Partial ID for Ordered Choices with Regressor measured in Bins

Chanhyuk Park

Abstract

As survey experiments and causal inference frameworks gain importance, individual-level survey data has become increasingly popular. These surveys often measure outcomes using ordered responses, and some control or independent variables are recorded in bins. Statistical analysis of such data requires careful consideration, as failing to account for differences in response interpretation, misspecification of underlying distributions, and uncertainties arising from the imprecise measurement of variables measured in bins can lead to biased and inconsistent results. Research designs focused on causal inference, such as survey experiments or DiD design with observational data, may be harmed more by these challenges, as biases can sometimes be large enough to alter the sign of estimated coefficients. This paper introduces a semiparametric partial identification method that is more robust to model misspecification and the presence of variables measured in bins than standard Orderd Probit or logit models. Monte Carlo simulation results demonstrate that the proposed approach better accounts for heterogeneity, accommodates different distributional forms of outcomes, and effectively addresses biases from variables measured in bins.

1 Introduction

Individual-level survey has been used to study micro-level political opinions. As the emphasis on causal inference frameworks increases and survey experiments become more common, survey data continues to gain more popularity. One common situation that researchers regularly encounter is the pair of ordinal outcome variable and at least one of variable that is only measured in bins. This paper shows that the standard method to deal with ordinal outcomes and binned covariates may lead to inconsistent and biased estimation, and introduces a semiparametric partial identification approach.

Ordinal outcomes, although the true values of these opinions would lie on a continuous scale, are typically measured using ordered-choice questions (e.g., *strongly disagree, disagree, neither agree nor disagree, agree, strongly agree*). Political opinons on variable policy areas are great examples of this: politicians' approval ratings (Canes-Wrone and De Marchi, 2002; Kriner and Schwartz, 2009), support for redistribution (Alt and Iversen, 2017; Magni, 2021), and foreign policy preferences (Scheve and Slaughter, 2001; Mayda and Rodrik, 2005). These outcomes put two obstacles in statistical analysis. First, respondents may not interpret response scales uniformly. Some individuals may have lower (or higher) thresholds than others, meaning that one person's strongly agree might be equivalent to another's agree King et al. (2004). Second, the true distribution of the outcome is unobservable, and misspecification of this distribution can lead to inconsistent and biased results (Manski, 1988; Greene and Hensher, 2010; Bond and Lang, 2018).

In addition to these two problems, binned variables add potential bias in estimation due to its imprecise measurement. Variables such as income, asset, education (schooling year), or age, are often measured in bins for privacy reasons and researchers usually observe only the lower and upper bounds instead of the exact values. Whether a bracketed variable serves as a primary independent variable or a control, the uncertainty from its imprecise measurement may introduce additional bias in estimation and identification.

Ordered logit and Orderd Probit models are by far the most popular choices in analyzing data with ordinal outcomes. They effectively incorporate the ordinal character of the outcome variable and also guarantees fast estimation through the maximum likelihood estimation (MLE). However, these strength do not come without costs. Both models rely on strong distributional assumptions (e.g., Orderd Probit assumes a normally distributed outcome), making them vulnerable to bias from distributional misspecification. Additionally, these models are generally not robust to uncertainty introduced by bracketed control or independent variables.

The magnitude of bias arising from these three challenges can be significant enough to alter the sign of estimated coefficients—an especially serious issue given that survey-based research designs, including survey experiments and observational studies using difference-in-differences (DiD) designs, often involve causal interpretation.

To address these issues, this paper introduces a semiparametric partial identification approach based on the Modified Maximum Score (MMS) estimation technique from Manski and Tamer (2002), further developed in Wang and Chen (2022). Wang and Chen (2022)'s approach offers several advantages over standard Orderd Probit and logit models. Being semiparametric, it avoids strong distributional assumptions and is therefore less prone to biases from model misspecification and response scale heterogeneity. Moreover, by explicitly accounting for bracketed variables – whether as key independent variables or as controls –this method provides more robust results in most survey analysis settings.

To evaluate the performance of this approach, this paper presents Monte Carlo simulation results. These simulations demonstrate that when the true distribution deviates significantly from normality, standard Orderd Probit and logit models yield biased estimates, whereas the proposed approach remains more robust. Furthermore, the proposed method offers more precise estimation even in the presence of bracketed regressors.

2 Statistical Models for Ordered Outcomes

Political science research frequently relies on outcomes measured in ordered responses, and its popularity continues to grow as causal inference methods and individual-level survey experiments receive increasing attention. Due to privacy concerns, these surveys often include bracketed variables, such as income and assets, where respondents select a range with lower and upper bounds rather than reporting exact values.

The two most commonly used statistical models for analyzing such data are Orderd Probit and Ordered logit. These models require strong distributional assumptions; for example, Orderd Probit assumes that the conditional distribution of outcomes follows a normal distribution. While these assumptions facilitate fast and efficient estimation using maximum likelihood, many researchers have criticized them for failing to account for three key challenges in survey data: (1) variations in individual interpretation of response scales, (2) misspecification of the true outcome distribution, and (3) uncertainty arising from imprecisely measured bracketed variables.

First, respondents may not interpret response scales uniformly. Political opinion questions are typically measured using five-point Likert scales, which are inherently subjective. Some individuals may have lower (or higher) thresholds than others, meaning that one person's *strongly agree* might be equivalent to another's *agree* King et al. (2004). As shown in Figure 1, although both A and B have same opinion denoted by the red arrow, they may answer differently because they interpret the scale differently.

Aldrich and McKelvey (1977) examined interpretability issues when recovering politicians' ideological positions from ordinal survey responses, though without directly addressing estimation concerns. King et al. (2004) proposed the "anchored vignettes" approach to mitigate these issues at the survey design stage. This method introduces standardized example questions designed to capture respondents' interpretations of key concepts, allowing researchers to adjust responses accordingly. King and Wand (2007) further refined the approach by developing methods to evaluate anchoring vignettes. However, a major limitation of this approach is its reliance on the assumption that respondents interpret both vignettes and primary survey questions consistently—an assumption that is difficult to test empirically.

Second, misspecification of the outcome distribution can lead to inconsistent and biased results (Manski, 1988; Greene and Hensher, 2010; Bond and Lang, 2018). Bond and Lang (2018) demonstrated that naive estimation using Ordered logit or Orderd Probit—which impose strong distributional assumptions—can sometimes lead to conclusions opposite to the true effect if the ac-

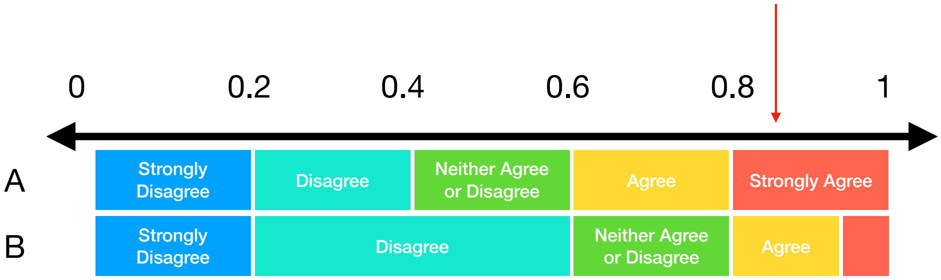

Figure 1: Different interpretation of the scale

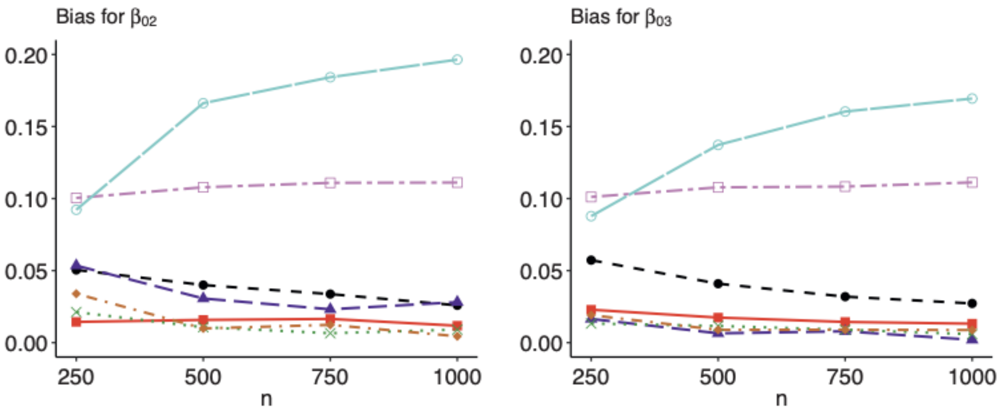

Figure 2: Biases in coefficients when the distribution of the error term is exponential (from Liu and Yu (2024)). Y-axis denotes the size of the bias of the coefficients, and X-axis denotes sample sizes. Light blue lines are biases of the results from Ordered logit models, and light purple lines are biases of the results from Orderd Probit models.

tual outcome distribution deviates significantly from the assumed model. Figure 2 shows the biases in coefficients when the true distribution of error term is exponential for the different sample sizes on the x-axis. Both Orderd Probit (light purple line) and Ordered logit (light blue line) models have large biases, and these biases do not decrease as the sample size increases. This issue is difficult to detect because Ordered logit and Orderd Probit models often produce similar results. Although statistical tests for distributional assumptions exist, they do not provide alternative models when those assumptions are violated (Bera and Jarque, 1982; Glewwe, 1997; Weiss, 1997).

Recently, Bloem (2022) introduced a sensitivity analysis for distributional assumptions. Building on Schroder and Yitzhaki (2017), Bloem (2022) proposed a robustness test for plausible monotonic increasing transformations of the observed ordinal scale's distribution. This framework is valuable because it allows researchers to assess how robust their findings are to certain distributional changes, such as globally concave and convex transformations or transformations with an inflection point. However, as the author notes, despite covering many theoretically plausible cases, the study remains limited to a restricted set of distributional forms.

On the other hands, semiparametric approaches have been developed to mitigate both response scale interpretation issues and potential distributional misspecification. The literature can be roughly divided in two categories. Kernel-based approaches estimate the error distribution nonparametrically using kernel estimation strategy. Lewbel (2000) was one of the first to attempt relaxing both assumptions in this approach. Klein and Sherman (2002) introduced a shift-restriction-based approach that uses kernel estimation for both cut points and regression coefficients, providing greater flexibility. One downside of this approach is that researchers to decide which kernels to be used, and the performance largely depends on this decision.

Some semiparametric approaches do not require input from researchers. Variants of the maximum score and maximum rank estimation are prime examples Lee (1992) extended Manski (1985)'s maximum score estimation model for binary outcomes to ordered choice settings. Liu and Yu (2024) built on Klein and Sherman (2002) by applying isotonic regression and maximum rank estimation strategy instead of kernel methods. Ito (2021) leveraged Monte Carlo resampling to construct likelihoods, enabling estimation without strict distributional assumptions.

Although previously mentioned semiparametric approaches offer greater robustness to both issues of different interpretation among respondents and distributional misspecification, most models care less about the additional complexity introduced by bracketed variables, which are frequently used in survey data. Whether a bracketed variable serves as a key independent variable or a control, its imprecise measurement may further bias the estimation of the parameters. One recent study by Chan, Matyas and Reguly (2024) attempts to address bracketed variables by leveraging multiple survey questions with different discretization schemes. Their findings suggest that access to alternative measures with varying bracket definitions improves distributional approximation, leading to point identification. While the results from their research suggest that researchers can achieve both robust estimation and enhanced respondent privacy by including multiple measures of the same variable, a key limitation is that such alternative measures are rarely available in practice.

Building upon these previous research efforts, this paper introduces a semiparametric partial identification approach developed in Wang and Chen (2022) that offers greater robustness to the three challenges discussed above in individual-level survey data. This approach extends one of the most prominent semiparametric methods – the Generalized Maximum Score estimator proposed by Lee (1992) – to explicitly account for cases in which independent or control variables are measured in brackets. Being another variant of the maximum score estimation, this approach does not require researchers to decide tuning parameters such as bandwidth for kernels.

3 Semiparametric Partial Identification Approach

This paper deals with the common setting in political science studies that deals with individual level survey data, where (a) the outcome is observed in ordinal way, (b) at least one of the regressors is measured in brackets.

Suppose a true or latent value of the outcome is continuous and denote it with Y^*. We further assume that the true data-generating process (DGP) is captured as:

$$Y^* = X^T\beta_1 + v^T\beta_2 + \epsilon$$

, where X is the vector of regressors and v is the bracket-valued regressor and ϵ is an error term.

We do not observe the Y^* directly, but observe the ordinal choices, Y, made based on Y^*.

$$Y = \begin{cases} 0 & \alpha_{-1} \leq Y^* \leq \alpha_0 \\ 1 & \alpha_0 \leq Y^* \leq \alpha_1 \\ \vdots & \\ k & \alpha_{k-1} \leq Y^* \leq \alpha_k \end{cases}$$

, where α_k denotes the cut points for each ordinal category. Conventionally, $-\infty = \alpha_{-1} \leq \alpha_0 \leq \ldots \leq \alpha_k = \infty$.

Also, we do not observe v directly, but only informed the lower bound of v_0 and the upper bound of v_1 for each unit. Usually, some sensitive information such as income, asset and education levels are measured in this manner.

Ordered probit and logit models estimate the modeling parameters, βs and cut-points (αs) based on distributional assumptions on error terms. However, the estimator from these models are inconsistent and biased if the distributional assumptions cannot be met. For now, let us assume that we observe the true v.

For example, Orderd Probit model assume that the error term (ϵ) follows the standard normal distribution. Then the probability of observing $Y = j$ is given by:

$$\begin{aligned}
\mathbb{P}\left(Y = j \mid X\right) &= \mathbb{P}\left(Y^* \leq \alpha_j\right) - \mathbb{P}\left(Y^* > \alpha_j - 1\right) \\
&= \mathbb{P}\left(\epsilon \leq \alpha_j - X_1^\beta - v^T\beta_2\right) - \mathbb{P}\left(\epsilon > \alpha_{j-1} - X_1^\beta - v^T\beta_2\right) \\
&= \Phi(\alpha j - X^T\beta_1 - v^T\beta_2) - \Phi(\alpha j - 1 - X^T\beta_1 - v^T\beta_2)
\end{aligned}$$

, where $\Phi(\cdot)$ is the standard normal cumulative distribution function (CDF).

The parameters can be easily estimated. The MLE gives:

$$(\beta_1, \beta_2, \alpha) = \arg \max_{\beta_1, \beta_2, \alpha} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log\left(\Phi(\alpha_j - X^T\beta_1 - v^T\beta_2) - \Phi(\alpha_{j-1} - X^T\beta_1 - v^T\beta_2)\right)\right]$$

Suppose that the true distribution of ϵ is not normal, and denote the CDF as $F(\cdot)$. If the distribution is symmetric, then the MLE under this true distribution would be:

$$(\beta_1, \beta_2, \alpha) = \arg \max_{\beta_1, \beta_2, \alpha} \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \log\left(F(\alpha_j - X^T\beta_1 - v^T\beta_2) - F(\alpha_{j-1} - X^T\beta_1 - v^T\beta_2)\right)\right]$$

Since $F \neq \Phi$, the estimators from Orderd Probit model are inconsistent, because increase in sample size does not make the two distributions closer, and therefore the estimators from two MLE will not converge. Also, the estimators are biased, but the size and direction of the bias depend on the shape of F.

Now we go back to the original setting and only observe v_0 and v_1, instead of v. It is easy to see that this imprecise measure of v may lead to another bias, by affecting the log likelihood function. The size and direction of the bias also depend on the shape of F.

4 Generalized Maximum Score Estimator

Generalized Modified Maximum Score (GMMS) estimator suggested by Wang and Chen (2022) can deal with the inconsistency and the bias caused by distributional misspecification and the bias from binned covariates. The estimator is based on the modified maximum score estimator (MMS) (Manski and Tamer, 2002) and the generalized maximum score (GMS) estimator (Lee, 1992), both of which are based on the maximum score estimator (Manski, 1975).

GMMS imposes below assumptions:

Assumption 1. $quantile_\alpha(\epsilon \mid X, v) = 0$

Assumption 2. $\mathbb{P}\left(\epsilon \mid X, v, v_0, v_1\right) = \mathbb{P}\left(\epsilon \mid X, v\right)$

Assumption 3. $\beta_2 > 0$

Following Manski (1985), Horowitz (1992), and Lee (1992), the first assumption, that the conditional α quantile of the error term equals 0, means that at least at one point, the outcome is fully explained by covariates without error. This is the only restriction on the distribution of the error term, thereby the conditional distribution of the outcome. This semiparametric nature of the estimator mitigates problem of different interpretation among respondents and the bias from misspecification. The second and third assumptions deal with the bias generated by imprecise measurement of some variables in the regression equation. For the rest of the paper, I assume that $median(\epsilon \mid X, v) = 0$.

The second assumption states that the error term are conditionally independent from v_0 and v_1 given both X and v. In other words, if the exact value of v is known, then v_0 and v_1 carries no further information. The last assumption requires v to have strict monotone relationship with the outcome, and this should be justified by the data, theory or results from previous research. If v is thought to have native relationship, then simple reverse coding will satisfy the assumption.

Before this, need to introduce partial identification. For the notatoinal ease define $Y_{mi} = \mathbf{1}\{Y_i > m\}$. Let $P_n(Y_{mi} = m \mid X_i, V_i^0, V_i^1)$ be consistent estimates of $P(Y_m = m \mid X, V^0, V^1)$ for $m = 1, \ldots, M-1$. Furthermore, define $\lambda_{mn}(X_i, V_i^0, V_i^1) = \mathbf{1}\left\{P_n(Y_i^m = m \mid X_i, V_i^0, V_i^1) > \frac{1}{2}\right\}$.

Then, the finite sample estimation of the identified set Θ_n using GMMS is given as:

$$\Theta_n = \{b : S_n(b) \geq \max_{C \in \mathcal{B}} S_n(c) - \varepsilon_n\} \tag{1}$$

, where

$$S_n(b^s) = \frac{1}{n} \sum_{i=1}^{n} \sum_{m=1}^{M-1} \left(Y_{mi} - \frac{1}{2} \right) \Big[\lambda_{mn}(X_i, V_{0i}, V_{1i}) \cdot \operatorname{sgn}(\tilde{X}_i' b + V_i^1 + b_{1m}) Y_{mi} + \\ (1 - \lambda_{mn}(X_i, V_{0i}, V_{1i})) \cdot \operatorname{sgn}(\tilde{X}_i' b + V_{0i} + b_{1m}) \Big] \tag{2}$$

, for some $\varepsilon_n > 0$. The basic intuition here is to maximize the sign consistency for each Y_j.

ε_n gives room for partial identification, so it is preferred to be as small as possible, but not to small. Chernozhukov, Hong and Tamer (2007) proves that when $\varepsilon_n \propto \frac{\ln(n)}{n}$, GMMS estimator is $\sqrt{n}$ consistent. To follow their proof, I choose to set $\varepsilon_n = \frac{\ln(n)}{n}$.

The conditional probability $P_n(Y_{mi} = m \mid X_i, V_i^0, V_i^1)$ can be estimated nonparametrically from the data. Wang and Chen (2022) suggested to use kernel based smoothing estimator, but I follow Hall, Racine and Li (2004), who suggested direct nonparametric estimation of conditional distribution of discrete variable via cross-validation.

5 Monte Carlo Simulations

I tried several Monte Carlo experiments to test the performance of GMMS, and compare it with those of standard Orderd Probit and Logit models. Here I consider the case where the outcome is a ordinal variable with 5 levels:

$$Y = \begin{cases} 1 & \text{if} \quad \alpha_{-1} \leq Y^* \leq \alpha_0 \\ \vdots & \\ 5 & \text{if} \quad \alpha_{-3} \leq Y^* \leq \alpha_4 \end{cases}$$

, where $Y^* = \beta_0 + X^T\beta_1 + v^T\beta_2$. X and v are one-dimensional, for simplicity. $X \sim \mathcal{N}(1,2)$, and $v \sim \mathcal{N}(0,2)$. v_1 is derived by round up v to the nearest integer, and v_0 equals $v_1 - 1$. The coefficients, $(\beta_0, \beta_1, \beta_2)$ are set to $(0.5, -0.5, 0.5)$.

I tested for two different distribution for the error term, ϵ, the exponential distribution and the student-t distribution. Both distributions have relatively fat tailed and skewed compared to the standard normal distribution and frequently observed in real world settings. I also consider three sample sizes of 200, 500, 1,000. The number of Monte Carlo replication is 100, and all estimations are the means of the 100 replication results.

Table shows the estimation of coefficients for X and v from GMMS, Ordered Probit and Ordered Logit models.

References

Aldrich, John H. and Richard D. McKelvey. 1977. “A Method of Scaling with Applications to the 1968 and 1972 Presidential Elections.” *The American Political Science Review* 71(1):111–130.

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

Bera, Anil K and Carlos M Jarque. 1982. “Model specification tests: A simultaneous approach.” *Journal of Econometrics* 20(1):59–82.

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.

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

Chan, Felix, Laszlo Matyas and Agoston Reguly. 2024. “Modelling with Discretized Variables.”.

Chernozhukov, Victor, Han Hong and Elie Tamer. 2007. "Estimation and Confidence Regions for Parameter Sets in Econometric Models." *Econometrica* 75(5):1243–1284.

Glewwe, P. 1997. "A test of the normality assumption in ordered probit model." *Econometric Reviews* 16(1):1–19.

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

Hall, Peter, Jeff Racine and Qi Li. 2004. "Cross-Validation and the Estimation of Conditional Probability Densities." *Journal of the American Statistical Association* 99(468):1015–1026.

Horowitz, Joel L. 1992. "A Smoothed Maximum Score Estimator for the Binary Response Model." *Econometrica* 60(3):505.

Ito, Takahiro. 2021. "Binary and Ordered Response Models in Randomized Experiments." *SSRN Electronic Journal* .

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 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.

Lee, Myoung-jae. 1992. "Median regression for ordered discrete response." *Journal of Econometrics* 51(1):59–77.

Lewbel, Arthur. 2000. "Semiparametric qualitative response model estimation with unknown heteroscedasticity or instrumental variables." *Journal of Econometrics* 97(1):145–177.

Liu, Ruixuan and Zhengfei Yu. 2024. "Simple Semiparametric Estimation of Ordered Response Models." *Econometric Theory* 40(1):1–36.

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. 1975. "Maximum score estimation of the stochastic utility model of choice." *Journal of Econometrics* 3(3):205–228.

Manski, Charles F. 1985. "Semiparametric analysis of discrete response: Asymptotic properties of the maximum score estimator." *Journal of Econometrics* 27(3):313–333.

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

Manski, Charles F. and Elie Tamer. 2002. "Inference on Regressions with Interval Data on a Regressor or Outcome." *Econometrica* 70(2):519–546.

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

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

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

Wang, Xi and Songnian Chen. 2022. "Partial Identification and Estimation of Semiparametric Ordered Response Models with Interval Regressor Data." *Oxford Bulletin of Economics and Statistics* 84(4):830–849.

Weiss, Andrew A. 1997. “Specification tests in ordered logit and probit models.” *Econometric Reviews* 16(4):361–391.