datasets can be interpreted as shifts in a common latent scale on which residual variation has unit variance. Substantive differences in coefficient magnitudes across studies then reflect differences in the relationship between treatment and the latent outcome, rather than differences in ad hoc scale choices imposed by the analyst.

2.3.2. Alternative Strategies: ALTE and Probabilities

While the NLTE is a convenient metric for comparing latent coefficients, we also consider two alternative strategies for identification and substantive interpretation.

First, we consider the **Anchored Latent Treatment Effect (ALTE)**. In case there are many treatments that can be identified, rather than fixing the variance of the unobserved error, the ALTE identifies the model by fixing the value of the effect of a treatment. Specifically, we can the value of the one coefficient as 1. Under this scheme, the error variance σ_ε is a free parameter, and all other treatment coefficients are interpreted relative to the effect of the treatment serves as an anchor.

Second, we utilize **Predicted Probabilities**. To move beyond latent units entirely, we can interpret the results by calculating the change in the predicted probability of a respondent falling into a specific category k, or a set of categories, given a treatment D:

$$\Delta P = P(Y_i \geq k \mid D_i = 1, X_i) - P(Y_i \geq k \mid D_i = 0, X_i)$$

Predicted probabilities are scale-invariant and provide the most stark evidence of model divergence. Because semiparametric models (KDE and NF) can identify non-linearities and non-normalities in the error distribution, they may yield significantly different probability shifts than parametric models even if the NLTE coefficients appear similar.

To clarify again, error-variance normalization does not solve the fundamental up-to-scale identification problem, and this is just one way to normalize the results. However, this provides a transparent and practically convenient metric. In the simulation and application sections that follow, we report NLTEs as our primary results, supplemented by predicted probabilities to highlight the substantive impact of relaxing parametric assumptions.

2.4. Heterogeneous Reporting Functions

The common reporting function assumption posits that all respondents use the same thresholds (α_j) to map their latent attitudes into ordinal responses. In practice, individuals may interpret response categories differently, leading to heterogeneity in the reporting function g. A simple way to incorporate such heterogeneity is to allow thresholds to depend on observed covariates X_i:

$$Y_i = j \quad \Longleftrightarrow \quad \gamma_{j-1}^{\top} X_i < Y_i^* \leq \gamma_j^{\top} X_i, \qquad j = 0, 1, \ldots, J,$$

where γ_j are parameter vectors capturing how X_i shifts the cutpoints. This specification is closely related to generalized ordered logit/probit and partial proportional odds models (Williams, 2006).

In our context, we assume that treatment indicators enter the latent outcome but not the threshold equations, so that heterogeneity in reporting is driven by background covariates rather than by the treatment itself. Under this restriction, the treatment effect τ still remains to be identified up-to-scale. Allowing for heterogeneous reporting functions can improve model fit and efficiency, and is important in applications concerned with interpersonal comparability of responses, but it does not by itself restore identification of the absolute scale of Y_i^*. We return to this specification later in the later sections.

3. Estimation

The previous section showed that, with ordinal outcomes, treatment effects on the latent outcome Y_i^* are identified only up to scale. Any estimator must therefore make an explicit choice of normalization, and any comparison across estimators or samples must respect that choice. This section develops and compares several estimators for the latent treatment effects τ.

We begin by discussing standard parametric ordinal regression models, which are widely used but rely on strong assumptions about the distribution of the latent error term. We then introduce two alternative estimators that relax these assumptions by estimating the error distribution from the data: a semiparametric kernel density estimation (KDE)–based estimator and a normalizing-flow-based estimator.

3.1. Parametric Baseline Models and Their Limitations

3.1.1. Ordered Logit, Ordered Probit, and IRT Models

The most common models for ordinal outcomes in political science are ordered logit and ordered probit. Both are based on the latent outcome representation, introduced in Section 2.1. These two models rely on strong parametric assumptions on the CDF of the latent error term F to estimate. Ordered probit assumes that the error term follows a standard normal distribution, while ordered logit assumes a standard logistic distribution.

Under either specification, the probability of observing category j conditional on (X_i, D_i) is

$$\mathbb{P}(Y_i = j \mid X_i, D_i) = F\big(\alpha_j - f(X_i, \beta) - D_i^\top \tau\big) - F\big(\alpha_{j-1} - f(X_i, \beta) - D_i^\top \tau\big)$$

, where F is the assumed CDF (normal for ordered probit, logistic for ordered logit). The log-likelihood for a sample of size n is

$$\ell(\alpha, \beta, \tau) = \sum_{i=1}^{n} \sum_{j=0}^{J} 1_{Y_i = j} \log \Big[F\big(\alpha_j - f(X_i, \beta) - D_i^\top \tau\big) F\big(\alpha_{j-1} - f(X_i, \beta) - D_i^\top \tau\big) \Big],$$

and maximum likelihood estimation yields $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\tau}$.

If multiple items measure the same latent construct, item response theory (IRT) models provide an alternative framework. Standard IRT models treat each respondent as having a latent trait θ_i and each item k as having discrimination and difficulty parameters. Ordered responses are modeled using item-specific thresholds and an assumed error distribution, typically normal or logistic. Treatment effects can then be estimated either in a second stage, by regressing estimated θ_i on D_i and X_i, or jointly in hierarchical IRT

models that combine measurement and causal components (Zhou, 2019; Stoetzer, Zhou and Steenbergen, 2025).

3.1.2. Consequences of Distributional Misspecification

When the true conditional error distribution F matches the assumed form and the latent outcome is correctly specified, MLE guarantees all parametric models discussed above to yield consistent and asymptotically efficient estimators of (α, β, τ) (up to scale). However, if the true F differs from the assumed distribution, the estimators generally converge to pseudo-true values that solve the misspecified likelihood equations rather than the true parameters (Manski, 1988; Greene and Hensher, 2010). The resulting bias can be substantial and may even change the sign of estimated effects in finite samples.

In applied settings, there are many reasons to doubt the normal or logistic error assumptions. The latent error term may be skewed, heavy-tailed, or multimodal. Unobserved heterogeneity, including unobserved confounders, may enter the latent outcome and induce heteroskedastic or non-normal residual variation. When the error distribution is misspecified, ordered logit/probit and many IRT models are no longer reliable for causal inference on the latent scale.

These concerns motivate estimators that treat the latent outcome structure as parametric but estimate the error distribution F flexibly from the data, avoiding strong distributional assumptions while retaining the interpretability of latent treatment effects.

3.2. A Semiparametric KDE-Based Estimator

Instead of assuming F, we can estimate F semiparametrically using kernel density estimation (KDE), following and extending ideas in Klein and Sherman (2002). The key idea is to construct a quasi-likelihood for (α, β, τ) by plugging a KDE-based estimate $\hat{F}$ into the ordinal likelihood expression and maximizing with respect to the parameters.

This approach treats $f(X_i, \beta) + D_i^\top \tau$ as a low-dimensional parametric component and F as an infinite-dimensional nuisance parameter estimated nonparametrically. Under appropriate smoothness and regularity conditions, the resulting quasi-maximum likelihood estimator is consistent for (α, β, τ) up to scale.

Formally, the KDE-based quasi-log-likelihood takes the form

$$\hat{\ell}(\alpha, \beta, \tau) = \sum_{i=1}^{n} \sum_{j=0}^{J} 1_{Y_i = j} \log \left[\hat{F}\big(\alpha_j - f(X_i, \beta) + D_i^\top \tau\big) - \hat{F}\big(\alpha_{j-1} - f(X_i, \beta) + D_i^\top \tau\big) \right],$$

possibly multiplied by a trimming function that down-weights extreme observations where KDE may be unstable. The KDE-based estimator $(\hat{\alpha}, \hat{\beta}, \hat{\tau})$ is defined as the maximizer of $\hat{\ell}(\alpha, \beta, \tau)$.

The construction of $\hat{F}$ involves standard choices of kernel, bandwidth, and (if desired) local smoothing and trimming parameters. Following Klein and Sherman (2002), we require that the conditional density of the index V_i be smooth and strictly positive over its support, and that bandwidths shrink at appropriate rates. The full procedure and regularity conditions are provided in the Appendix.

Under mild regularity conditions, the KDE estimator $\hat{F}$ converges uniformly to the true F, and the resulting quasi-log-likelihood $\hat{\ell}(\alpha, \beta, \tau)$ converges uniformly to the infeasible log-likelihood $\ell(\alpha, \beta, \tau; F)$. Standard arguments for M-estimators then imply that the maximizer $(\hat{\alpha}, \hat{\beta}, \hat{\tau})$ is consistent for (α, β, τ) up to a positive scale factor. Under additional smoothness and moment conditions, the estimator is asymptotically normal with a variance that reflects both sampling noise and the first-stage nonparametric estimation error.

Intuitively, by letting the data determine the shape of F rather than imposing a normal or logistic form, the KDE-based estimator avoids the bias that arises when ordered logit or probit are misspecified. In large samples, it recovers the same latent treatment effects (up to scale) that would be obtained if F were known.

3.3. A Normalizing-Flow-Based Estimator

3.3.1. Normalizing Flows: Basic Idea

Normalizing flows are a flexible class of models for probability distributions that represent a complex density as the image of a simple base density under a sequence of invertible, differentiable transformations (Chen and Gopinath, 2000). Let Z be a random variable with a simple base distribution (such as standard normal), and let T_θ be an invertible transformation parameterized by θ. Define

$$h = T_\theta(Z).$$

The density of h is then given by the change-of-variables formula:

$$f_\theta(h) = f_Z\big(T_\theta^{-1}(h)\big) \left|\det \mathrm{d}\, T_\theta^{-1}(h)\right|,$$

where f_Z is the base density and $\mathrm{d}\, T_\theta^{-1}$ is the Jacobian of the inverse transformation. The corresponding CDF F_θ can be obtained by integrating f_θ. By choosing a sufficiently rich family for T_θ, normalizing flows can approximate a wide range of error distributions while preserving tractable likelihoods and gradients.

Again, instead of assuming a specific F, we model the error term as

$$\varepsilon_i = T_\theta(Z_i), \qquad Z_i \sim \mathcal{N}(0, 1),$$

where T_θ is a normalizing flow. The induced error distribution has density f_θ and CDF F_θ.

Given parameters $(\alpha, \beta, \tau, \theta)$, the probability of observing category j is

$$\mathbb{P}(Y_i = j \mid X_i, D_i; \alpha, \beta, \tau, \theta) = F_\theta\big(\alpha_j - f(X_i, \beta) + D_i^\top \tau\big) - F_\theta\big(\alpha_{j-1} - f(X_i, \beta) + D_i^\top \tau\big).$$

The log-likelihood for the sample is defined analogously

$$\ell(\alpha, \beta, \tau, \theta) = \sum_{i=1}^{n} \sum_{j=0}^{J} 1_{Y_i = j} \log \Big[F_\theta\big(\alpha_j - f(X_i, \beta) + D_i^\top \tau\big) - F_\theta\big(\alpha_{j-1} - f(X_i, \beta) + D_i^\top \tau\big) \Big].$$

We estimate $(\alpha, \beta, \tau, \theta)$ jointly by maximizing this likelihood using gradient-based optimization. Because T_θ is invertible and differentiable, both f_θ and F_θ are tractable, and the gradients of the log-likelihood with respect to all parameters can be computed efficiently.

In practice, one must choose a specific flow architecture and hyperparameters. For example, since the error term is usually one-dimensional, we may use a rational quadratic spline flow, which represents T_θ as a piecewise rational quadratic function with monotone constraints, or a coupling-based flow with affine or spline transformations. The base distribution Z is typically standard normal. The parameters θ are trained jointly with (α, β, τ) by maximizing the ordinal log-likelihood, using stochastic gradient descent or related algorithms. In the simulations and application, we use relatively simple flows that balance flexibility with computational tractability.

3.4. Practical Considerations and Comparison

Ordered logit and ordered probit are simple, fast, and familiar. When their distributional assumptions are believable, they yield efficient and unbiased estimates. However, when the true error distribution deviates from logistic or normal, or when unobserved heterogeneity induces non-normal residual variation, these models can be seriously biased (Manski, 1988; Greene and Hensher, 2010; Johnston, McDonald and Quist, 2020; Smits et al., 2020).

The KDE-based estimator retains a parametric structure for the latent outcome but estimates the error distribution nonparametrically. It avoids imposing a specific functional form for F and is consistent under mild smoothness and support conditions, at the cost of additional tuning (choice of kernel and bandwidth) and potentially higher variance in small samples. It is most attractive in moderate to large samples where the KDE of the error distribution can be estimated reliably.

The normalizing-flow-based estimator also relaxes the parametric error assumption, but does so within a fully parametric likelihood framework. By representing the error distribution via an invertible transformation

of a simple base distribution, it can flexibly approximate complex shapes while preserving tractable likelihood evaluation and gradient computation. This approach can deliver both robustness to misspecification and efficiency, especially in settings with sufficient sample size and computational resources.

In all cases, error-variance normalization is essential for comparability. By expressing all coefficients in units of the latent error's standard deviation, we ensure that treatment effects from ordered probit, ordered logit (after rescaling), the KDE-based estimator, and the normalizing-flow-based estimator are on the same latent scale. The next section uses Monte Carlo simulations to compare these estimators under a variety of latent error distributions and sample sizes.

4. Monte Carlo Simulation

Several Monte Carlo simulations have been conducted to test the surrogate-based residuals and compare the performance of the Klein and Sherman semiparametric estimator and NF-based estimator compared with standard ordered probit and logit models. Here I consider the case where the outcome is a ordinal variable with 5 levels.

The latent variable Y^* follows the relationship:

$$Y^* = D^T\tau + X^T\beta + \varepsilon$$

, where D is a binary variable denoting randomized treatment status, X is a matrix of covariate values. X has one binary variable and two continuous variable generated from the standard normal distributions.

We observe Y that is constructed as:

$$Y = \begin{cases} 1 & \text{if} \quad \gamma_{-1} \leq Y^* \leq \gamma_0 \\ 2 & \text{if} \quad \gamma_0 \leq Y^* \leq \gamma_1 \\ 3 & \text{if} \quad \gamma_1 \leq Y^* \leq \gamma_2 \\ 4 & \text{if} \quad \gamma_2 \leq Y^* \leq \gamma_3 \\ 5 & \text{if} \quad \gamma_3 \leq Y^* \leq \gamma_4 \end{cases}$$

, where $\gamma = -1 = -\infty$, and $\gamma 4 = \infty$. Thus, the thresholds are dependent on covariates X.

The performances of the KDE-based estimator and the NF-based estimator introduced in the previous section were compared against ordered logit, and probit models, and OLS with common cardinalization of 1 to 5. I tested for two different distribution for the error term, ε, the standard normal distribution $\mathbb{N}(0, 1)$ and

the lognormal-distribution with the scale factor of 1. The lognormal-distribution has heavier tail regions than both the standard logistic and the standard normal distribution. Thus, by construction, the normal error represents the most favorable DGP for the ordered probit model, and the lognormal-distribution error term case represents DGP where the distributional assumptions for both ordered logit and ordered probit models are violated. Four sample sizes of 500, 1,000 and 5,000 are considered. I first generate a population of size $1e+6$ and draw random samples for Monte Carlo replication from this population. The number of Monte Carlo replication is 100, and all estimations are the means of these 100 replication results.

For the NF-based model, I choose to use rational quadratic spline based flow which has closed form invertible structure and ensures fast optimization. Hyperparameters for the flow are the number of bins and bounds, and the number of bins is set to 32 and bounds are set to $[-10, 10]$. These two parameters work similar to the bandwidth for KDE, thus smaller bounds or greater number of bins increases flexibility. The optimization of the log-likelihood for the Normalizing flow based estimator is done by BFGS optimizer accompanied by 1,000 epochs of stochastic optimation using Adam optimizer.

Regarding the KDE-based estimator, I used standard normal kernel and set $\delta = \frac{1}{15}$, pilot bandwidth to be $2/3$th point of the suggested interval, the final base bandwidth to be the $2/5$ point of the suggested interval, and fix all other parameters to the middle points of the suggested intervals. For instance, ϵ in the damping function is set to be $\frac{0+(\frac{1}{40}-\frac{\delta}{20})}{2} \approx 0.33$. Also, the trimming was done at the 0.02 level, so data points within 0.02 quantile and 0.98.

Since the ordinal regression estimators only identify τ up to scale, without lose of generality, I normalize all the coefficients including τ with the variance of the error term. For the flow model, the variance is calculated by Monte Carlo sampling from the learned flow for the error term. For the KDE-based model, the error term is estimated by another round of KDE again and Monte Carlo sampling from the KDE estimate of the error term is used to estimate the variance. This allows easy comparison across different models and the interpretation for the coefficients will be relative to the variance of the error term. τ^{NLTE} is set to be 1.0 for the standard normal error case and 0.46 for the lognormal error case. The exact value of $\tau = 1$ is used to calculate bias and RMSE in OLS case, since it is argued to estimate the ATE directly.

Figure 1 summarizes the Monte Carlo simulation results for the standard normal error term case. Panel (a) shows the mean bias from 100 simulations of OLS, ordered logit model, ordered probit model, KDE-based estimator, and Normalizing flow based estimator. The biases of the KDE-based and NF-based estimators are comparable to the true model, the ordered probit. Panel (b) shows RMSE for each estimator, and again both KDE-based estimator and NF-based estimator perform comparable to ordered probit model. Considering that the standard normal error case meets the distributional assumption for ordered probit model, the great performance of the ordered probit model is expected. Also, the CDF of standard logistic is not too different

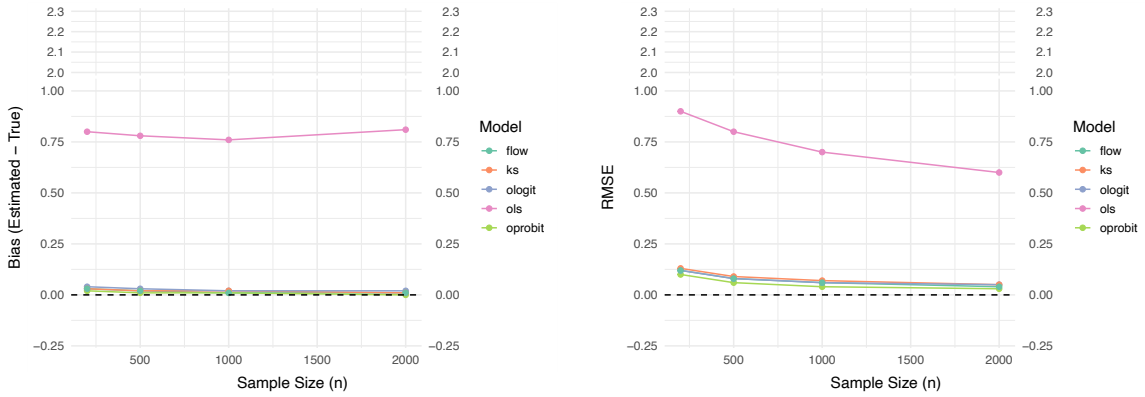

FIGURE 1. Monte Carlo Simulation Results: Standard Normal Error

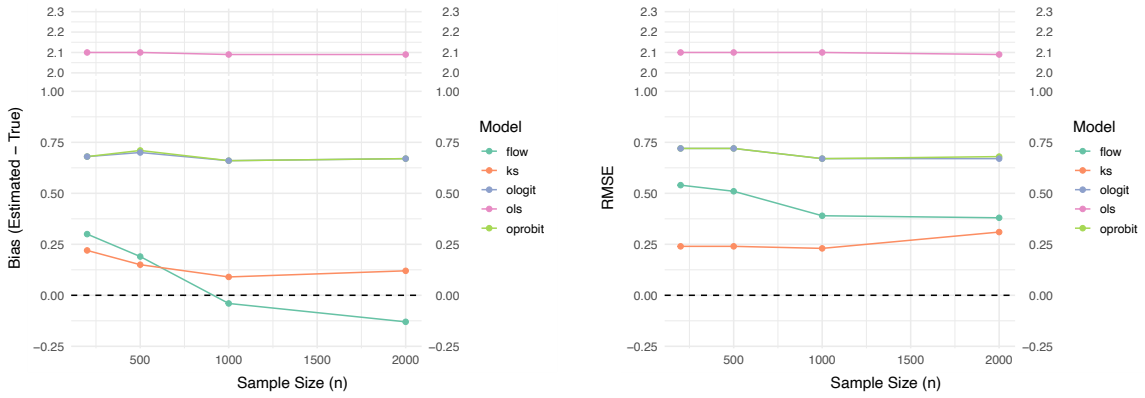

FIGURE 2. Monte Carlo Simulation Results: Lognormal distribution Error

from that of standard normal, ordered logit model also retains small bias. This demonstrates strong finite sample performance of both KDE-based and NF-based estimators.

Figure 2 describes the results from the lognormal-distribution error term case. Again OLS shows the worst results, clearly overestimating the treatment effect. Ordered probit and ordered logit resulted in very similar estimates and almost indistinguishable in the figure. However, these standard ordinal regression models also show large over estimation of NLTE, and the bias dies not shrink as the sample size increases. This clearly demonstrates that they can be inconsistent estimators if the distributional assumptions are violated. Both of the flexible ordinal estimators introduced in this paper shows much better performance in terms of bias, and as expected, the biases seem to converge to 0 as the sample size increases. One panel (b) we observe a similar story, while OLS has very RMSE, standard ordered logit and probit maintains the RMSE around 0.7 to 0.74. Both KDE-based and NF-based estimators achieves the best performance in terms of RMSE.

Overall, the Monte Carlo simulation demonstrates that cardinalization approach depends on arbitrary

labels and thus unreliable to capture the causal effect, standard ordered regression models may perform poorly when the distributional assumptions are violated. KDE-based estimator and NF-base estimator, however, demonstrate better performance across two cases and different sample sizes.

5. Application 1: Tomz and Weeks (2020)

The first example replicates the analysis by Tomz and Weeks (2020). The study looked into the question of how the human rights practices of foreign adversaries affect domestic public support for war through survey experiments. The main argument was that in democracies with well-established human rights, people are less willing to go to war with other countries that also respect human rights than with comparable countries that violate them. Thus they are focusing on two factors that can impact people's support for war: the political system of the foreign country and how much the foreign country respect human rights.

The paper uses total of 5 surveys, but this paper focus on the first, which contains an experiment for their main result. The survey was conducted by YouGov with a nationally representative sample of 1,430 US adults in October 2012. The treatment was given as in a form of vignettes on a country that is developing nuclear weapons. The vignettes also include trade and military relationship of the country and the US, and its conventional military strength which is set to the half of that of the US. In the vignettes, information about the political system (democracy = 1, not = 0) and human rights practices (respect = 1, not = 0) were binary treatments and they were randomized independently.

The outcome of interest here is the support for physical strike of the US armed forces to the country. Specifically, respondents were asked to answer in 5-point scale: *Favor strongly, Favor somewhat, Neither favor nor oppose, Oppose somewhat*, and *Oppose strongly*. In the paper, the authors take the cardinalization approach and OLS. They first recode the outcome to take numeric values from 1 to 5, then they used this to run OLS. As discussed in Section 2, the cardinalization is arbitrary and we only can interpret the ATE as suggested by authors when the cardinalization is capturing the unknown transformation function, which is untestable.

The model they used in their main results (model 2 of their Table 1) can be expressed as following:

$$Y^* = \tau_{DEM} \cdot DEM + \tau_{HR} \cdot HR + X^T \beta + \varepsilon$$

$$Y \in \{1, 2, 3, 4, 5\}$$

where Y^* is the support for the military strike in latent scale and Y is the observed outcome with cardinalization from 1 to 5, DEM and HR are binary variables indicating political system treatment and

human rights practices treatment respectively, X for the control variables and ε for the error term.

Since they are interested in whether citizens living in democracy where human rights are respected (the US, to be more specific) are more reluctant to support military strike to a country who is developing nuclear weapons if the country also respecting human rights or the country has democratic political system, the causal parameters we are interested in the above equation are τ_{HR} and τ_{DEM}. We only can observe the ordinal outcome, we can only identify the effects up to scale, and here I adopted error-variance normalization we discussed in the previous section. I replicate their result on Table S3 of their paper, and use ordered logit, ordered probit and KDE-based estimator as a comparison.

	ATE	NLTE			
	Original – OLS	Ordered Logit	Ordered Probit	KDE-based	NF-based
Respect HR	-12.6^{***}	-0.46^{***}	-0.49^{***}	-0.48	-0.49^{***}
	(1.47)	(0.10)	(0.06)	(0.27)	(0.06)
Democracy	-9.5^{***}	-0.31^{**}	-0.34^{***}	-0.33	-0.34^{***}
	(1.47)	(0.10)	(0.06)	(0.22)	(0.06)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 4. Replication of Tomz and Weeks (2020) Table S3. ATE represents the original binarized OLS estimates (0-100 scale). NLTE columns report standardized latent coefficients ($\sigma = 1$).

Table 4 reports the results across all five specifications. Substantively, all models agree on the direction of the treatment effects: information that a country respects human rights or is a democracy significantly decreases domestic support for military action. However, the interpretation and statistical information utilized differ. The original OLS interpretation (a 12.6 percentage point drop for the Human Rights treatment) relies on the validity of the binarization and the cardinal 0–100 scale. By collapsing the 5-point scale into two categories, the original analysis discards potentially valuable information regarding the intensity of respondent preferences.

In contrast, NLTE models utilize the full ordinal range of the data. The standardized coefficients remain remarkably consistent across the Ordered Probit, KDE-based, and NF-based models. For instance, the effect of respecting human rights is estimated at approximately -0.49 standard deviations across these specifications. The fact that the flexible semiparametric estimators (KDE and NF) closely track the Ordered Probit results serves as a diagnostic, suggesting that the latent error distribution for this specific survey approximates a normal distribution. In this instance, the parametric assumption appears robust, though the semiparametric approach provides the necessary validation that the original cardinalization was not distorting the underlying direction of the effects.

6. Application 2: Mattingly et al. (2025)

The second application replicates the large-scale cross-national study by Mattingly et al. (2025), which examines how authoritarian regimes utilize state-produced media to promote their political and economic models to foreign audiences. We focus our replication on the authors' primary hypotheses: (H1) US government messaging increases preference for the American model, (H2) Chinese Communist Party (CCP) messaging increases preference for the Chinese model, and (H3) in the presence of competing messages, public preference shifts toward the Chinese model.

To test the hypothesis, authors launched multiple survey experiments took place in 19 different countries (N = 6,000) in June 2022. Respondents were assigned via block randomization to one of four conditions: viewing US state-produced videos (USA), CCP-produced videos (China), a combination of both (Competition), or nature-themed placebo videos (Control). The primary outcomes were measured using a 6-point Likert scale, ranging from *Strongly Prefer the US* (1) to *Strongly Prefer China* (6). The original analysis relied on OLS regression, which cardinalizes these ordinal categories by treating the 1–6 scale as an interval measure.

After the treatment, respondents were asked to answer two questions which measures main outcome variables (whether they prefer the US (political / economic) system or (political / economic) Chinese system) in 6 categories from *Strongly Prefer the US* to *Strongly Prefer China*. The cardinalize the outcomes by assigning numeric values, so that *Strongly Prefer the US* was recoded as 1 and *Strongly Prefer China* was recoded as 6. The original results came from OLS results with these cardinalized outcomes.

The model Mattingly et al. (2025) used for their main results can be expressed as follows

$$Y_i^* = \tau_{CHINA} \cdot CHINA_i + \tau_{USA} \cdot USA_i + \tau_{COMP} \cdot COMP_i + X_i^T \gamma + \varepsilon_i$$

$$Y \in \{1, 2, 3, 4, 5, 6\}$$

where $CHINA_i$, USA_i and $COMP_i$ are indicator variables for the respective experimental conditions, and X_i is a vector of pre-treatment covariates including age, gender, education and family income.

While the original study estimates the Average Treatment Effect (ATE) by applying OLS directly to the categorical labels (1–6), our approach focuses on identifying the latent treatment parameters (τ). As is standard in semiparametric ordinal models, the latent scale is identified only up to location and scale; hence, we apply the error-variance normalization ($\sigma = 1$). This allows for a direct comparison between the parametric specifications (Ordered Probit and Logit) and the flexible KDE and NF-based estimators. Under this framework, the coefficients represent the shift in the latent preference distribution in units of standard

deviations, providing a measure of treatment intensity that is robust to the arbitrary cardinalization of the 6-point scale.

Table 5 reports the replication of the original OLS results, and results of NLTE estimation using different ordinal regression models including KDE-based model and flow-based model. For preferences over the political system (upper panel), we find a stark divergence between parametric and semiparametric models. While the direction of the effects remains consistent across all specifications, the magnitude of the estimates is significantly inflated in the cardinal (OLS) and parametric ordinal (Logit/Probit) models. Specifically, the estimated effect of the Chinese treatment in the NF-based model (0.37) is nearly two-thirds smaller than the OLS estimate (1.04) and less than half the size of the Ordered Probit estimate (0.76).

Notably, the *Competition* effect, which the original study identified as a robust driver of preference for the Chinese model (0.36, $p < 0.001$), is substantially attenuated in the flexible models. In the NF-based specification, this effect drops to 0.11, suggesting that when the assumption of a Normal error distribution is relaxed, the persuasive power of competitive rhetoric appears much more fragile than previously reported. This suggests that the OLS and Probit models may be over-attributing shifts in the latent distribution to the treatment, rather than accounting for the non-normal dispersion of political preferences.

In contrast, the lower panel of Table 5 reports preferences over economic systems, where the models show greater consensus. While OLS still yields the largest point estimates, the gap between the Probit (0.57) and the NF-based model (0.58) for the China treatment is negligible. This indicates that the latent distribution of economic preferences in this sample likely approximates normality more closely than political preferences. However, the OLS estimates for all treatments remain consistently higher than their latent counterparts, reinforcing the argument that cardinalizing Likert scales may lead to substantially different conclusions.

7. Conclusion

This paper has examined the problem of causal inference when the outcome of interest is measured on an ordinal scale, a situation that arises frequently in political science research on attitudes and preferences. Building on a standard latent-variable framework, we showed that, even under conventional causal assumptions such as SUTVA, ignorability, and positivity, treatment effects on the underlying continuous outcome are in general identified only up to scale. The joint distribution of ordinal responses is invariant to positive affine transformations of the latent outcome and thresholds, so the absolute magnitude and location of the latent average treatment effect cannot be recovered from the data alone. Any attempt to compare coefficients (across models or across studies) must therefore adopt an explicit normalization. To address this, we proposed the *Normalized Latent Treatment Effect (NLTE)* as an alternative causal estimand, ensuring that latent treatment

Outcome: Preference over Political System					
	ATE	NLTE			
	Original – OLS	Ordered Logit	Ordered Probit	KDE-based	NF-based
China	1.04^{***}	0.73^{***}	0.76^{***}	0.36^{***}	0.37^{***}
	(0.05)	(0.04)	(0.04)	(0.07)	(0.04)
USA	-0.43^{***}	-0.35^{***}	-0.38^{***}	-0.18^{***}	-0.17^{***}
	(0.04)	(0.04)	(0.04)	(0.05)	(0.03)
Competition	0.36^{***}	0.21^{***}	0.25^{***}	0.13^{*}	0.11^{**}
	(0.05)	(0.04)	(0.04)	(0.06)	(0.04)
Outcome: Preference over Economic System					
	ATE	NLTE			
	Original – OLS	Ordered Logit	Ordered Probit	KDE-based	NF-based
China	0.87^{***}	0.56^{***}	0.57^{***}	0.59^{***}	0.58^{***}
	(0.05)	(0.04)	(0.04)	(0.07)	(0.06)
USA	-0.57^{***}	-0.39^{***}	-0.43^{***}	-0.42^{***}	-0.43^{***}
	(0.05)	(0.04)	(0.04)	(0.05)	(0.05)
Competition	0.28^{***}	0.16^{***}	0.18^{***}	0.19^{**}	0.18^{**}
	(0.06)	(0.04)	(0.04)	(0.06)	(0.06)

$^{***}p < 0.001$; $^{**}p < 0.01$; $^{*}p < 0.05$

TABLE 5. Replication of Mattingly et al. (2025) Table C3. ATE represents original OLS estimates. NLTE columns report standardized latent coefficients from various ordinal specifications.

effects are expressed in a transparent and comparable metric.

We argued that common strategies for analyzing ordinal outcomes are problematic in light of this limitation. Cardinalization, which assigns numeric scores to categories and applies tools such as OLS or difference-in-means, produces estimates whose size and sometimes even sign can depend sensitively on arbitrary labeling choices (Schröder and Yitzhaki, 2017; Bond and Lang, 2018; Bloem, 2022). Binarization discards information about intermediate categories, leading to inefficient estimators and potentially obscuring substantively important changes in the distribution of responses. Parametric ordinal models, including ordered logit and ordered probit, avoid arbitrary numeric labels and respect the ordered nature of the data, but rely on strong assumptions about the latent error distribution. When these assumptions are violated the resulting estimators converge to pseudo-true parameters and can suffer substantial bias (Manski, 1988; Greene and Hensher, 2010; Johnston, McDonald and Quist, 2020; Smits et al., 2020).

This paper introcused two estimators that retain the interpretability of latent treatment effects while

relaxing parametric assumptions on the error distribution. The first is a semiparametric KDE-based estimator that uses kernel density estimation to approximate the latent error CDF and constructs a quasi-likelihood for the ordinal model. The second employs normalizing flows to model the error distribution as an invertible transformation of a simple base distribution within a maximum likelihood framework. Both estimators treat the latent outcome as parametric, estimate the error distribution flexibly from the data.

Monte Carlo simulations demonstrated that these choices matter in practice. When the latent error distribution is correctly specified (for example, normal for ordered probit), parametric ordinal models perform well, as expected. However, under misspecification such as lognormal error, ordered logit and probit can be noticeably biased, whereas the KDE-based and NF-based estimators remain consistent and exhibit favorable finite-sample performance. The simulations also showed that cardinalization yields highly unstable estimates across different labeling schemes, and that binarization is less efficient and can miss meaningful shifts among categories. In this sense, the density-estimation-based approaches provide a robust alternative that better respects the ordinal nature of the data and the limited information it contains about the latent scale.

The empirical applications further underscore the substantive importance of these methods. Reanalyzing Tomz and Weeks (2020) on human rights and military force, we found that standard binarized OLS tends to understate the magnitude of the latent treatment effect. Conversely, our replication of Mattingly et al. (2025) regarding authoritarian propaganda reveals that the perceived "superiority" of Chinese state cues over American cues in the original study is, in part, a methodological artifact. By accounting for the floor effects and non-normal dispersion of political preferences, our proposed estimators show that the relative persuasive power of these cues is much closer than cardinal OLS suggests. These examples demonstrate how the assumption of cardinality or normality can inadvertently shape—and potentially distort—substantive conclusions in top-tier political science research.

Several directions for future work follow from these results. First, while we focused on single-item ordinal outcomes, extending density-estimation-based methods to multi-item settings and integrating them more tightly with IRT-style measurement models would be valuable. Second, developing practical diagnostic tools for assessing error distribution misspecification in ordinal models, building on surrogate residuals or related techniques, would complement the estimators proposed here. Third, applications with clustered, panel, or hierarchical data structures pose additional challenges for both identification and estimation that merit further study. More broadly, the analysis reinforces the importance of taking the ordinal nature of many political science outcomes seriously and of being explicit about the assumptions and normalizations that underlie estimates of causal effects on latent attitudes.

References

Abramson, Ian S. 1982. "On Bandwidth Variation in Kernel Estimates-A Square Root Law." *The Annals of Statistics* 10(4).

Alt, James and Torben Iversen. 2017. "Inequality, Labor Market Segmentation, and Preferences for Redistribution." *American Journal of Political Science* 61(1):21–36.

Armstrong, Ben G. and Sloan Margaret. 1989. "Ordinal Regression Models for Epidemiologic Data." *American Journal of Epidemiology* 129(1):191–204.

Bloem, Jeffrey R. 2022. "How Much Does the Cardinal Treatment of Ordinal Variables Matter? An Empirical Investigation." *Political Analysis* 30(2):197–213.

Bond, Timothy N. and Kevin Lang. 2018. "The Sad Truth about Happiness Scales." *Journal of Political Economy* 127(4):1629–1640.

Breen, Richard, Kristian Bernt Karlson and Anders Holm. 2018. "Interpreting and Understanding Logits, Probits, and Other Nonlinear Probability Models." *Annual Review of Sociology* 44(Volume 44, 2018):39–54.

Canes-Wrone, Brandice and Scott De Marchi. 2002. "Presidential Approval and Legislative Success." *Journal of Politics* 64(2):491–509.

Chen, Scott and Ramesh Gopinath. 2000. Gaussianization. In *Advances in Neural Information Processing Systems*, ed. T. Leen, T. Dietterich and V. Tresp. Vol. 13 MIT Press.

D'Amico, Gennaro, Juan G. Abraldes, Paola Rebora, Maria Grazia Valsecchi and Guadalupe Garcia-Tsao. 2020. "Ordinal Outcomes Are Superior to Binary Outcomes for Designing and Evaluating Clinical Trials in Compensated Cirrhosis." *Hepatology* 72(3):1029–1042.

Greene, William H. and David A. Hensher. 2010. *Modeling Ordered Choices : A Primer*. Cambridge: Cambridge University Press.

Johnston, Carla, James McDonald and Kramer Quist. 2020. "A generalized ordered Probit model." *Communications in Statistics - Theory and Methods* 49(7):1712–1729.

King, Gary, Christopher J.L. Murray, Joshua A. Salomon and Ajay Tandon. 2004. "Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research." *American Political Science Review* 98:191–207.

King, Gary and Jonathan Wand. 2007. "Comparing Incomparable Survey Responses: Evaluating and Selecting Anchoring Vignettes." *Political Analysis* 15(1):46–66.

Klein, Roger W. and Richard H. Spady. 1993. "An Efficient Semiparametric Estimator for Binary Response Models." *Econometrica* 61(2):387–421.

Klein, Roger W. and Robert P. Sherman. 2002. "Shift Restrictions and Semiparametric Estimation in Ordered Response Models." *Econometrica* 70(2):663–691.

Kriner, Douglas and Liam Schwartz. 2009. "Partisan Dynamics and the Volatility of Presidential Approval." *British Journal of Political Science* 39(3):609–631.

Magni, Gabriele. 2021. "Economic Inequality, Immigrants and Selective Solidarity: From Perceived Lack of Opportunity to In-group Favoritism." *British Journal of Political Science* 51(4):1357–1380.

Manski, Charles F. 1988. "Identification of Binary Response Models." *Journal of the American Statistical Association* 83(403):729–738.

Mattingly, Daniel, Trevor Incerti, Changwook Ju, Colin Moreshead, Seiki Tanaka and Hikaru Yamagishi. 2025. "Chinese state media persuades a global audience that the "China model" is superior: Evidence from a 19-country experiment." *American Journal of Political Science* 69(3):1029–1046.

Mayda, Anna Maria and Dani Rodrik. 2005. "Why are some people (and countries) more protectionist than others?" *European Economic Review* 49(6):1393–1430.

Rubin, Donald B. 1974. "Estimating causal effects of treatments in randomized and nonrandomized studies." *Journal of educational Psychology* 66(5):688.

Scheve, Kenneth F and Matthew J Slaughter. 2001. "What determines individual trade-policy preferences?" *Journal of International Economics* 54(2):267–292.

Schröder, Carsten and Shlomo Yitzhaki. 2017. "Revisiting the evidence for cardinal treatment of ordinal variables." *European Economic Review* 92:337–358.

Silverman, Bernard W. 1986. *Density estimation for statistics and data analysis*. Routledge.

Smits, Niels, Oguzhan Ogreden, Mauricio Garnier-Villarreal, Caroline B Terwee and R Philip Chalmers. 2020. "A study of alternative approaches to non-normal latent trait distributions in item response theory models used for health outcome measurement." *Statistical Methods in Medical Research* 29(4):1030–1048.

Stoetzer, Lukas F., Xiang Zhou and Marco Steenbergen. 2025. "Causal inference with latent outcomes." *American Journal of Political Science* 69(2):624–640.

Tomz, Michael R. and Jessica L. P. Weeks. 2020. "Human Rights and Public Support for War." *The Journal of Politics* 82(1):182–194.

Williams, Richard. 2006. "Generalized Ordered Logit / Partial Proportional Odds Models for Ordinal Dependent Variables." *The Stata Journal* 6(1):58–82.

Zhou, Xiang. 2019. "Hierarchical Item Response Models for Analyzing Public Opinion." *Political Analysis* 27(4):481–502.

8. Appendix

A. KDE-Based Ordinal Regression

This appendix provides additional details for the semiparametric KDE-based estimator introduced in Section 3. We first define the quasi-likelihood, then describe the kernel estimation of the error distribution and state a consistency result.

A.1. Model and Quasi-Likelihood

Recall the latent outcome model

$$Y_i^* = f(X_i, \beta) + D_i^\top \tau + \varepsilon_i,$$

and the threshold-based reporting function

$$Y_i = j \quad \Longleftrightarrow \quad \alpha_{j-1} < Y_i^* \leq \alpha_j, \qquad j = 0, 1, \ldots, J,$$

with $-\infty = \alpha_{-1} < \alpha_0 < \cdots < \alpha_J = \infty$. For notational simplicity, define the latent outcome

$$V_i(\beta, \tau) = f(X_i, \beta) + D_i^\top \tau.$$

Let F denote the (unknown) CDF of ε_i. Conditional on (X_i, D_i), the probability of observing category j is

$$p_j(X_i, D_i; \alpha, \beta, \tau, F) = F\big(\alpha_j - V_i(\beta, \tau)\big) - F\big(\alpha_{j-1} - V_i(\beta, \tau)\big).$$

If F were known, the population log-likelihood would be

$$\ell(\alpha, \beta, \tau; F) = \mathbb{E}\left[\sum_{j=0}^{J} 1_{Y_i = j} \log p_j(X_i, D_i; \alpha, \beta, \tau, F)\right].$$

In the sample of size n, the infeasible sample log-likelihood is

$$\ell_n(\alpha, \beta, \tau; F) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=0}^{J} 1_{Y_i = j} \log p_j(X_i, D_i; \alpha, \beta, \tau, F).$$

When F is unknown, we replace it by a nonparametric estimator $\hat{F}$ constructed via KDE, obtaining a

quasi-log-likelihood

$$\hat{\ell}_n(\alpha, \beta, \tau) = \frac{1}{n} \sum_{i=1}^{n} \hat{m}(X_i) \sum_{j=0}^{J} 1_{Y_i = j} \log \hat{p}_j(X_i, D_i; \alpha, \beta, \tau),$$

where

$$\hat{p}_j(X_i, D_i; \alpha, \beta, \tau) = \hat{F}\big(\alpha_j - V_i(\beta, \tau)\big) - \hat{F}\big(\alpha j - 1 - V_i(\beta, \tau)\big),$$

and $\hat{m}(X_i)$ is a trimming function that downweights observations in extreme regions of the covariate space where KDE may be unstable. The KDE-based estimator $(\hat{\alpha}, \hat{\beta}, \hat{\tau})$ is defined as any maximizer of $\hat{\ell}_n(\alpha, \beta, \tau)$.

A.2. Kernel Estimation of the Error Distribution

To construct $\hat{F}$, we exploit the relationship between the distribution of the index V_i and the error term ε_i. Let

$$g_1(v \mid Y \leq j) = \text{density of } V \text{ given } Y \leq j, \qquad g_0(v \mid Y > j) = \text{density of } V \text{ given } Y > j,$$

and let

$$\pi_j = \mathbb{P}(Y_i \leq j), \qquad 1 - \pi_j = \mathbb{P}(Y_i > j).$$

Then by the Bayes' rule,

$$\mathbb{P}(Y_i \leq j \mid V_i = v) = \frac{\pi_j g_1(v \mid Y \leq j)}{\pi_j g_1(v \mid Y \leq j) + (1 - \pi_j) g_0(v \mid Y > j)}.$$

Since $Y_i^* = V_i + \varepsilon_i$, and $Y_i \leq j$ iff $Y_i^* \leq \alpha_j$, we have

$$\mathbb{P}(Y_i \leq j \mid V_i = v) = \mathbb{P}\big(\varepsilon_i \leq \alpha_j - v\big) = F\big(\alpha_j - v\big).$$

Thus, if we can estimate π_j, $g_1(\cdot)$, and $g_0(\cdot)$ from the data, we can construct an estimator $\hat{F}$ such that

$$\hat{F}(\alpha_j - v) \approx \widehat{\mathbb{P}}(Y_i \leq j \mid V_i = v).$$

The probabilities π_j and $1 - \pi_j$ can be estimated by sample proportions:

$$\hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} 1_{Y_i \leq j}, \qquad 1 - \hat{\pi}_j = \frac{1}{n} \sum_{i=1}^{n} 1_{Y_i > j}.$$

The conditional densities g_1 and g_0 are estimated using KDE. For concreteness, suppose we use a univariate kernel $K(\cdot)$ (e.g. Gaussian) and bandwidths that may depend on j. Let $n_1(j) = \sum_{i=1}^{n} 1_{Y_i \leq j}$ and $n_0(j) =$

$\sum_{i=1}^{n} 1_{Y_i > j}$. Define the subsamples

$$V_i : Y_i \leq j, \qquad V_i : Y_i > j,$$

and estimate the conditional densities as

$$\hat{g}_1(v \mid Y \leq j) = \frac{1}{n_1(j) h_{1j}} \sum_{i: Y_i \leq j} K\left(\frac{v - V_i}{h_{1j}}\right),$$

$$\hat{g}_0(v \mid Y > j) = \frac{1}{n_0(j) h_{0j}} \sum_{i: Y_i > j} K\left(\frac{v - V_i}{h_{0j}}\right),$$

where h_{1j} and h_{0j} are bandwidths chosen according to standard rules (possibly location-adaptive). Then the estimated conditional probability is

$$\widehat{\mathbb{P}}(Y_i \leq j \mid V_i = v) = \frac{\hat{\pi}_j \hat{g}_1(v \mid Y \leq j)}{\hat{\pi}_j \hat{g}_1(v \mid Y \leq j) + (1 - \hat{\pi}_j) \hat{g}_0(v \mid Y > j)}.$$

We interpret this as an estimator of the CDF of ε_i evaluated at $\alpha_j - v$:

$$\hat{F}(\alpha_j - v) \equiv \widehat{\mathbb{P}}(Y_i \leq j \mid V_i = v).$$

In practice, the KDE is implemented with local smoothing and trimming to reduce boundary bias and instability in regions with sparse data. Following Abramson (1982) and Silverman (1986), one can use pilot density estimates to define location-specific bandwidths and damping functions that prevent the bandwidth from shrinking too quickly in low-density regions. We adopt these techniques, as in Klein and Sherman (2002), to ensure $\sqrt{n}$ consistency.

A.3. Consistency of the KDE-Based Estimator

This section sketches the main consistency result for the KDE-based estimator. Fuller proofs follow from standard arguments combined with the consistency of locally smoothed KDEs (Klein and Spady, 1993; Abramson, 1982; Silverman, 1986).

Assumption 5 (Smoothness of the Index Distribution). *The conditional density of the index $V_i(\beta, \tau)$ given (X_i, D_i) exists, is strictly positive on its support, and is continuously differentiable up to a sufficiently high order. The support of V_i is compact or can be truncated to a compact set via trimming without affecting the asymptotic behavior of the estimator.*

Assumption 6 (KDE Regularity Conditions). *The kernel $K(\cdot)$ is a bounded, symmetric density with compact support. Bandwidths h_{1j} and h_{0j} satisfy $h_{1j} \to 0$, $h_{0j} \to 0$, and $nh_{1j} \to \infty$, $nh_{0j} \to \infty$ as $n \to \infty$, with additional conditions ensuring the consistency of locally smoothed KDEs (e.g. rates controlling the pilot bandwidth and damping parameters).*

Under these conditions, the KDE estimators $\hat{g}_1$ and $\hat{g}_0$ converge uniformly to g_1 and g_0, and the estimated CDF $\hat{F}$ converges uniformly to the true F on compact subsets of the support. Consequently, the quasi-log-likelihood $\hat{\ell}_n(\alpha, \beta, \tau)$ converges uniformly to $\ell_n(\alpha, \beta, \tau; F)$, and the maximizer of $\hat{\ell}_n$ converges to the maximizer of the infeasible log-likelihood (up to the usual positive affine transformation).

Theorem 2 (Consistency of the KDE-Based Estimator). *Suppose the latent outcome model and threshold structure are correctly specified, and Assumptions 1–2 hold. Then, as $n \to \infty$,*

$$\left|(\hat{\alpha}, \hat{\beta}, \hat{\tau}) - (\alpha, \beta, \tau)\right| = o_p(1),$$

up to an arbitrary positive affine transformation $(a + c\alpha, c\beta, c\tau)$ with $c > 0$. In particular, the latent treatment effects τ are consistently estimated up to scale.

Because the ordinal model is only identified up to scale, we impose the error-variance normalization $\mathrm{Var}(\varepsilon_i) = 1$ ex post by rescaling the coefficients using the estimated error variance $\hat{\sigma}_\varepsilon^2$ as described in Section 3. Under this normalization, the KDE-based estimator yields treatment effects that are directly comparable to those from ordered probit, ordered logit (after rescaling), and the normalizing-flow-based estimator.

B. Normalizing Flows

This appendix provides additional detail on the normalizing-flow-based estimator described in Section 3. We briefly review the change-of-variables formula, define the ordinal likelihood with a learned error distribution, and state a consistency result.

B.1. Change-of-Variables Formula

Let Z be a random variable with a known base density $f_Z(z)$ (for example, the standard normal density), and let $T_\theta : \mathbb{R} \to \mathbb{R}$ be an invertible, differentiable transformation parameterized by θ. Define

$$\varepsilon = T_\theta(Z).$$

Then, by the change-of-variables formula, the density of ε is

$$f_\theta(\varepsilon) = f_Z\big(T_\theta^{-1}(\varepsilon)\big)\left|\frac{d}{d\varepsilon}T_\theta^{-1}(\varepsilon)\right|.$$

The corresponding CDF F_θ is

$$F_\theta(u) = \int_{-\infty}^{u} f_\theta(t), dt.$$

We refer to the family $T_{\theta\theta\in\Theta}$ as a *normalizing flow*. By choosing a sufficiently flexible family (e.g. coupling layers, spline flows), we can approximate a wide class of univariate distributions with tractable densities and gradients.

B.2. Ordinal Likelihood with a Flow-Based Error Distribution

We embed the flow-based error distribution into the latent outcome model,

$$Y_i^* = f(X_i,\beta) + D_i^\top\tau + \varepsilon_i, \qquad \varepsilon_i = T_\theta(Z_i), \quad Z_i \sim \mathcal{N}(0,1),$$

with thresholds α defining the observed ordinal outcome:

$$Y_i = j \quad\Longleftrightarrow\quad \alpha_{j-1} < Y_i^* \le \alpha_j.$$

Let f_θ and F_θ denote the density and CDF of ε_i implied by the flow T_θ. Conditional on (X_i, D_i), the probability of category j is

$$\mathbb{P}(\,Y_i = j \mid X_i, D_i; \alpha, \beta, \tau, \theta) = F_\theta\big(\alpha_j - V_i(\beta,\tau)\big) - F_\theta\big(\alpha_{j-1} - V_i(\beta,\tau)\big),$$

where $V_i(\beta,\tau) = f(X_i,\beta) + D_i^\top\tau$. The sample log-likelihood is

$$\ell_n(\alpha,\beta,\tau,\theta) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j=0}^{J} 1_{\,Y_i=j}\log\left[F_\theta\big(\alpha_j - V_i(\beta,\tau)\big) - F_\theta\big(\alpha_{j-1} - V_i(\beta,\tau)\big)\right].$$

We estimate $(\alpha,\beta,\tau,\theta)$ jointly by maximizing $\ell_n(\alpha,\beta,\tau,\theta)$ with respect to all parameters. Because T_θ is invertible and differentiable, both f_θ and F_θ are tractable, and gradients of the log-likelihood with respect to $(\alpha,\beta,\tau,\theta)$ can be computed efficiently using automatic differentiation.

B.3. Consistency of the Flow-Based Estimator

This section briefly states a consistency result for the Normalizing-Flow-Based estimator. Let $(\alpha_0, \beta_0, \tau_0, F_0)$ denote the true parameters and error distribution. Suppose that the true error distribution F_0 belongs to the closure of the flow family $F_\theta : \theta \in \Theta$, and that the latent outcome model and threshold structure are correctly specified.

Assumption 7 (Flow Approximation and Identification). *The flow family $F_\theta : \theta \in \Theta$ is rich enough that there exists $\theta_0 \in \Theta$ such that $F_{\theta_0} = F_0$ (or F_{θ_0} approximates F_0 arbitrarily well). The parameter vector $(\alpha, \beta, \tau, \theta)$ is identified up to a positive affine transformation $(a + c\alpha, c\beta, c\tau)$ with $c > 0$.*

Assumption 8 (Regularity Conditions). *The parameter space for $(\alpha, \beta, \tau, \theta)$ is compact or can be restricted to a compact subset; the log-likelihood function $\ell_n(\alpha, \beta, \tau, \theta)$ satisfies standard regularity conditions (continuity, differentiability, integrability); and the model is correctly specified in the sense that the true data-generating process is in the model class for some $(\alpha_0, \beta_0, \tau_0, \theta_0)$.*

Under these conditions, standard maximum likelihood theory implies that the flow-based estimator is consistent (up to scale) and asymptotically normal.

Theorem 3 (Consistency of the Flow-Based Estimator). *Suppose the latent outcome model and threshold structure are correctly specified, and Assumptions 3–4 hold. Then, as $n \to \infty$,*

$$\left|(\hat{\alpha}, \hat{\beta}, \hat{\tau}, \hat{\theta}) - (\alpha_0, \beta_0, \tau_0, \theta_0)\right| = o_p(1),$$

up to an arbitrary positive affine transformation $(a + c\alpha_0, c\beta_0, c\tau_0)$ with $c > 0$. In particular, the latent treatment effects τ_0 are consistently estimated up to scale.

As with the KDE-based estimator, we impose the error-variance normalization $\mathrm{Var}(\varepsilon_i) = 1$ ex post by rescaling the coefficients using the estimated error variance. Given $\hat{\theta}$, we estimate

$$\hat{\sigma}\varepsilon^2 = \mathrm{Var}\hat{F}\theta(\varepsilon_i) \approx \frac{1}{M} \sum m = 1^M \left(T_{\hat{\theta}}(Z_m)\right)^2,$$

where $Z_m \sim \mathcal{N}(0, 1)$ are independent draws from the base distribution. We then report

$$\hat{\tau}^{\text{unit-var}} = \frac{\hat{\tau}}{\hat{\sigma}\varepsilon}, \qquad \hat{\beta}^{\text{unit-var}} = \frac{\hat{\beta}}{\hat{\sigma}\varepsilon},$$

so that all treatment effects are expressed on the same unit-error-variance latent scale as in the main text.