Journal of Quantitative Criminology
https://doi.org/10.1007/s10940-025-09627-5

ORIGINAL PAPER

When the Mean is Misleading: a Guide to Ordered Regression for Meaningfully Modeling Ordinal Outcomes

Jonathan R. Brauer[1] · **Jacob C. Day**[2]

Received: 7 May 2025 / Accepted: 11 August 2025

Abstract

Objectives We demonstrate the utility of ordered regression models for analyzing ordinal outcomes frequently encountered in criminology, comparing their predictive accuracy and estimation against conventional linear models.

Methods The study employs simulation to illustrate how linear and ordered probit models represent ordinal data distributions. Subsequently, it analyzes national survey data ($N = 1{,}150$) on racial differences in fear of police from Pickett et al. (2022). Linear models (item-specific, multilevel, normed-scale) are compared to Bayesian cumulative probit models using cross-validation, visualizations (mosaic plots; predicted probabilities), and a comprehensive R tutorial.

Results Simulations confirm linear models poorly represent non-normal ordinal distributions, while ordered probit models accurately recover observed patterns. Cross-validation reveals dramatically superior predictive performance for ordinal models (e.g., $\Delta\text{ELPD} \approx 1{,}920$, $\text{SE} = 48.6$). While both approaches yield convergent average effect estimates ($d = 0.98$ vs. $d = 0.97$), they generate substantially different category-specific predictions. Ordered probit models estimate 32.8% of Black participants report being "very afraid" of police killing compared to 7.6% of White participants, while linear models predict only 11.0% and 1.1%, respectively. Observed proportions were 32.1% and 8.6%.

Conclusions Ordered regression models provide superior predictive accuracy and precise quantification of distributional patterns while yielding equivalent average effects when linear models perform adequately. This demonstrates that methodological choice affects substantive understanding of policy-relevant phenomena. Although the paper limits its focus to cumulative probit models and a single substantive domain, the tutorial, supplementary material, and review of relevant literature are meant to encourage practical adoption and elaboration of ordinal methods in criminology.

Keywords Ordinal data · Likert · Cumulative probit · Linear model · Bayesian modeling · Precision · Fear of police

Extended author information available on the last page of the article

Published online: 10 December 2025

Introduction

Regression modeling techniques are central to quantitative criminological research, enabling researchers to describe patterns, test theories, and evaluate policies. The reliability and validity of our findings, however, depend heavily on the appropriate application of statistical techniques to the types of data commonly collected in the field. A frequent challenge arises from the prevalence of ordinal data, particularly from surveys using Likert-type scales to measure attitudes, perceptions, or behavioral frequencies. While widely used, the analysis of such data frequently defaults to treating ordinal responses as if they were continuous, typically employing linear regression models.

This common practice, while convenient, rests on assumptions – such as equal intervals between categories and normally distributed errors – that may not align well with the nature of ordinal measurements. When model assumptions mismatch the data structure, the resulting estimates of relationships and effect magnitudes may be less accurate or potentially misleading. Ensuring our analytical methods faithfully represent the information contained in our data is fundamental to building a robust and cumulative science. Striving for accuracy and precision in estimation allows us to move beyond simple directional statements ("X is positively related to Y") towards quantifying *how much* X relates to Y on meaningful response scales, a step crucial for developing more refined theories and informing evidence-based practice (cf. Tukey 1977).

Ordered regression models (e.g., ordered probit, ordered logit) offer a suite of techniques specifically designed for analyzing dependent variables with ordered categories. These models make different assumptions than linear models, explicitly accounting for the ranked nature of the data and modeling the process through latent thresholds. By aligning statistical assumptions more closely with the properties of ordinal data, these methods can provide a more accurate representation of underlying patterns and potentially yield different substantive conclusions compared to linear approaches, particularly regarding the magnitude of effects.

Despite their long history and availability in standard software, ordered regression models arguably remain underutilized in criminology relative to the prevalence of ordinal data. This manuscript aims to provide a practical tutorial and demonstration for criminologists on the application and interpretation of ordered regression models, using the cumulative probit model as a primary example. We first use simulated data to illustrate key differences in how linear and ordered probit models represent ordinal distributions. We then conduct a detailed analysis of data from Pickett et al.'s (2022) study on racial differences in fear of the police, comparing findings from linear models (including one replicated estimate from the original study) with those from ordered probit models. Our core goals are to illustrate:

- Using mosaic plots for visualizing bivariate ordinal data.
- Comparing the fit of intercept-only linear and ordered probit models to observed ordinal distributions.
- Comparing estimates of group differences (Black vs. White fear of police) derived from both model types, both on the latent scale and, crucially, as predicted probabilities for specific response categories.
- Comparing estimates from multilevel ordered probit models versus linear models using composite scales.

Through this comparative demonstration, we aim to equip researchers with a clearer understanding of how ordered regression models work, the types of insights they can provide (especially regarding precise effect magnitudes and distributional patterns), and the potential consequences of applying linear models to ordinal data. The goal is not to universally condemn linear models, but rather to illustrate the value of ordered regression as an important tool for enhancing analytical accuracy when working with ordinal outcomes, thereby contributing to more precise and reliable criminological knowledge.

Background

Why Bother With Ordered Regression Modeling?

Ordered regression models have deep roots in over a century of methodological development. The pioneers of probit and logit modeling (e.g., Fechner 1860; Bliss 1934; Berkson 1944) grappled with the fundamental challenge of accurately measuring and analyzing outcomes like human perception and response variation, particularly when responses fall into ordered categories but the "distance" between those categories is unknown or unequal. This challenge remains highly relevant for criminologists today, given the frequent use of ordinal measures like Likert scales. While linear regression is often applied to such data, it is important to understand its underlying assumptions and how they relate to the properties of ordinal data. Applying models whose assumptions are strongly violated by the data can potentially lead to inaccurate or misleading inferences regarding effect magnitudes or even direction.

Numerous scholars have detailed the potential issues when applying linear ("metric") models to ordinal data (e.g., Agresti 1989, 2010; Bürkner & Vuorre 2019; Greene & Hensher 2010; Liddell & Kruschke, 2018; Winship & Mare 1984). These discussions highlight several key areas where the assumptions of linear models may clash with the properties of ordinal data:

1. **Unequal Intervals Between Categories:** Linear models treat the numerical coding of ordinal categories (e.g., 1, 2, 3, 4, 5) as if they represent equal intervals on an underlying scale. For instance, they assume the difference between "Strongly Disagree = 1" and "Disagree = 2" is equivalent to the difference between "Disagree = 2" and "Neither = 3." This assumption is rarely justified empirically for subjective ratings and, when violated, means that a one-unit change on the predictor does not correspond to a constant change on the outcome scale, potentially distorting estimates of relationships.
2. **Inappropriate Error Distribution Assumptions:** Linear models typically assume that measurement errors (residuals) are normally distributed around the predicted value. This assumption is often incompatible with the discrete, bounded nature of ordinal responses, especially when data are skewed or exhibit floor/ceiling effects. This mismatch can lead to issues like heteroscedasticity (unequal error variance), potentially resulting in biased standard errors and impacting the validity of statistical tests.
3. **Distortion from Ceiling and Floor Effects:** Criminological data, particularly survey responses about rare behaviors or strong attitudes, often cluster at the extremes of an ordinal scale (e.g., many respondents reporting "Never" or "Very Afraid"). Also known

as censoring issues (cf. Osgood et al. 2002), linear models struggle to accurately represent these clusters due to the inaccurate assumption of an unbounded latent variable. This can lead to underestimation of the true strength of relationships involving those scale extremes, a phenomenon we illustrate later in this tutorial.

4. **Latent Threshold Shifts:** Liddell and Kruschke (2018) also demonstrate how different groups might use response categories differently – that is, they might have different internal thresholds for selecting, say, "Agree" versus "Strongly Agree" – even if their mean level on the underlying latent trait is identical. Linear models cannot easily account for such threshold differences and may consequently indicate spurious group differences that are artifacts of response style rather than substantive differences in the construct.

These potential issues are not merely theoretical concerns; they can manifest in practical research contexts, potentially leading to incorrect inferences about the presence, direction, or magnitude of effects (e.g., false positives or false negatives).

Ordered regression models (like cumulative probit and logit) are designed to address these challenges. They do so by explicitly modeling an underlying latent continuous variable and estimating threshold parameters that map segments of this latent variable onto the observed ordinal categories. This approach avoids assuming equal intervals, inherently handles the bounded nature of the scale, accommodates different threshold patterns, and can better represent skewed distributions and ceiling/floor effects. Importantly, when data *do* approximate interval-level measurement with normally distributed errors, linear and ordered regression models typically generate comparable estimates.

Criminology boasts a long tradition of applying sophisticated statistical tools to address theoretical and empirical challenges (e.g., Osgood et al. 2002; Greenberg 2016; Britt et al. 2018), including the use of ordered regression models for ordinal outcomes (e.g., Weisburd et al. 2022; Torres et al. 2024). However, the application of linear models to ordinal outcome data remains prevalent throughout the field (e.g., Jones et al. 2022; Svensson & Oberwittler 2021; Madon et al. 2023; Herman & Pogarsky 2024). It is also common practice to apply linear models to composite scales created by averaging ordinal items and then "normalizing" them (e.g., to a 0–100 range), with results often interpreted as percentage point changes (e.g., Mummolo 2018; Peyton et al. 2019; Metcalfe & Pickett 2022; Wozniak et al. 2022; Hannan et al. 2023; Lee et al., 2024; Pickett et al. 2024). Furthermore, researchers sometimes report running both linear and ordinal models but focus interpretation on the linear results, citing "ease of interpretation" or noting only that coefficient signs and statistical significance align across models (e.g., Baranauskas, & Drakulich 2018; Silver 2020; Wheeler et al. 2020; Munksgaard et al. 2023; Smith et al. 2025).

As we will show, these common practices can result in distorted inferences and important yet overlooked patterns in data. Thus, while linear models are valuable tools, understanding their assumptions and limitations alongside the capabilities of alternatives like ordered regression is crucial for making informed methodological choices. This tutorial aims to contribute to that understanding by providing a practical demonstration for criminologists, illustrating *how* ordered regression models work and *what differences* in estimation and interpretation can arise compared to linear models. By showcasing these techniques, we hope to encourage their application to potentially enhance the accuracy and nuance of data-based inferences from the ordinal data prevalent in our field.

Why Ordinality Remains Problematic in Multi-Item Scales

Many researchers appear to assume the problems associated with applying metric models to ordinal data are limited to cases where metric models are applied to *individual* ordinal items. The common practice of scaling such items together presumably makes them interval or interval enough to be robust to the issues we outline above (cf. Carifo and Perla 2008). For a detailed rebuttal of this inaccurate assumption, we refer readers to Liddell and Kruschke (2018), who explicitly demonstrated that averaging across multiple ordinal measurements "does not solve these problems" and confirmed that composite scores share the same fundamental flaws as single ordinal items. For example, their simulations showed how linear models applied to averaged ordinal items could produce false alarms and failures to detect effects, with these problems "not ameliorated" by increasing the number of items. Aggregation preserves the fundamental problem: Non-linear relationships between underlying latent variables and observed ordinal responses create systematic distortions, particularly at extreme values.

Empirical Evidence Favors Ordinal Methods Methodological research typically favors ordinal-specific methods over linear methods for analyzing ordinal data. For example, in SEM contexts, methods like Weighted Least Squares and Diagonally Weighted Least Squares tend to provide more accurate parameter estimates than traditional linear approaches while preserving the multi-item measurement benefits that justify composite scaling (e.g., Flora & Curran, 2004; DiStefano & Morgan, 2014). Though frequently cited for suggesting that items with five or more categories might be treated as continuous in some CFA contexts, Rhemtulla, Brosseau-Liard, and Savalei (2012) found categorical methods tend to be superior with fewer categories and/or when items exhibit asymmetric distributions.

Recent work reinforces concerns about conventional linear aggregation approaches. Donnellan, Usami, and Murayama (2023) demonstrate that ignoring item-specific variations in relationships with predictors, or what they term "random item slopes," leads to inflated Type I (false positive) error rates, especially in large samples. Their real-world analyses show these problematic variations exist in commonly used psychological scales, with standard psychometric indices (Cronbach's alpha; SEM fit statistics) insufficient to detect them. Moreover, and consistent with Liddell and Kruschke's (2018) findings, Kołczyńska and colleagues (2024) note that common approaches like linear re-scaling of ordinal data "come at a cost," leading to "information loss" and potentially "inflated error rates, distorted effect sizes, and even inversions of effects." The robustness of linear models to ordinal outcome variables applies only under highly restrictive conditions rarely met in practice, especially for skewed 5-point Likert scales common in criminological research.

These issues extend to various methodological contexts, including multilevel modeling. For example, Bauer and Sterba (2011) investigated "When can the results of a multilevel linear model fit to an ordinal outcome be trusted?" Their results suggest the answer: "Rarely." Only when responses were roughly normal with seven categories did linear model bias decrease to acceptable levels; they conclude: "In almost all cells of the design, the linear model estimates were inferior to the cumulative logit model estimates." Importantly, multilevel ordinal models can achieve the same analytical goals as traditional multi-item scaling – pooling information across items, estimating latent constructs, controlling for

measurement error – while respecting the categorical nature of the data. Kołczyńska et al. (2024) demonstrate this approach using Bayesian explanatory IRT modeling for ordinal outcomes, explicitly modeling varying thresholds across items within multilevel frameworks while estimating underlying latent traits.

Beyond Average Effects: Estimands, Model Fit, and Ordinal Data

Another relevant perspective, particularly common in econometric analysis, sometimes favors linear models even when the outcome is discrete or ordered. The argument often centers on the desirable property that linear regression can provide unbiased estimates of average effects (e.g., average treatment effects or coefficients representing average relationships between variables), even if the model itself is misspecified in other ways, such as producing predicted probabilities outside the logical bounds or poorly representing the underlying distribution (cf. Angrist & Pischke 2009). From this viewpoint, if the primary goal is estimating the average marginal effect coefficient for causal or descriptive inference about linkages, the simplicity and direct interpretability of the linear model coefficient might be preferred over non-linear models like probit or logit, where marginal effects are conditional or require further calculation.

While the unbiased estimation of an average effect coefficient is a valuable statistical property, its substantive utility depends critically on the appropriateness and meaningfulness of the *estimand* – the specific quantity we aim to estimate – in the context of the data and research question (Lundberg et al. 2021). For ordinal data, the coefficient from a linear regression model represents the estimated average change in the *mean* of the numerically coded ordinal outcome associated with a one-unit change in the predictor. The central issue then becomes: Is the *mean* of an ordinal variable the most meaningful summary or the most relevant target for inference? Fundamentally, this is a question that researchers must answer in the context of their specific research programs, though we suspect that analytical conventions are preventing many from fully engaging with this issue.

As discussed in "Why Bother with Ordered Regression Modeling?" section, treating ordinal data as interval-level and calculating a mean ignores the inherently ranked nature of the categories and assumes equal spacing that rarely holds. Furthermore, the mean can be a poor or even misleading descriptor of central tendency for distributions that are highly skewed, bimodal, or subject to ceiling/floor effects, which are all common features of ordinal data in criminology (as we later show in Fig. 1 with simulated examples). Consequently, obtaining an unbiased estimate of the effect on a potentially uninterpretable summary statistic (the mean of ordered categories) may offer limited insight into the substantive processes generating data. What good is an unbiased estimate of the effect on the mean *when the mean itself is a fundamentally flawed representation of an ordinal response distribution?*

This highlights the importance of considering the *fit* of the model to the data and clearly defining the *substantive quantity of interest* beyond just the average effect on the mean. Are we interested in:

- The effect on an underlying *latent propensity* assumed to generate the ordered responses?
- The effect on the *probability* of observing a specific response category (e.g., "Very Afraid") or exceeding a certain threshold (e.g., "Afraid" or "Very Afraid")?
- How the entire *distribution* of responses shifts across levels of a predictor?

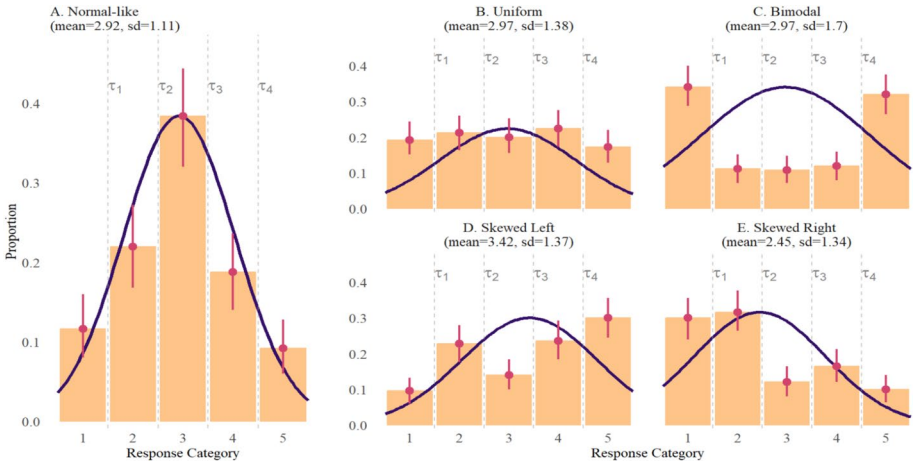

Fig. 1 Visualizing How Metric and Ordered-Probit Models View Ordinal Data. Note: Light bars show observed (simulated) proportions of responses. Dark line overlays linear model-implied normal curve. Point intervals display ordered probit model-implied predicted probabilities with 95% bootstrap confidence intervals. Grey dashed lines illustrate how ordered probit model cuts a 5-category item into four latent thresholds (τ_j). For color figures, see article's online version of record

If these latter quantities are the primary focus, then models explicitly designed to capture the ordinal structure and estimate these probabilities or distributional shifts, such as ordered probit or logit models, are more suitable. While linear models might provide an unbiased estimate of the *average* slope across a latent continuous scale, ordinal models also provide comparable, typically convergent, and more interpretable estimates of the average slope across a latent continuous scale. Meanwhile, linear models demonstrably fail to accurately recover the probabilities within specific categories when the underlying distribution is non-normal (as shown in "Comparing Linear and Ordered Probit Models" section and Fig. 3), undermining their utility for research questions focused on predicted probabilities or specific parts of the outcome distribution. The focus solely on the unbiasedness of the mean-effect coefficient overlooks the crucial aspect of whether the model accurately and completely represents the data generating process across its full range while ignoring the fact that these same quantities can be estimated as or more accurately in ordinal models.

To support broader adoption of techniques that can help us both accurately estimate and move beyond average effects, alongside our reported analysis and primary supplement, we provide a comprehensive R tutorial comparing frequentist and Bayesian linear and multilevel approaches to ordinal modeling (see Appendix 1 table in manuscript for a summary of R tutorial file contents). This standalone tutorial demonstrates practical challenges with frequentist multilevel ordinal models (including common convergence failures), shows how Bayesian alternatives provide more robust solutions, and includes step-by-step code for model fitting, diagnostic testing, and predicted probability generation. The tutorial illustrates how ordered regression techniques allow researchers to more accurately represent observed ordinal distributions and to estimate comparable average effects as well as other substantively meaningful quantities, such as category-specific probabilities and shifts in those probabilities. This appendix is distinct from our primary online supplement, which contains all code and output for the specific results presented in this paper. Ultimately, our

aim is to illustrate how choosing models that better align with the data structure enables more complete and accurate descriptive and inferential analysis.

Assumptions and Limitations of Ordinal Models

Before illustrating their application, below we explicitly present key assumptions of ordinal regression models and compare them with more familiar linear model assumptions:

- *Proportional Odds Assumption.* Cumulative ordinal models assume predictor effects are consistent across threshold boundaries (e.g., the effect of race should be proportionally similar whether distinguishing the lowest category from all others, the two lowest from the three highest, etc.). Linear models make a similar assumption that effects are constant across the latent treated-as-continuous scale. Violations of proportional odds can be diagnosed through formal tests (see Liu et al., 2023) and can be addressed via category-specific modeling (see Section 10 in Appendix 1).
- *Latent Variable Distribution.* Cumulative probit models assume there exists an underlying continuous scale (e.g., like an internal 'fear thermometer') that follows a normal distribution, which gets converted into discrete ordinal responses through threshold boundaries. Linear models make similar normality assumptions (for residuals) while additionally requiring equal intervals and unbounded scales.
- *Threshold Ordering and Monotonicity.* Models assume ordered thresholds ($\tau_1 < \tau_2 < \tau_3 < \tau_4$), which naturally holds for properly coded ordinal data, and monotonic predictor effects, meaning higher predictor values systematically increase or decrease the probability of higher ordinal categories. Instead of estimating thresholds, linear models assume the numerical coding of ordinal categories reflects true equal intervals on a continuous scale, and they make similar monotonicity assumptions about the relationship between predictors and outcomes.

Key Limitations and Comparative Perspective As Allison (1999) and Williams (2009) explain, comparing ordinal regression coefficients across groups can be problematic due to rescaling based on residual variation. This concern also applies to linear models with heteroskedastic errors, but it can be addressed in both cases by focusing on predicted probabilities over raw coefficients (as we do below). When proportional odds assumptions fail, standard models provide averaged effects that may obscure threshold-specific patterns. However, linear models applied to ordinal data violate normality, homoskedasticity, and equal-intervals assumptions simultaneously, often more severely.

Despite limitations, ordinal model assumptions are generally more appropriate for ordinal data than those of linear alternatives. Indeed, given the extensive evidence, Liddell and Kruschke (2018, p.337) directly counter arguments that applying metric models to ordinal data is "innocuous and desirable for its simplicity" by demonstrating that it leads to "systemic errors." Given this, they argue that ordinal models should be the "default treatment of ordinal data" and offer a potent analogy: Dismissing these problems as rare is "like arguing it's okay to drive drunk because accidents rarely happen." They emphasize researchers "cannot know in advance if any given set of data will yield different conclusions when ana-

lyzed with a metric model and an ordered-probit model," but fundamentally, "an ordered-probit model will better describe the data than a metric model."

Ultimately, the choice of modeling approach should be guided by assumption plausibility rather than computational convenience, and ordinal models relax the linear model's restrictive equal-intervals and unbounded-scale assumptions while adding only the proportional odds constraint, which can be tested and relaxed as needed. Below, we attempt to provide the readers with examples that we hope will help them to imagine how the consequences of these choices might affect their own analyses and illustrate the analytical advantages of (multilevel) ordered regression when working with ordinal outcomes.

Comparing Linear and Ordered Probit Models

We begin with a brief comparison of linear and ordered probit models with an accompanying analysis of simulated data to illustrate how they model or "view" data differently. For more detailed comparisons of linear and ordered regression models, we refer readers to the Liddell and Kruschke (2018) and to the many other interdisciplinary topical treatments (e.g., Agresti 1989, 2010; Bürkner & Vuorre 2019; Greene & Hensher 2010; Winship & Mare 1984).

Classic Linear Regression

First, consider the classic linear regression equation:

Equation 1. (Classic linear regression)

$$y_i \sim \text{Normal}\,(\mu_i, \sigma)$$
$$\mu_i = X\beta$$

Where:

- y_i is the observed metric continuous outcome variable for individual *i*.
- $\text{Normal}(\mu_i, \sigma)$ indicates that y_i is assumed to be normally distributed with mean μ_i and standard deviation σ.
- μ_i is the predicted mean of y_i for individual *i*.
- X represents the predictors.
- β represents the regression coefficients (the effect of the predictors on y_i).
- σ represents the standard deviation of the error term.

Cumulative Probit Regression

Under the hood, the ordered probit model posits an underlying continuous latent variable that is transformed into discrete ordered categories through a set of thresholds. Mathematically, the model can be expressed as:

Equation 2. (Cumulative Probit Regression)

$$y*_i \sim \text{Normal}\,(\mu_i, 1)$$
$$\mu_i = X\beta$$
$$y_i = j \text{ if } \tau_{j-1} < y*_i \ \leq \tau_j$$

Where:

- $y*_i$is the unobserved latent continuous variable for individual *i*.
- $\text{Normal}\,(\mu_i, 1)$ indicates that $y*_i$ is assumed to be normally distributed with mean μ_i and standard deviation 1 (the standard deviation is fixed to 1 for identification purposes in ordered probit models).
- $\mu_i$is the predicted mean of $y*_i$ for individual *i*.
- *X* represents the predictors.
- β are regression coefficients.
- $y*_i$ is the observed ordinal category for individual *i*.
- τ_j values represent threshold parameters that divide the latent space into ordered categories.
- $y_i = j$ if $\tau_{j-1} < y*_i \leq \tau_j$ defines how the latent variable $y*_i$ is mapped to the observed ordinal category y_i using thresholds τ_j.

Put simply, these models rely on fundamentally different ways of characterizing outcome data. A linear model assumes that the outcome can be adequately summarized by its (conditional) mean and variance, treating deviations from the mean as normally distributed residuals. While this assumption formally applies to residuals rather than the observed response itself, in practice, the model implicitly treats the outcome distribution as normal when making predictions – generating fitted values that assume a smooth, continuous spread even if the actual data are discrete or clustered. In contrast, the ordered probit model does not assume a normal outcome distribution but instead models an underlying latent variable that follows a normal distribution. It estimates thresholds that map this latent scale onto discrete ordinal categories, producing predictions that align more closely with the categorical nature of the data rather than assuming a continuous normal approximation.

This threshold-based approach allows the ordered probit model to capture the nuances of ordinal data, such as skewness and multimodality, that linear regression cannot. Thus, it is unsurprising that tutorials for various routine to complex applications abound in methodological literatures, such as multilevel modeling (e.g., Bauer and Sterba, 2011), or dynamic SEM (McNeish et al. 2024), or causal mediation with ordinal mediators (Nguyen et al. 2016). In fact, ordered regression techniques can even be applied usefully to continuous or count outcome data (Cameron and Trivedi 2013; Liu et al. 2017) and may improve model fit and reduce estimate bias compared to popular linear or parametric count modeling techniques (e.g., Poisson or negative binomial regression), particularly when outliers, zero-inflation, or overdispersion are present (cf. Jung et al. 2006; Valle et al. 2019).

To help readers visualize how linear and ordered regression models differently "view" data, we simulated different ordinal data distributions and then plotted linear and ordered probit regression predictions with observed response distributions.[1] Specifically, the simulation generates data from five distinct ordinal distributions, each with N=250 simulated observations: (A) Symmetric normal-like distribution; (B) Uniform distribution; (C) Bimodal

[1] While we use Bayesian regression techniques in our primary analysis, here we generate frequentist cumulative probit regression predictions using the *polr* function from *MASS* R package (see Ripley et al. 2013).

distribution; (D) Left-skewed distribution; (E) Right-skewed distribution.[2] The first three distributions were simulated with nearly identical mean values, while the fourth and fifth display opposite skew patterns. Figure 1 provides a visual comparison of these approaches, revealing critical differences in how each method recovers the underlying response distribution.

Visually, the differences are most pronounced in non-normal distributions. For skewed and bimodal distributions, the linear model's assumption of equal intervals and normal distribution can severely misrepresent the underlying data structure, as demonstrated by the normal curve failing to capture the observed response distributions. The consequences of such misfit are quite dire for analyses built upon such assumptions. For example, even simple linear-based analyses of group differences in means (e.g., *t*-tests) would generate near-null point estimates and critical values that fail to statistically distinguish between (A) normal-like, (B) uniform, and (C) bimodal distributions despite obvious distributional differences and distinct data generating processes. Even where differences might be statistically detected, such as between (D) skewed left and (E) skewed right distributions, estimates are difficult to interpret and fail to represent observed distributional differences, and inferences are susceptible to substantial magnitude and even directional errors. For example, the linear model only "sees" about ~ 1 response category difference between D & E distributions, with means implying the typical observation in D is a '3' or '4' and the typical observation in E is a '2' or '3.' Yet, most observations in D were in categories '4' and '5' (not '3' and '4'), whereas most observations in E were in '1' & '2' (not '2' and '3').

The ordered probit model, by contrast, adapts to the observed category proportions, providing a more accurate representation of each data-generating process. This is clearly seen in the point interval estimates that accurately recover the observed proportions. For example, notice in panel (C), the linear model's normal curve fails to capture the bimodal nature of the data, while the ordered probit model's point estimates accurately reflect the two peaks. Similarly, in panels (D) and (E), the linear model's symmetrical curve struggles to represent the skewed distributions, whereas the ordered probit model's thresholds adjust to the observed skew.

In essence, the simulation demonstrates that relying on linear regression for ordinal data, especially when distributions deviate from normality, can lead to a distorted view of the data and potentially misleading conclusions. This visual comparison underscores the importance of choosing a modeling approach that aligns with the nature of the data, a nature that in many criminological contexts is imposed by common survey measurement practices. In doing so, it highlights the advantages of ordered regression (e.g., cumulative probit) models for ordinal outcomes. This improved representation of the underlying data directly addresses the goal of improving precision by allowing researchers to more accurately capture and communicate the true patterns in their data, rather than forcing complex ordinal distributions into ill-fitting linear frameworks.[3]

[2] Think of these as five subgroups of n = 250 observations per distributional group (total sample size of N = 1,250). This generates a total sample that is comparable to large-N survey research in criminology. Three distributions are simulated with identical population means (3.0) but, with sufficient sampling variability that observed sample means may differ slightly. The sample provides > 95% power to detect meaningful mean differences between the skewed distributions, while effectively demonstrating how distributions with similar means can exhibit dramatically different response patterns.

[3] We conducted additional simulations examining ordinal regression performance under some conditions that might be typical of real-world criminological data, including scenarios with polarized response distributions and zero-inflated crime outcomes. These simulations demonstrate how linear models can miss important heterogeneous treatment effects that ordinal models successfully detect, even when average treatment effects appear similar across modeling approaches. Full simulation details, code, and results are available in the online supplement (Supplemental Simulations 1A and 1B).

From Simulations to Real-World Data

Pickett et al.'s (2022) data on racial differences in fear of the police, published in *Criminology*, provide a compelling real-world parallel to the issues described above and depicted in Fig. 1. Pickett and colleagues conducted a nationally representative survey to investigate racial and ethnic differences in fear of the police, using a 10-item Likert-type scale to measure personal fear. Their findings revealed clear racial disparities, with Black Americans reporting higher levels of fear compared to White Americans.[4] Their survey data reveal distinctly different ordinal response distributions between Black and White Americans, which is precisely the kind of group-based distributional difference that challenges linear modeling assumptions and where ordinal methods should demonstrate clear advantages.

We selected Pickett et al.'s (2022) data because their study represents high-quality, impactful research addressing a critical social issue using methods common within the field, including the analysis of Likert-type ordinal items via linear regression and composite scales. The original authors usefully presented some item-specific distributions and kernel densities and effectively extracted important insights regarding racial disparities within the framework of these conventional techniques. Indeed, we believe their analysis derived more substantive understanding from the data using these methods than is sometimes achieved, particularly when interpretations focus primarily on coefficient signs and statistical significance.

In addition to its substantive importance, Pickett et al.'s (2022) study has other features that make it amenable to our purposes: the public availability of their data and code (facilitating open science practices), its use of widely employed linear modeling techniques, and its analysis of multi-item Likert-type ordinal data typical in criminological surveys. This makes it an ideal case for demonstrating the application and potential value of ordered regression models in a familiar and relevant context.

Ultimately, our goal here is illustrative and pedagogical. By applying ordered probit models to the same data, we aim to:

1. Demonstrate concretely how the outputs and interpretations from ordered regression differ from those of linear regression when applied to the same substantive question.
2. Showcase how ordered models can provide additional, potentially more precise, insights into distributional patterns and category-specific probabilities that linear models are less equipped to reveal.
3. Provide a practical example for criminologists considering the use of these techniques in their own work.
4. Illustrate the value of critically considering common aggregation/scaling practices for multiple ordinal items and assessing item-specific heterogeneity with multilevel models.

[4] Our central aim is to illustrate the utility of applying ordered regression modeling in this type of analysis, not to reproduce all models or reanalyze all outcomes reported in Pickett et al.'s (2022) original manuscript. Likewise, in some figures, Pickett and colleagues also reported results for a combined "other" race/ethnicity group, though that group's small subsample size generated much noisier estimates (i.e., very wide confidence bounds). We restrict our focus here to personal fear of police contrasts between Black and White participants. With that said, our analysis does replicate two specific estimates from their study (see Sect. 5.6): a bivariate race coefficient ($\beta = 32.09$; Model 1 in their supplementary Table S3) and corresponding standardized effect estimate ($d = 0.97$; 2022, p.302).

Application: Racial Differences in Fear of Police

Data

The survey data collected by YouGov in 2021 were obtained from: (1) a nationally representative sample (N = 700), and (2) an oversample of the Black population (N = 450), for a total sample size of N = 1,150. Both samples were collected using a two-step sample matching procedure meant to match participants to from YouGov's online panel to a synthetic sampling frame based on the American Community Survey to ensure representativeness.

Personal fear of police Pickett and colleagues' measure of personal fear was based on responses to ten items following the question stem:

> "*Now, we want to ask about EMOTIONAL FEAR. How afraid or unafraid are you that the POLICE will do the following things to you WITHOUT GOOD REASON in the next five years?"* (2022, p.320)

The ten items asked how afraid the participants were that police would: (1) *Stop you*; (2) *Search you*; (3) *Yell at you*; (4) *Handcuff you*; (5) *Punch or kick you*; (6) *Pin you to the ground*; (7) *Pepper spray you*; (8) *Use a taser on you*; (9) *Shoot at you with a gun*; and (10) *Kill you*.

Participants answered each item using a 5-point Likert type ordinal response scale. Following Pickett and colleagues, we recoded each of the 10 items so that higher values indicate greater fear as: 0 = *very unfraid*; 1 = *unafraid*; 2 = *neither afraid nor afraid*; 3 = *afraid*; 4 = *very afraid*.

Pickett et al. then "averaged the ten items to construct an index (loadings: 0.823 to 0.958, $\alpha = 0.98$) on which higher values indicated greater personal fear" (2022, p.300). This procedure creates a unidimensional composite "latent" fear of police index ranging from 0 to 4 (in increments of 0.1). Subsequently, they "normed" or re-scaled the measure to range from 0 to 100 so they could reportedly interpret coefficients as effects on a "percentage" scale (i.e., percentage of the total response distribution; 2022, p.302).

Analysis Plan

We focus on results from Bayesian linear and cumulative probit regression models. The posterior distributions that Bayesian models generate are advantageous for model convergence, model checking, and flexible visualization of results (Bürkner & Vuorre 2019). However, traditional frequentist approaches for applying ordinal regressions are widely available in popular statistical programs.[5] In fact, our simulation above and our R tutorial (Appendix 1) provide example applications of frequentist cumulative probit modeling (e.g., using the `polr` function from the MASS R package).

Bayesian models require specification of prior distributions. We use the `rstanarm` R package for linear models (Goodrich et al. 2020), which specifies diffuse prior distributions that are scaled by default to the outcome's response distribution. We use the `brms` R package (Bürkner 2017) for cumulative probit models and manually specify equal probability

[5] Some examples include: `polr()` from the MASS package in R; `ologit` or `oprobit` in Stata; `proc logistic` or `proc probit` in SAS; and `PLUM` in SPSS.

(flat) priors for ordinal models. These priors will allow for a wide range of likelihood-driven estimates and ensure results will be comparable to those that would be generated using similar frequentist techniques.[6]

Our analysis is organized around five primary tutorial goals[7]:

1. Illustrate the use of mosaic plots for visualizing bivariate ordinal response distributions, which better represent categorical frequencies than the continuous scatterplot.
2. Compare relative fit of intercept-only linear and cumulative probit regression models to observed ordinal response distributions.
3. Compare linear and cumulative probit regression results of item-specific analyses estimating race differences in fear of police on latent response scale (regression betas).
4. Extract race-specific predicted probabilities, calculate black-white contrasts, and then visualize predicted probabilities and contrasts from linear and cumulative probit models estimating race differences in item-specific fear of police response data.
5. Compare predicted probabilities representing latent fear of police from a cumulative probit multilevel model of all ten items, a linear multilevel model of all ten items, and a linear model of a normed mean scale (comparable to Pickett et al.'s analysis).

Together, these analyses will demonstrate how more appropriate modeling and visualization of ordinal data can reveal patterns and relationships that might be obscured or misrepresented by conventional linear approaches, thereby contributing to more precise and accurate criminological knowledge.

Ordinal Modeling and Visualization

Mosaic Plots

Goal: Illustrate the Use of Mosaic Plots for Visualizing Bivariate Ordinal Response Distributions, Which Better Represent Categorical Frequencies Than the Continuous Scatterplot

Before modeling, it is always a good idea to visualize one's data. After viewing the ten observed item distributions using frequency bar graphs, we visualize the bivariate relationship between race and each fear item.[8]

[6] Alternatively, one might specify stronger or "more informative" prior distributions that, for instance, specify a certain range of values as highly implausible or even as impossible; doing so would generate posterior estimates that deviate from likelihood-only estimates in predictable, prior-specified ways. For a simple example, a classical frequentist model of self-reported crime counts might generate impossible negative predicted counts for some subgroups and improbably extreme positive counts for other subgroups. In a Bayesian model, one could specify a prior distribution of plausible values for which negative counts are impossible and extreme positive counts are improbable, which would result in regularized posterior estimates with better out-of-sample predictive performance.

[7] Our primary online supplement contains all R code, results, and supplementary output for analyses presented in this paper, while Appendix 1 provides a general-purpose tutorial for implementing ordinal modeling techniques in R. These online supplements are available at the authors' website (https://reluctantcriminologists.com/supp/bd-supp/) and the project's OSF page (https://osf.io/n52kc/).

[8] Frequency bar graphs for all ten items can be found in the online supplement.

Unfortunately, scatterplots are notoriously bad at generating useful visual descriptions of bivariate relationships between ordinal or categorical variables. In contrast, mosaic plots are designed for this task. Will Ball (2020) describes a mosaic plot as "similar to a heatmap, except that the frequency of an observation… is proportional to the area of a tile." For example, the width of each bar in Fig. 2 represents the proportion of the sample within that racial group; this strength is not immediately apparent in this sample since the bars are approximately equal due to Pickett and colleagues' oversampling approach. Meanwhile, similar to a proportional stacked bar chart, the height or "thickness" of each tile in Fig. 2 represents the proportion of responses within each fear response category for that racial group, with thicker (or thinner) tiles indicating a larger (or smaller) proportion of participants reporting in that fear response category.

These mosaic plots help us clearly see substantial race differences exist in the ordinal response distribution for each fear of police item. For those that prefer numbers, we can also summarize these race-specific proportions in a table. However, including all estimates would be a lot of information. For simplicity, we summarize the extreme categories of *very unafraid* ($y=0$) and *very afraid* ($y=4$) in Appendix Table 2.

Across all ten police actions listed, Black participants consistently reported a lower percentage of "very unafraid" responses and a higher percentage of "very afraid" responses compared to White participants. The more physically invasive or dangerous actions, such as *kick/hit you*, *pin you down*, *pepper spray you*, *tase you*, *shoot at you*, and *kill you*, all show the largest disparities in perceived safety. For example, between 44.1% and 44.9% of White participants reported being "very unafraid" of these items compared to between 10.3% and 11.4% of Black participants who felt the same way, resulting in more than a 30-percentage point difference for each item. Overall, these plots effectively illustrate the heightened sense of fear among Black participants regarding potentially violent police interactions.

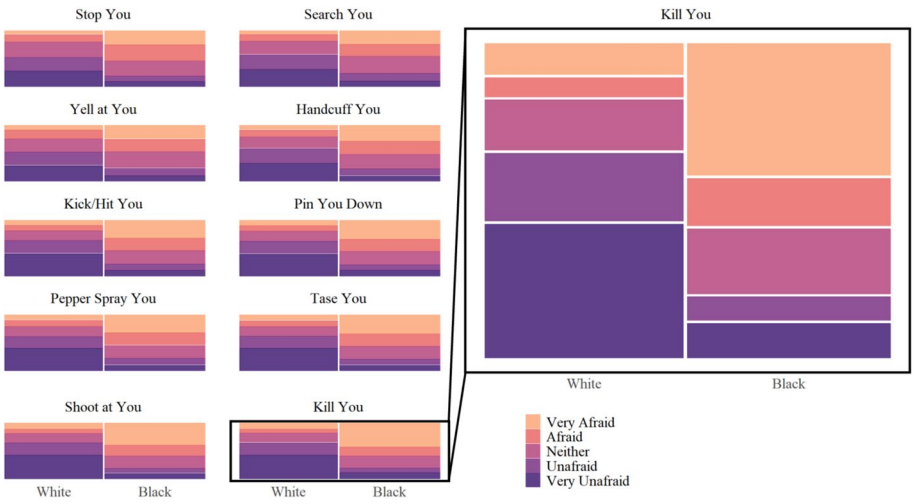

Fig. 2 Mosaic Plots of "Fear of Police" Ordinal Item Response Distributions by Race. Note: $N=504$ White and $N=516$ Black participants from Pickett et al.'s (2022) data. Higher values indicate greater fear. The width of each bar represents the proportion of the sample within that race. The height of each tile represents the proportion of responses within each category for that race

While the mosaic plots and Appendix 2 provide valuable descriptive information, quantifying observed differences and demonstrating the magnitude of those differences, we typically need statistical modeling to go further. Specifically, we often want to estimate the degree of uncertainty surrounding our inferences, permit extensions such as adjustment for potential confounders, and allow for formal hypothesis testing if desired. Therefore, the usual next step would be to estimate these joint distributions with a regression model.

Comparing Fit of Item-Specific Intercept-Only Models

Goal: Compare Relative Fit of Intercept-Only Linear and Cumulative Probit Regression Models to Observed Ordinal Response Distributions

We begin regression modeling with a comparison of the Bayesian equivalent of linear and ordinal simple means or intercept-only models for each of the ten fear items. As implied by the name, each intercept-only model unconditionally estimates only the intercepts (e.g., mean values or ordinal thresholds) for each fear item without including race as a predictor. The intercept-only linear model generates a single parameter estimate intended to recover the mean of the latent continuous response variable, which in this case is equivalent to the mean of each ordinal fear item. While the mean of a 5-category Likert-type item can provide a general sense of central tendency, it fails to completely capture distributional characteristics of ordinal data, such as skewness and multimodality. Different ordinal distributions can have the same mean, making it difficult to draw meaningful conclusions from linear models alone (e.g., see Fig. 1 above).

Equation 3. (Intercept-only Bayesian linear model)

Where:

- y_i is the observed value of the ordinal item for individual *i*.
- $\text{Normal}(\mu,\ \sigma)$ indicates that y_i is assumed to be normally distributed with mean μ and standard deviation σ.
- μ is the intercept, representing the mean of the latent continuous variable.
- $\mu \sim \text{Normal}(y_{mean,}\ 2.5\ ^*\ y_{sd})$ is the prior distribution for the intercept μ, assumed to be normally distributed with mean y_{mean} (the sample mean of y) and standard deviation $2.5\ ^*\ y_{sd}$ (2.5 times the sample standard deviation of y). This is rstanarm's data-dependent default scaling.
- σ is the standard deviation of the normal distribution.
- $\sigma \sim \text{Exponential}(1\ /\ y_{sd})$ is the prior distribution for the standard deviation σ, assumed to be exponentially distributed with rate $1\ /\ y_{sd}$ (the inverse of the sample standard deviation of y).

By contrast, the intercept-only cumulative probit model generates multiple threshold parameter estimates that divide the latent space into ordered categories, each representing the space between observed response values. Thus, the model generates *j*–1 thresholds for an ordinal item with *j* response categories, or four thresholds for a 5-category Likert-type item.

Unlike the linear mean, each ordered threshold is fundamentally interpretable upon transformation to the probability scale as the (unconditional) cumulative proportion of responses in the preceding response categories. Using these cumulative threshold estimates, we can then calculate the specific (predicted with uncertainty)[9] proportion of responses in each ordinal response category.

Equation 4. (Intercept-Only Bayesian Cumulative Probit Model)

$$y*_i \sim \text{Normal}(\mu, 1)$$
$$y_i = j \text{ if } \tau_{j-1} < y*_i \leq \tau_j$$
$$\tau_1 \sim \text{Normal}(-0.84,\ 1)$$
$$\tau_2 \sim\ \text{Normal}(-0.25,\ 1)$$
$$\tau_3 \sim\ \text{Normal}(0.25,\ 1)$$
$$\tau_4 \sim\ \text{Normal}(0.84,\ 1)$$

Where:

- $y*_i$is the latent continuous variable for individual *i*.
- $\text{Normal}(\mu, 1)$ indicates that $y*_i$ is assumed to be normally distributed with mean μ and standard deviation 1 (the standard deviation is fixed to 1 for identification purposes in ordered probit models).
- y_i is the observed ordinal category for individual *i*.
- $y_i = j$ if $\tau_{j-1} < y*_i \leq \tau_j$ defines how the latent variable $y*_i$ is mapped to the observed ordinal category y_i using thresholds τ_j.
- μ is the intercept, representing the mean of the latent continuous variable.
- $\tau_1 \sim \text{Normal}(-0.84,\ 1)$, $\tau_2 \sim\ \text{Normal}(-0.25,\ 1)$, $\tau_3 \sim\ \text{Normal}(0.25,\ 1)$, and $\tau_4 \sim\ \text{Normal}(0.84,\ 1)$ are priors representing expected thresholds for equal category probabilities

For each of the ten items, the linear (metric normal) model intercept estimate accurately recovers the unconditional mean (e.g., $M_{\text{stopyou}} = 1.99$), which is what the model is designed to do. However, posterior predictive density plots in Fig. 3 show that the underlying normal distribution assumption results in linear model predictions (dark purple normal curves) that fail to recover the observed ordinal item values (light orange ordinal density curves).

In contrast, Fig. 3 shows the intercept-only cumulative probit model predictions (dark purple point-intervals) are a much better fit to the data (light orange bars) than are the comparable intercept-only linear models. This is unsurprising, as additional thresholds parameters are estimated in the cumulative probit model to more completely summarize each item's ordinal distribution.

[9] With classical frequentist modeling techniques, these intercept or threshold estimates would be computed directly from the data or likelihood. In a Bayesian model, these estimates represent the center of a posterior sample of estimates that each are generated by combining the likelihood (e.g., observed mean or ordinal thresholds in the data) with a prior distribution of plausible values for the parameter in question. As explained earlier, we specified diffuse or flat priors so posterior point and uncertainty estimates would converge with frequentist model estimates, ensuring that our Bayesian results are comparable to those obtained using traditional frequentist methods (see Table 3).

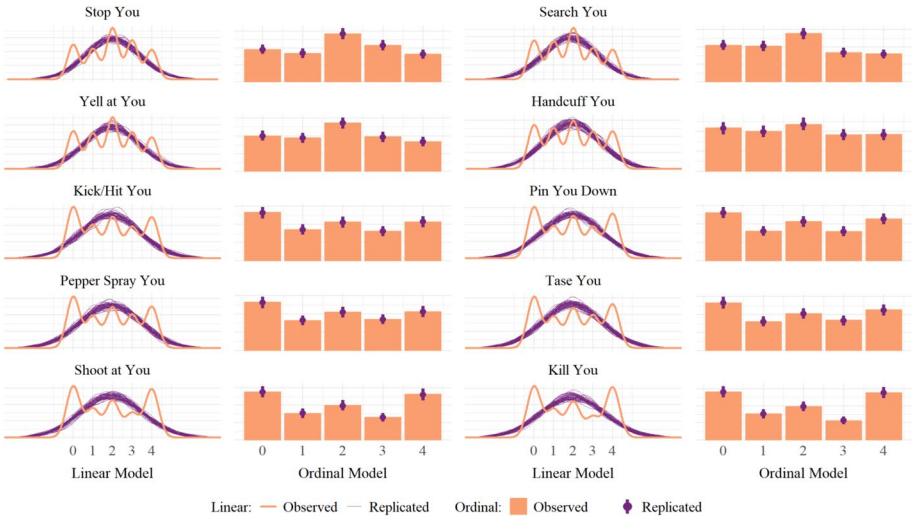

Fig. 3 Posterior Predictive Checks for Linear vs. Ordinal Models. Note: Pairs show linear model density checks (left) and ordinal model bar checks (right).

We can quantify and compare the relative fit of the intercept-only linear regression and cumulative probit models for each fear of police item using Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO-CV; see Vehtari et al. 2017) and calculating expected log predictive density (ELPD) for each model. The ELPD provides a measure of (out-of-sample) predictive accuracy, with higher values indicating better predictive performance and, by comparing its predictive performance to unseen data, it penalizes models for extra parameters. PSIS-LOO-CV provides a rigorous method for comparing the predictive accuracy of different models, helping us to select the model that best generalizes to unseen data.

Table 1 reports PSIS-LOO-CV results to compare model fits with ELPD. The "Difference" column shows the difference in ELPD between the ordinal and linear models (ordinal minus linear), with positive values favoring the ordinal model. The standard error (SE) quantifies our uncertainty in these differences.

These tables confirm what we observed in the posterior predictive checks (Fig. 3), where ordinal models better captured the discrete nature of the response categories compared to the continuous approximation used in linear models. The ordinal models showed consistently better predictive performance across all items, with ELPD differences ranging from about 115 to 319 units. The improvements were particularly substantial for more severe items like *shoot at you* (Diff ≈ 289) and *kill you* (Diff ≈ 317), items which exhibited more extreme bimodal distributions. Generally, the performance of modeling ordinal variables using continuous Gaussian approximation will be better for variables with mean-centered distributions and several response categories that exhibit symmetric decay about the mean (e.g. "Normal-like" distribution in Fig. 1).

The "Diff/SE" ratio provides an approximation of statistical significance that is similar to a *z*-test. Ratios that exceed approximately 2 indicate statistically significant differences at the $p < 0.05$ level. In this case, all differences are well over 10 standard errors from zero (around 12–20), indicating superior fit of the ordinal models.

Table 1 Comparing relative fit of linear and ordinal intercept-only models

Outcome	Linear ELPD	Ordinal ELPD	Difference	SE	Diff/SE
Stop You	−1,738.41	−1,623.50	114.90	11.42	10.06
Search You	−1,748.40	−1,625.31	123.09	11.83	10.41
Yell at You	−1,751.51	−1,630.70	120.81	11.65	10.37
Handcuff You	−1,788.46	−1,640.93	147.53	13.01	11.34
Kick/Hit You	−1,847.02	−1,629.02	218.00	16.57	13.15
Pin You Down	−1,856.14	−1,626.34	229.81	17.09	13.45
Pepper Spray You	−1,848.66	−1,630.11	218.55	16.55	13.20
Tase You	−1,858.72	−1,627.75	230.97	17.10	13.50
Shoot at You	−1,893.01	−1,603.83	289.18	19.50	14.83
Kill You	−1,906.10	−1,589.35	316.74	20.42	15.51

Table reports expected log predictive density (ELPD) values calculated from Pareto-smoothed importance sampling leave-one-out cross-validation (PSIS-LOO-CV) to assess relative fit of intercept-only linear regression and cumulative probit models for each fear of police item (see also Fig. 3). The ELPD provides a measure of predictive accuracy, with higher values indicating better predictive performance. The "Diff" column shows the difference in ELPD between the ordinal and linear models (ordinal minus linear), with positive values favoring the ordinal model. The standard error (SE) quantifies uncertainty in these differences. For statistical testing with LOO-CV comparisons, differences are typically considered meaningful when they exceed twice their standard error (i.e., similar to a *z*-test, when "Diff/SE" > 1.96)

Now that we have established the poor fit of intercept-only linear models to ordinal outcome items and the substantial improvements we gain in predictive fit to ordinal outcome distributions by using ordered (e.g., cumulative probit) modeling techniques, we are ready to add race as a predictor in our models.

Comparing Estimated Item-Specific Race Effects on Latent Response Scale

Goal: Compare Linear and Cumulative Probit Regression Results of Item-Specific Analyses Estimating Race Differences in Fear of Police on Latent Response Scale (Regression Betas)

Our next step involved adding "Black" race as a predictor to each item-specific linear or cumulative probit model. The equation for the linear model in Eq. 3 now becomes:

Equation 5. (Bayesian linear model with race predictor)

$$y_i \sim \text{Normal}(\mu_i, \sigma)$$
$$\mu_i = \alpha + \beta * \text{Black}_i$$
$$\alpha \sim \text{Normal}(y_{mean,}\ 2.5 * y_{sd})$$
$$\beta \sim \text{Normal}(0, 2.5 * y_{sd} / x_{sd})$$
$$\sigma \sim \text{Exponential}(1 / y_{sd})$$

The equation for the cumulative probit model in Eq. 4 changes to:

Equation 6. (Bayesian cumulative probit model with race predictor)

$$
\begin{aligned}
&y*_i \text{ i} \sim \text{Normal}(\mu_i, 1)\\
&\mu_i = \alpha + \beta * \text{Black}_i\\
&y_\text{i} = j \text{ if } \tau_{j-1} < y*_i \text{ i} \leq \tau_j\\
&\tau_1 \sim \text{Normal}(-0.84, 1)\\
&\tau_2 \sim \text{Normal}(-0.25, 1)\\
&\tau_3 \sim \text{Normal}(0.25, 1)\\
&\tau_4 \sim \text{Normal}(0.84, 1)\\
&\beta \sim \text{Normal}(0, 1)
\end{aligned}
$$

In both models, the beta coefficient (β) represents the change in the latent variable for a one-unit increase in the "Black" predictor (i.e., comparing Black participants to White participants). For the linear model, β represents the difference in means between the two groups on the unbounded latent metric scale. As before, we will use the rstanarm package's default data-dependent scaling priors for the linear model, where y_{mean} is the mean of the response variable, y_{sd} is the standard deviation of the response variable, and x_{sd} is the standard deviation of the predictor (approximately 0.5 for a binary predictor). For the ordered probit model, β represents the difference in means on the latent scale in standard deviation units, which is constrained by the thresholds that map it to the observed ordinal categories. Again, we specify a flat prior distribution over the thresholds and a diffuse Normal(0, 1) prior for β coefficients.

After estimating the models, we extracted the point estimate (regression beta coefficient) and credible interval (95% CrI) representing the model-predicted effect of Black race on the outcome item's latent response scale and plotted them in Fig. 4.

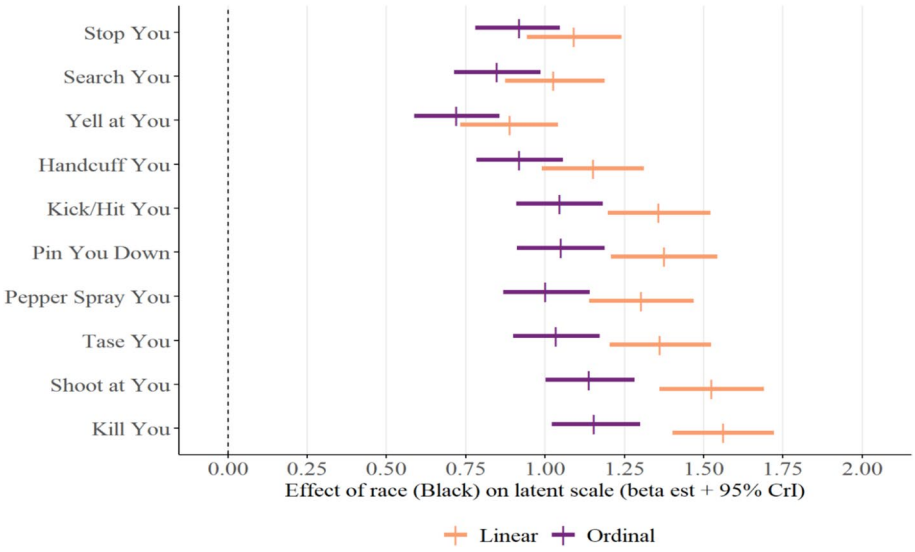

Fig. 4 Estimated Effects of Race (Black) on Fear of Police Responses, Item-Specific Latent Response Scale. Note: Plot shows regression beta point estimates and 95% credible intervals for the effect of Black race on the latent response scale of each fear of police item, comparing linear and cumulative probit models. The plot displays the beta coefficients from both models, highlighting differences in the estimated effects on the latent scales.

Initially, Fig. 4 appears to show that the linear model estimates larger effects of race on the latent response scale for all 10 items compared to the cumulative probit model. Yet, the linear and cumulative probit models assume very different scales for each outcome, making direct comparisons of beta coefficients challenging. Recall that the linear model is assuming an unbounded continuous outcome, so *larger* effects could be relative to a much larger set of model-implied outcome values (i.e., with "plausible" predictions extending below $y=0$ and above $y=4$). Meanwhile, the cumulative probit model accurately assumes a discrete distribution with y-values bounded between 0 and 4, so the smaller cumulative probit (CP) estimates on the latent y scale are interpreted relative to a more narrowly restricted range of y values.

In short, though the CP beta coefficient is intended to be somewhat comparable to the beta from a linear model, where a one-unit increase in X (or Black vs. White contrast) is associated with a beta-unit increase in y, it is difficult to precisely interpret differences in results across these models. However, though the linear coefficient is largely uninterpretable with ordinal outcomes (i.e., as a difference in means on a response variable for which calculating the mean is inappropriate), the same is not true for the CP coefficient. Since it is on the latent standardized normal scale, the CP coefficient can be interpreted as a latent *Cohen's d*, representing the difference in latent means between groups in standard deviation units. For example, Fig. 4 shows the point estimate for *pepper spray you* at $\beta=1.0$, indicating that Black participants on average have a fear of being pepper sprayed by police that is one standard deviation higher than that of White participants. Compared to popular rule-of-thumb thresholds for interpreting *Cohen's d* effect magnitudes, these are extremely large effects by social science standards. This underscores the substantive significance of these racial disparities and highlights another serious consequence regarding analytical precision: In other research contexts, scholars using linear models to analyze ordinal outcomes with similar underlying distributional shapes might fail to detect statistically significant effects that are smaller in magnitude but practically important, even with relatively large samples (false negative).

With that said, relying solely on beta coefficients can lead to misleading conclusions. It is understandable that researchers might observe estimates from different modeling specifications that are in the same direction, implying similar statistical inferential conclusions (e.g., all 95% CrIs exclude zero), and apparently somewhat similar magnitudes, and then interpret both modeling techniques as generating "similar substantive conclusions." However, to achieve more precise scientific conclusions, it is essential to move beyond latent scales and examine predicted probabilities.

Extracting and Visualizing Predicted Probabilities and Contrasts

Goal: Extract Race-Specific Predicted Probabilities, Calculate Black-White Contrasts, and Visualize Predicted Probabilities and Contrasts From Linear and Cumulative Probit Models Estimating Race Differences in Item-Specific Fear of Police Response Data

As we discussed in the previous section, directly comparing beta coefficients from linear and ordered probit models can be misleading due to differences in the latent scales. To gain a clearer understanding of the impact of race on fear of police, we will now focus on extracting and visualizing predicted probabilities and contrasts.

Specifically, we will focus on the predicted probability of reporting being "afraid" or "very afraid" ($y >= 3$). This aligns with the implied estimand of the exemplar study, which aimed to estimate group-based differences in "fear" of police. Understanding fear of police among different demographic groups is crucial for addressing issues of trust and legitimacy in criminological research, and reporting precise differences in fear responses is important for accurately assessing the magnitude of disparities and understanding the true impact of policing practices on community perceptions.

Linear Model Predicted Probabilities and Contrasts To make fair comparisons between models, we calculated the predicted probability of responding "afraid" or "very afraid" ($y >= 3$) for each racial group from both linear and ordered probit models. For linear models, this required transforming the continuous normal predictions into probabilities of exceeding the threshold value of 3 on the 0–4 scale.[10] We then calculated Black-White contrasts by taking the difference between these predicted probabilities.

To quantify the difference in predicted probabilities between Black and White participants, we calculate the Black-White contrast as: Contrast = P($y >= 3$ | Black) – P($y >= 3$ | White). We also calculate the posterior distribution of the contrast and summarize it to obtain a point estimate and credible interval.

Ordered Probit Model Predicted Probabilities and Contrasts For the ordered probit models, we use the posterior distribution of the model parameters to directly calculate the predicted probabilities of $y >= 3$ for Black and White participants. We then calculate the Black-White contrast in the same way as for the linear models.

Visualization Figure 5 presents a comprehensive visualization of the predicted probabilities and Black-White contrasts for the "afraid" or "very afraid" responses ($y >= 3$) across the ten fear of police items. The plots are structured to allow for direct comparisons of model predictions with observed data, as well as comparisons between the linear and ordered probit models.

Race-Specific Predicted Probabilities (Left Panel) The left panel displays the race-specific predicted probabilities for both linear (orange point-intervals) and ordered probit (purple point-intervals) models, alongside the observed proportions (diamonds). Each point-interval represents the posterior mean and 95% credible interval for the predicted probability of $y >= 3$ for either White or Black participants, as estimated by the respective model. The observed proportions, shown as white or black diamonds, represent the raw proportions of participants in each racial group who reported $y >= 3$.

Black-White Contrasts (Right Panel) The right panel of Fig. 5 depicts the Black-White contrasts, which are the differences in predicted probabilities between Black and White participants, for both linear and ordered probit models. The observed contrast (black diamond) is

[10] Since our Bayesian linear models generate a full posterior distribution, we used the model's normality assumption to calculate P($y >= 3$) = 1—*pnorm*(3, mean = μ_i, sd = σ), where μ_i is the mean of the latent variable for individual i and σ is the standard deviation. This allowed us to estimate the conditional predicted probabilities of responding *afraid* or *very afraid* for each group and the uncertainty around those estimates. Technical details and code are available in our online supplement.

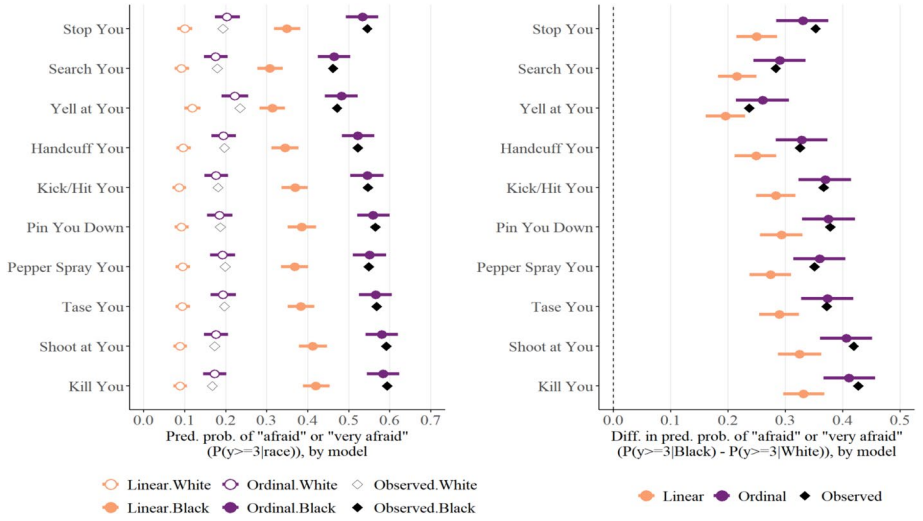

Fig. 5 Race-Specific Probabilities of "Afraid" or "Very Afraid" Responses and Black-White Probability Contrasts, Observed and from Linear or Ordinal Model Predictions. Note: Left plot shows race-specific observed and model-predicted probabilities of reporting being "afraid" or "very afraid" (y=4) of each police outcome. Right plot displays contrasts (Black – White) in observed or model-predicted probabilities for each outcome.

also included. Each point-interval represents the posterior mean and 95% credible interval for the contrast. The dashed vertical line at $X=0$ serves as a reference point, indicating no difference between the two groups. Points to the right of this line suggest a higher probability of $y>=3$ for Black participants, while points to the left indicate a higher probability for White participants.

Key Observations *First, as expected, the CP model's point-intervals accurately reflect observed data, confirming its appropriate fit to ordinal outcome data. In stark contrast, while a naive interpretation of beta estimates might have suggested larger race effects in the linear model (see Fig. 4), converting to predicted probabilities reveals this to be a misleading conclusion. Linear model predictions consistently underestimate observed values, with no interval successfully capturing the observed race-specific proportions or Black-White contrasts for any outcome item.*

Specifically, the linear model's point estimates, on average, underestimated the item-specific probabilities of "afraid" or "very afraid" responses by 0.14, with underestimates ranging from 0.08 to 0.20 across outcomes. Similarly, the linear model underestimated the item-specific Black-White contrasts by an average of 0.08, with underestimates ranging from 0.042 to 0.1. This pattern aligns with the ceiling and floor effects described by Kruschke and Liddell (2018), where linear models struggle to represent data clustered at extreme values due to their inherent assumption of normality. Overall, these plots clearly demonstrate the different ways the two models represent the observed data, highlighting the closer alignment of the CP model's predicted probabilities with the empirical category proportions in this specific case.

Estimating Latent Fear With Multilevel Models or Normed Scale

Goal: Compare Predicted Probabilities Representing Latent Fear of Police from a Cumulative Probit Multilevel Model of All Ten Items, a Linear Multilevel Model of All Ten Items, and A Linear Model of a Normed Mean Scale (Comparable to Pickett et al.'s Analysis)

Is a Composite "Latent" Fear Measure Appropriate? Recall, Pickett and colleagues (2022) constructed a composite scale by averaging ten ordinal items, then rescaled the measure to range from 0 to 100, and interpreted coefficients from linear regression models predicting this normed scale as effects on a "percentage" (of the outcome) scale. As noted earlier, this appears to be relatively common practice in criminology. However, this approach invalidly assumes equal intervals between response categories, and rescaling to 0–100 implies metric properties that do not exist in the original ordinal data. This approach also loses item-specific information and assumes the appropriateness of modeling a unidimensional "latent" fear construct. Alternatives include cumulative probit models that respect the ordinal nature of the items and multilevel modeling to account for item clustering. A cumulative probit multilevel model should fit the ordinal data better and allow us to simultaneously examine whether item-specific predicted probabilities and contrasts indeed cluster across the ten items as we would expect if each were an indicator of a unidimensional latent fear construct.

Cumulative Probit Multilevel Model (CP MLM) The equation for the cumulative probit multilevel model is:

Equation 7. (Bayesian cumulative probit multilevel model with race predictor)

$$
\begin{aligned}
& y*_{ijk} \sim \text{Normal}(\mu_{ijk},\ 1) \\
& \mu_{ijk} = \alpha_{\ jk} + \beta * \text{Black}_i + u_{0jk} + u_{1jk} * \text{Black}_i \\
& \mu_{0jk},\ \mu_{1jk} \sim \text{MVNormal}(0,\ \textstyle\sum_{\ j}) \\
& \text{y}_{ijk} = \text{k if } \tau_{k-1} < \text{y}*_{ijk} \leq \tau_k \\
& \tau_1 \sim \text{Normal}(-0.84,\ 1) \\
& \tau_2 \sim \text{Normal}(-0.25,\ 1) \\
& \tau_3 \sim \text{Normal}(0.25,\ 1) \\
& \tau_4 \sim \text{Normal}(0.84,\ 1) \\
& \beta \ \sim \text{Normal}(0,\ 1)
\end{aligned}
$$

Where:

- $y*_{ijk}$ is the latent continuous variable for individual *i*, item *j*, and threshold *k*.
- μ_{ijk} is the mean of $y*_{ijk}$.
- $\alpha_{\ jk}$ is the item-specific intercept for threshold *k*.
- β is the population-level effect of Black_i.
- Black_i is a binary indicator for race (Black = 1, White = 0).
- μ_{0jk} and μ_{1jk} are item-specific random intercepts and slopes, respectively.
- MVNormal indicates a multivariate normal distribution.
- Σ*j* is the covariance matrix for the random effects of item *j*.
- y_{ijk} is the observed ordinal response, where *k* represents the ordered category (0 to 4).
- τ_k are the threshold parameters.

Visualizing Item-Specific and Multilevel Model Estimates Figure 6 presents a comparison of predicted probabilities and Black-White contrasts from our ten item-specific CP models (semi-transparent point-intervals) with comparable estimates from the CP MLM (violin plots). The MLM provides pooled estimates of the effect of race on each response category on a latent fear scale; overlaying those estimates on all ten item-specific estimates helps us to see how these MLM estimates are generated and to visually determine whether the ten items appear to assess a common latent construct.

Figure 6 reveals that, overall, item-specific models generate probability and contrast estimates that closely cluster across the ordinal response distribution. This clustering supports the notion that the individual items measure a common "latent" fear of police construct. However, notable exceptions emerge, particularly in the extreme response categories (e.g., "very unafraid" for Whites and "very afraid" for Blacks), where substantial heterogeneity across items is observed. Multilevel model estimates are precise where underlying item-specific predictions closely cluster, and uncertainty intervals are wider where there is more variation in item-specific predictions. This pattern underscores the model's ability to capture both commonalities and variations across items.

Linear Model with Normed Mean Scale To more directly compare with Pickett et al.'s approach, we averaged the ten fear items and rescaled the resulting composite to a 0–100 scale. A linear model with race as a predictor was fit to this normed scale.

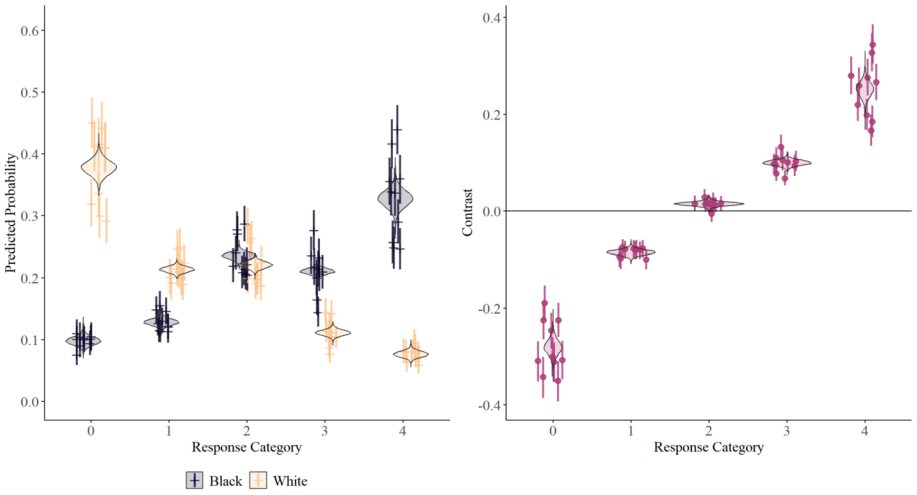

Fig. 6 Race-specific predicted probabilities of latent fear of police for all ordinal response categories and black-white probability contrasts, from item-specific and multilevel ordinal models. Note: Left plot shows race-specific predicted probabilities of reporting each ordinal response category (0 = *very unafraid* to 4 = *very afraid*) from ten item-specific ordered probit models (semi-transparent point-intervals) and one multilevel ordered probit model predicting latent fear of police (violin plot). Right plot displays contrasts (Black – White) in predicted probabilities for each response category from the ten item-specific models (semi-transparent point-intervals) and the multilevel model (violin plots)

Equation 8. (Linear model with race predictor, normed *y* scale)

$$y_i \sim \text{Normal}(\mu_i,\ \sigma)$$
$$\mu_i = \alpha + \beta * \text{Black}_i$$
$$\alpha \sim \text{Normal}(y_{mean},\ 2.5 * y_{sd})$$
$$\beta \sim \text{Normal}(0, 2.5 * y_{sd} / x_{sd})$$
$$\sigma \sim \text{Exponential}(1 / y_{sd})$$

Where:

- y_i is the normed mean scale value for individual *i*.
- μ_i is the mean of y_i.
- α is the intercept.
- β is the population-level effect of Black_i.
- Black_i is a binary indicator for race (Black = 1, White = 0).
- $\alpha \sim \text{Normal}(y_{mean},\ 2.5 * y_{sd})$ is the prior distribution for the intercept α.
- $\beta \sim \text{Normal}(0, 2.5 * y_{sd} / x_{sd})$ is the prior distribution for the regression coefficient β.
- α is the standard deviation of the error term.
- $\sigma \sim \text{Exponential}(1 / y_{sd})$ is the prior distribution for the standard deviation σ.

As expected, this model revealed a substantial difference in reported fear between Black and White participants. The coefficient for race (Black) was β = 31.65 (95% CI: 28.19, 35.28), indicating that Black participants scored, on average, 31.65 points higher than White participants on the 0–100 normed scale, where White participants scored a mean of 32.17. While this 31.65-point difference on the 0–100 scale indicates a substantial disparity, translating this linear coefficient back into the context of the original ordinal responses requires care. As noted previously, interpreting this as a direct "percentage point" difference rests on assumptions about equal intervals and the nature of the scale transformation that differ from the assumptions of ordinal models. Comparing the insights from this approach with those from the ordinal models (below) highlights how different methodological assumptions can lead to different quantitative conclusions about the magnitude of disparities.

To better understand these results, we back-transformed the predicted means to the original 0–4 scale. For White participants, the predicted mean was 1.29, and for Black participants, it was 2.55, for a predicted difference of 1.26 units on the original scale. Importantly, 31.65% of a theoretically possible four-threshold increase (i.e., from 0 to 4) on a latent ordinal fear response also equals 1.266. It should be clear by now that the "percentage change" summarized by linear model coefficients using a normalized (0 to 100) ordinal composite scale is neither a true nor meaningful percentage metric; it provides the illusion of a metric interpretation while obscuring the true interpretation as an inappropriate and potentially biased metric-approximation of predicted change in raw ordinal units. On the original scale (0 = *very unafraid*, 4 = *very afraid*), these conditional means imply that the typical White participant falls between "unafraid" and "neither," while the typical Black participant falls between "neither" and "afraid." How accurate are these predictions? Fig. 5 (above) revealed that over half of Black participants reported being "afraid" or "very afraid" on eight out of ten items; the CP MLM also predicts just over half of Black participants feel "afraid" or "very afraid" (see categories '3' and '4' in Fig. 6). As with the item-specific results, a linear model fit to a mean-scaled and normed composite measure comprised of ordinal outcome items substantially underestimates the predicted race effect on latent fear.

Linear Multilevel Model To enable a direct comparison with the cumulative probit multilevel model (CP MLM) using leave-one-out cross-validation (LOO), we also fit a linear multilevel model (linear MLM). This was necessary because we cannot use LOO to directly compare the CP MLM with the linear model using the normed mean scale.

Equation 9. (Bayesian linear multilevel model with race predictor)

$$
\begin{aligned}
y_{ij} &\sim \text{Normal}(\mu_{ij}, \sigma) \\
\mu_{ij} &= \alpha + \beta * \text{Black}_i + \mu_j \\
\mu_j &\sim \text{Normal}(0,\ \sigma_j) \\
\alpha &\sim \text{Normal}(0,\ 1) \\
\beta &\sim \text{Normal}(0,\ 1) \\
\sigma &\sim \text{Exponential}(1) \\
\sigma_j &\sim \text{Exponential}(1)
\end{aligned}
$$

Where:

- y_{ij} is the fear value for individual *i* and item *j*.
- μ_{ij} is the mean of y_{ij}.
- α is the global intercept.
- β is the population-level effect of Black_i.
- Black_i is a binary indicator for race (Black = 1, White = 0).
- μ_j is the random intercept for item *j*.
- $\mu_j \sim \text{Normal}(0,\ \sigma_j)$ is the prior distribution for the random intercept.
- σ is the standard deviation of the error term.
- Normal(0, 1) is a diffuse prior distribution for the global intercept α and regression coefficient β.
- Exponential(1) is the prior distribution for the standard deviation σ and the standard deviation of the random intercepts σ_j.

Comparing MLM and Linear Latent Model Predictions for "Very Afraid" To compare each of the three models' predictions of "very afraid" ($y = 4$), we calculated the probabilities and contrasts implied by both the mean-scaled normed linear model and the multilevel linear model using two thresholds: 100 (precise) and 87.5 (generous, corresponding to $y >= 3.5$). This is necessary because the metric normal distributional assumption of the linear models permits ambiguity (and potential researcher degrees of freedom) in defining the precise cutoff on the latent scale that is nominally equivalent to "very afraid." In contrast, the CP MLM model is explicitly designed to generate precise estimates for each ordinal response category.

Table 2 presents a comprehensive comparison of predicted probabilities and contrasts for "very afraid" responses across all models. "Observed" proportions of latent "very afraid" responses were also included by calculating an unweighted average of the race-specific proportions across all ten items.

Key Observations *Examining Table 2 reveals that the ordinal MLM's predicted probabilities and contrasts closely align with the observed average proportions, with the observed values falling well within the model's 95% credible intervals. In comparison, the estimates derived from the linear models (both normed scale and multilevel, using either precise or*

Table 2 Race-specific predicted probabilities of "Very Afraid" responses and contrasts from ordinal MLM and linear scale or MLM models predicting latent fear of police

		95% CrI	
	Estimate	**Lower**	**Upper**
Probability: P("Very Afraid"|Black)			
Observed Proportions	0.321	–	–
Ordinal MLM (y=4)	0.328	0.300	0.356
Linear Scale: Precise (normed y>=100)	0.110	0.092	0.131
Linear MLM: Precise (y>=4)	0.133	0.126	0.140
Linear Scale: Generous (normed y>=87.5)	0.211	0.185	0.239
Linear MLM: Generous (y>=3.5)	0.233	0.224	0.243
Probability: P("Very Afraid"|White)			
Observed Proportions	0.086	–	–
Ordinal MLM (y=4)	0.076	0.067	0.087
Linear Scale: Precise (normed y>=100)	0.011	0.007	0.015
Linear MLM: Precise (y>=4)	0.019	0.017	0.021
Linear Scale: Generous (normed y>=87.5)	0.030	0.023	0.040
Linear MLM: Generous (y>=3.5)	0.045	0.042	0.049
Contrast: P("Very Afraid"|Black)—P("Very Afraid"|White)			
Observed Contrast	0.235	–	–
Ordinal MLM (y=4)	0.251	0.218	0.284
Linear Scale: Precise (normed y>=100)	0.099	0.082	0.119
Linear MLM: Precise (y>=4)	0.114	0.108	0.121
Linear Scale: Generous (normed y>=87.5)	0.181	0.156	0.208
Linear MLM: Generous (y>=3.5)	0.188	0.179	0.198

Observed proportions of "very afraid" (y=4) responses calculated as the mean of race-specific observed proportions across all ten items; observed contrast is the Black–White difference in these proportions. All other estimates represent predicted probabilities and contrasts for the "very afraid" (y=4) response category on a latent "fear of police" construct. "Ordinal MLM" estimates are from a Bayesian ordered probit (ordinal) multilevel model with thresholds and race effects varying across all ten outcomes, calculated as population-level predictions marginalized over the random effects of outcomes; interval estimates incorporate uncertainty from both fixed and random effects in the predictions (see also response category=4 in Fig. 5). "Linear Scale" estimates are from a Bayesian linear model predicting a composite latent fear scale calculated as the normed mean of all ten outcomes (see Pickett et al. 2022). Linear MLM are estimates from a Bayesian linear multilevel model with otherwise comparable specifications to the ordinal MLM. Since the linear models do not directly predict ordinal categories but assumes underlying latent metric responses, two alternative estimates are calculated using different latent response thresholds: Precise: "y=4 or above" (>=100 on normed scale); Generous: "y=3.5 or above" (>=87.5 on normed scale).

generous thresholds) show notable differences from the observed values, with the observed parameters falling outside the 95% credible intervals from these models. For example, the ordinal MLM predicted a probability of Black participants being very afraid of police to be 0.328, while the linear normed model (precise) predicted 0.110; the observed proportion was 0.321. The ordinal MLM predicted a probability of white participants being very afraid of police to be 0.076, while the linear normed model (precise) predicted 0.011; the observed proportion was 0.086. These comparisons highlight the ordinal MLM's greater accuracy in recovering the observed category proportions for "very afraid" in this dataset.

LOO Comparison of Linear MLM and CP MLM The LOO comparison revealed a dramatic difference in model fit (see Appendix 3), with the linear MLM exhibiting a substantially worse ELPD (−17,182.0) compared to the CP MLM (-15,263.0). This difference of -1,919.0 (SE = 48.6; see online supplement), or more than 39 standard errors, strongly supports the notion that the ordinal nature of the data is crucial and that the CP MLM provides a substantially better representation of the underlying response structure.

Direct Replication and Comparison with Pickett et al. (2022)

To directly address concerns about the comparability of our Bayesian estimates with those reported in the original study, we replicated Pickett et al.'s key bivariate analysis (reported in Model 1 of Table S3 in their online supplement) and compared it with our Bayesian linear and multilevel ordinal alternatives. Key results are summarized in Table 3.

Exact (OLS) and Conceptual (Bayesian) Replication of Original Analysis Following Pickett et al.'s specifications, we created the 0–100 normed fear scale and applied OLS regression with robust standard errors. Our frequentist replication yielded a *Black race* coefficient of 31.62 (95% CI: 29.78, 33.46), closely matching our Bayesian linear estimate of 31.65 (see "Estimating Latent Fear with Multilevel Models or Normed Scale" section) and their original estimate of 32.09. Though we observe a very slight difference between the original published estimate and our replicated estimates, which likely reflects a minor difference in sample composition (we identified one missing observation that required a decision about inclusion), these results generally confirm the reproducibility of their findings while validating our analytical approach.

Ordinal Multilevel Alternative In comparison, our multilevel cumulative probit model using the same data treated each item as a separate observation nested within respondents while retaining the ordinal Likert (0 to 4) response scale. Unsurprisingly, this approach yielded a very different fixed effect of $\beta = 0.98$ (95% CrI: 0.86, 1.10), along with a random effect variability estimate of 0.17 (95% CrI: 0.10, 0.29) across items.

Convergent Findings, Complementary Insights Both approaches detect nearly identical effect magnitudes on standardized latent scales. The linear approach suggests Black participants score 31.6 points higher on the 0–100 normed scale, equivalent to about a 1.26-unit difference on the original 0–4 ordinal response scale. The ordinal multilevel approach yields a latent Cohen's $d = 0.98$, virtually matching Pickett et al.'s reported $d = 0.97$ (2022, p.302).

Table 3 Replication and ordinal analysis of bivariate race difference in fear of police from Pickett et al. (2022)

Approach	Method	Scale	Race Coeff [95% CI/CrI]	Interpretation	What this means
Pickett et al. (Original)	OLS + Robust SE	0–100 normed	32.09[a] [30.24, 33.94]	"32.1 percentage point difference"	Percentage of full (ordinal) scale, or: (32.1/100) * 4 = Race difference of 1.28 metric units on original 0–4 ordinal response scale
Exact Replication	OLS + Robust SE	0–100 normed	31.62[b] [29.78, 33.46]	"31.6 percentage point difference"	Percentage of full (ordinal) scale, or: (31.6/100) * 4 = Race difference of 1.26 metric units on original 0–4 ordinal response scale
Bayesian Replication	Bayesian Linear	0–100 normed	31.65[b] [28.19, 35.28]	"31.7 percentage point difference"	Percentage of full (ordinal) scale, or: (31.7/100) * 4 = Race difference of 1.27 metric units on original 0–4 ordinal response scale (~1.3 vs ~2.6, or increase from *unfraid* for Whites to *neither afraid nor unafraid* for Blacks)
Ordinal MLM Alternative	Multilevel Cumulative Probit	0–4 latent scale	FE: 0.98[c] [0.86, 1.10] RE: 0.17[c] [0.10, 0.29]	0.98 SD difference on latent scale; category-specific probabilities (see Fig. 6)	Difference of nearly one full standard deviation (latent Cohen's $d = 0.98$) on latent continuous scale, with substantial item-level effect heterogeneity

[a] See "Black" coefficient from Model 1 in "*Table S3. Comparison of Police-Related Fear, by Race/Ethnicity*" in Pickett et al.'s (2022) online supplement

[b] See Sect. 11.4 (*Lin: Latent effects on normed mean scale*) of our online supplement

[c] See "*MLM: Fit Summary*" and "*MLM: Random effects (re)*" output in Sect. 11.3 (*CP: Multilevel model*) of our online supplement

However, the ordinal multilevel model reveals important distributional insights obscured by scale averaging. The substantial random effect variability (SD = 0.17) indicates that racial disparities vary considerably across different types of police actions. Specifically, disparities are most pronounced for the most serious actions (killing: + 0.24; shooting: + 0.20) while being smaller for more routine interactions (yelling: -0.25; searching: -0.17; stopping: -0.13).[11] This substantial item-level heterogeneity suggests that the racial divide in police-related fear is not uniform across all police behaviors but is especially concentrated in fears of the most severe police violence.

Methodological Implications In this case, both approaches reach convergent substantive conclusions about the existence, direction, and overall magnitude on a *latent* scale of racial disparities in police-related fear. Yet, the ordinal approach provides additional precision about how these disparities manifest across different categories of police behavior, demonstrating that while scale-based approaches might effectively capture overall group differences in these data, ordinal models can reveal substantively important distributional

[11] All random effects coefficients are reported in "*MLM: Random effects (re)*" output in Sect. 11.3 (*CP: Multilevel model*) of our primary online supplement.

patterns. In this case, such differences might inform our understanding of the specific nature of police-related fears in different racial communities. In other cases, differences might reveal assumption violations, effect heterogeneity, or threshold-specific patterns that fundamentally alter substantive conclusions about relationships.

Likewise, the convergence in effect sizes observed here between linear and ordinal approaches is not guaranteed across all datasets or research contexts. Convergence depends on factors such as the degree of skewness, the presence of ceiling/floor effects, the number of response categories, and the underlying distributional patterns. In cases with more extreme ordinal distributions or smaller effect sizes, for instance, the approaches may yield more divergent results, making the choice of analytical approach more consequential for substantive conclusions.

Discussion

Key Findings and Substantive Implications

Our analysis and tutorial provided a practical demonstration of applying and interpreting ordered regression models for ordinal data frequently encountered in criminological research. Through simulation and detailed analysis of Pickett et al.'s (2022) "fear of police" data, we illustrated how conventional linear models can lead to different quantitative conclusions compared to models explicitly designed for ordinal data, particularly regarding predicted probability effect magnitudes and distributional representation on the response scale.

Our analysis of racial differences in fear of police revealed substantively important findings. The ordered probit models suggested a substantially larger proportion of Black participants reporting being "very afraid" of lethal police force compared to estimates derived from linear models – a finding that aligned closely with observed data proportions. When linear models imply probabilities that underestimate the proportion of Black participants who are "very afraid" of police by over 20 percentage points (32.8% vs. 11.0%), this represents more than a statistical technicality. Accurately and completely quantifying such estimates may be crucial for building theories grounded in precise understanding of phenomena, developing effective policy interventions, and accurately representing the lived experiences of research participants.

Methodological Considerations and Practical Guidelines

Convergent Yet Distinct Findings The convergence in latent effect sizes between linear and ordinal approaches ($d \approx 0.97$–0.98) demonstrates that both methods can detect similar overall magnitudes of group differences on standardized latent scales. However, this convergence should not obscure fundamental differences in model adequacy. Our posterior predictive checks and probability comparisons showed that linear models consistently failed to recover actual probability distributions of fear responses, systematically underestimating category-specific probabilities and missing distributional patterns that ordinal models captured accurately.

Model Adequacy and Method Selection The fear of police data exhibited highly skewed ordinal distributions with substantial ceiling and floor effects at the item level, which are precisely the conditions where ordinal models should outperform linear alternatives. That linear models failed to adequately represent probability distributions and missed critical substantive insights even when aggregate effect sizes converged underscores that summary statistic agreement does not guarantee model adequacy or equivalent substantive understanding.

Our findings provide concrete guidance for methodological decision-making. Linear models may suffice for research questions focused solely on detecting the existence and general magnitude of group differences on latent scales measured with items exhibiting normal-like distributions across many response categories. However, our posterior predictive checks and cross-validation results show ordinal models can provide dramatically superior predictive accuracy even when average effects converge, suggesting they better capture underlying ordinal data generating processes. Moreover, ordinal models become essential when research questions concern: (1) accurate representation of response probability distributions, (2) category-specific effects or probabilities, (3) item-level heterogeneity in multi-item constructs, or (4) precise prediction of response patterns.

It is also crucial to remember that the consequences of applying linear models to ordinal data can vary. Though not observed in these data, in other contexts, such mismatches might lead to effect overestimation, Type I (false positive) or Type II (false negative) inference errors, or even incorrect conclusions about the direction of an effect (Liddell and Kruschke, 2018). These possibilities underscore the importance of selecting models whose assumptions plausibly match the data structure or, at minimum, of assessing the potential impact of assumption violations on the conclusions drawn. Choosing methods well-suited to the data structure helps ensure that statistical analyses provide complete and accurate representations of the underlying phenomena rather than potentially incomplete aggregate summaries (cf. Brauer 2025).

Item-Level Heterogeneity Our comprehensive supplementary analyses (e.g., Appendix 1) revealed that race effects varied substantially across individual items, ranging from 0.72 to 1.16 on the latent scale, with stronger effects for more severe police actions (kill; shoot; physical violence). This heterogeneity was successfully captured through Bayesian multilevel modeling with partial pooling, but it would be completely missed by aggregate scaling and linear modeling approaches that assume uniform effects across items.

Proportional Odds and Model Flexibility Diagnostic testing revealed that the proportional odds assumption was violated in these data, with a category-specific model showing substantially better predictive performance (ELPD difference = -84.5 ± 14.7). This indicates that race effects vary meaningfully across ordinal thresholds, with the data suggesting that racial disparities are not uniform across the fear scale (e.g., effects may be larger for extreme fear responses than for moderate ones). The violation highlights the importance of exploring category-specific patterns through predicted probabilities rather than relying solely on averaged coefficients and suggests that researchers should consider category-specific modeling approaches when proportional odds assumptions fail substantially. However, even when this assumption is violated, cross-validation confirms cumulative ordinal models remain vastly superior to linear alternatives that ignore the categorical nature of ordinal data entirely.

The "Default Method" Question If ordinal models perform equally well when linear models are adequate and substantially better when linear models fail, why not simply default to ordinal approaches as Liddell & Kruschke (2018) advocate? The conventional practice of using linear models for computational simplicity or interpretational convenience represents choosing a method that might fail when a superior alternative is readily available. Just as our demonstration showed convergent effect sizes when both methods worked adequately, ordinal models yield results similar to linear models under favorable conditions while providing superior insights when distributional assumptions are violated.

Broader Implications

Precisely Analyzing Imprecise Things Given that Likert-type measures are inherently ordinal and imprecise approximations of presumed underlying constructs, one might ask: Why focus so intensely on analytical precision? This question, while understandable, rests on a false equivalence between measurement precision and analytical precision. Our demonstration shows that even imprecise measures can yield precise insights when analyzed appropriately – and overlooked or misleading conclusions when analyzed inappropriately. The convergence in aggregate effect sizes masked fundamental differences in model adequacy and substantive understanding. A key premise of ours is that criminologists conducting quantitative research *do* care about precision – of their estimates, measures, and theories. If we do not care about precision, then why quantify at all?

Our perspective, echoing Tukey (1977, p. vi), is that quantitative inquiry fundamentally seeks answers to "by how much?" not just "in which direction?" questions. While ordinal models cannot transform imprecise measures into precise ones, they aim to *more faithfully represent the information contained within the existing ordinal data*. Choosing analytical techniques whose assumptions align poorly with the data risks adding unnecessary statistical distortion on top of any inherent measurement limitations. Striving for accurate representation of the data we *do* have seems a prerequisite for building more precise descriptions and, ultimately, theories.

The capacity of cumulative probit models to explicitly estimate thresholds mapping onto observed response categories allows for a more granular understanding of phenomena compared to the broad directional statements or average effects often derived from linear models applied to ordinal data. This highlights the importance of critically examining whether our conventional estimands – typically mean differences on assumed continuous scales – are the most substantively meaningful targets, or whether alternative estimands like category-specific probabilities, threshold effects, or distributional shifts might provide more theoretically- and policy-relevant insights (see Brauer 2025; Brauer and Day 2025). Ordinal methods allow us to estimate traditional average effects when appropriate, while also opening up new estimands that were previously inaccessible or inaccurate under linear approaches.

Scale-Level vs. Item-Level Analysis An apparent objection to our emphasis on analyzing individual ordinal items concerns the well-established preference across social science disciplines for multi-item scales.[12] This preference rests on compelling psychometric arguments: Multi-item scales are expected to improve reliability compared to single items, offer

[12] This objection was raised by a reviewer and the associate editor.

enhanced content validity, and provide more stable measurements by averaging out item-specific error variance. Given these apparent advantages, one might reasonably ask: Why forgo the benefits of established scaling practices in favor of analyzing individual items? Wouldn't such an approach create substantial multiple comparison problems while sacrificing the psychometric gains that justify multi-item measurement?

This objection mischaracterizes our approach. We do not advocate abandoning multi-item measurement models; rather, we demonstrate a methodological progression that begins with item-level analysis to illustrate ordinal methods' superiority for recovering underlying data distributions, then advance to multilevel modeling approaches that can achieve the same psychometric benefits as traditional scaling while respecting ordinality. Multilevel models and SEM/CFA approaches can generate equivalent multi-item measurement models (cf. Mehta & Neale 2005). The key difference is employing methods designed for ordinal data rather than forcing ordinal responses into inappropriate linear frameworks. This approach recognizes both the methodological costs of inappropriate aggregation and the importance of precise quantitative description.

Note, too, that distributional problems survive scale-level aggregation. Šimkovic and Träuble (2019) demonstrate that ceiling and floor effects persist even in aggregated measures, creating "biased noisy inference" and compromising core statistical properties. Micceri's (1989) analysis of 440 large-sample psychological datasets revealed pervasive non-normality, including extreme skewness, multimodality, and heavy tails, even in well-established multi-item scales. This demonstrates that aggregation does not normalize underlying ordinal distributions but often preserves and amplifies their problematic characteristics. Moreover, transforming aggregated ordinal items into seemingly continuous scales, especially when 'normed' (e.g., to a 0–100 range), creates an illusion of metric interpretation where researchers interpret coefficients as 'percentage point' changes when this is fundamentally inappropriate.

Concerns about multiple comparisons in item-specific analyses also reflect misunderstanding of our methodological progression.[13] Our specific approach addresses multiple comparison concerns through hierarchical pooling while maintaining item-specific analytical precision (cf. Gelman et al. 2012; Garcia 2024; but see Ogle et al. 2019). Yet, the choice is not between item-specific versus multi-item analysis; it is between analytical approaches that accurately represent data structure versus those that sacrifice precision for convenience.[14]

When the primary aim is rich quantitative description to identify robust phenomena, which is a core tenet of addressing the "theory crisis" in social science (cf. Eronen & Bringmann 2021; Brauer & Day 2025), our progression from item-level to multilevel ordinal modeling provides accurate representation of underlying phenomena while preserving the psychomet-

[13] This was another objection raised by the associate editor.

[14] This perspective aligns with broader critiques of uncritical latent variable reification in psychological measurement (e.g., Ropovik 2015; Flake & Fried, 2020; Rhemtulla et al. 2020). Revelle (2024), a respected figure in psychometrics, argues that factors derived from aggregation often represent "convenient fictions" serving as organizational tools rather than reflecting underlying causal entities, emphasizing that "aggregation should be purposeful" and that "our latent variables are just descriptive summaries of the items rather than causal." "Similarly, Fried's work demonstrates that summing ordinal symptom ratings can obscure critical information about individual symptoms and their distinct etiologies, correlates, and consequences – information entirely lost when items are aggregated (Fried & Nesse 2015; Fried et al. 2016; Fried et al. 2022." Measurement invariance issues, where differential item functioning can systematically bias group comparisons, also are not resolved but hidden within composite scores (Rhemtulla et al. 2020).

ric advantages that justify multi-item measurement. As our empirical demonstration shows, these methodological choices can yield substantially different quantitative conclusions about important social phenomena – differences that may matter for theory development, accurate representation of lived experiences, policy formation, and intervention design.

Do We Really Need This? Our emphasis on achieving accurate quantitative description is not intended as methodological pedantry but stems from a view that this is a fundamental component of cumulative empirical science (Brauer & Day 2025). Recommendations to consider models like ordered regression for ordinal data arise from their potential to provide more accurate and complete information compared to alternatives whose assumptions may be violated. Rather than viewing such recommendations merely as reviewer hurdles, we suggest they represent opportunities to potentially strengthen inferences and extract more meaning from valuable, hard-earned data. Within research areas where theoretical predictions may lack metric precision, adopting analytical practices that yield detailed and accurate quantitative descriptions can be particularly valuable for building a solid empirical foundation and advancing theoretical development (Eronen and Bringmann 2021; Scheel et al. 2021).

We think such progressive problem shifts will require us collectively to move beyond default analyses that primarily offer dichotomous "significant/nonsignificant" conclusions or vague directional statements (Brauer 2025). Efficient knowledge accumulation often depends on precisely quantifying effect magnitudes and characterizing distributional patterns completely and accurately. Importantly, ordinal models might facilitate richer descriptive and predictive insights through full distributional modeling, natural predicted probability calculations, and threshold-specific estimates. At the same time, they fully support conventional hypothesis testing, now applied to estimates generated by faithful modeling of underlying ordinal data structures.

Conclusion

Our analyses and tutorial demonstrated the value of ordered regression models for enhancing accuracy and completeness when modeling ordinal outcome data in criminological research. By embracing techniques specifically designed for such data, criminologists gain powerful tools to accurately capture distributional patterns and estimate effects on specific, meaningful quantities such as category-specific probabilities. The practical steps demonstrated, from visualizing ordinal distributions with mosaic plots to extracting predicted probabilities, provide concrete tools applicable across diverse criminological contexts including offense severity ratings, risk assessments, and attitudinal scales.

As the field confronts challenges of replication and reproducibility while seeking to improve theories and empirical knowledge, the pursuit of analytical precision through appropriate methodology becomes paramount. Rigorous analyses that carefully map statistical models to underlying data generating processes are essential for accurately describing the complexities of our social world. Equally important is critically evaluating whether our chosen estimands are the most substantively meaningful for building a quantitative criminology that is increasingly robust in its findings and effective in informing policy. The choice should not be 'when is linear close enough?' but rather 'how can we accurately extract the most complete information from our hard-earned data?' Using methods tailored to its complexity is a first but crucial step.

Appendix 1

Table 4 Overview of R tutorial structure and contents[1]

Tutorial Section	Statistical Approach	Model Family	Data Structure	Convergence Status	Purpose & Key Insights
Model 1	Frequentist OLS	Gaussian	Single-level	Replicates original	Baseline comparison with linear approach
Model 2	Bayesian Linear	Gaussian	Single-level	Converges	Bayesian equivalent of original analysis
Model 3a	Frequentist CP MLM	Cumulative probit	Multilevel	Convergence issues	Frequentist multilevel ordinal (often fails to converge)
Alternative 3a	Item-Specific CP Models	Cumulative probit	Item-specific	Frequentist alternative	Simpler & more reliable non-MLM fallback if frequentist MLM convergence fails
Model 3b	Bayesian CP MLM	Cumulative probit	Multilevel	Converges successfully	Primary recommended multilevel ordinal approach
Proportional Odds Test	Category-Specific Model	Categorical logit	Multilevel	For proportional odds assumption testing	Tests whether race effects vary across thresholds
Predicted Probabilities	Probability Estimation	N/A	From fitted models	Demonstrates interpretation	Shows interpretable outputs: item-specific vs population-level

[1]This table summarizes the modeling approaches used to analyze ordinal fear of police data that we cover in our R tutorial (referred to as Appendix 1). The tutorial also demonstrates why Bayesian multilevel ordinal models are preferred over linear approaches. Complete R code and detailed explanations ("Appendix-1-Tutorial" Quarto file) and data (in "Data" folder) for our tutorial, which is distinct from yet complements our primary online supplement, can also be found at the authors' website (https://reluctantcriminologists.com/supp/bd-supp/) and the project's OSF page (https://osf.io/n52kc/)

Appendix 2

Table 5 Extreme fear responses by race and police action

	Very Unafraid ($y=0$)			Very Afraid ($y=4$)		
Police Action	White	Black	B−W Diff	White	Black	B−W Diff
Stop You	28.8	8.9	−19.9	6.3	25.4	19.0
Search You	31.5	9.9	−21.7	6.5	24.6	18.1
Yell at You	29.2	10.3	−18.9	7.9	24.6	16.7
Handcuff You	33.1	9.7	−23.4	7.3	28.7	21.3
Kick/Hit You	41.9	10.5	−31.4	8.7	32.6	23.8
Pin You Down	41.1	10.5	−30.6	9.5	34.7	25.2
Pepper Spray You	41.7	10.5	−31.2	8.9	32.6	23.6
Tase You	41.9	10.3	−31.6	9.5	34.3	24.8
Shoot at You	44.2	10.5	−33.8	10.7	40.3	29.6
Kill You	44.2	11.4	−32.8	10.3	43.6	33.3

Observed percentages reporting highest and lowest levels of fear of police in Pickett et al. (2022) data

Appendix 3

Table 6 Multilevel model comparisons using leave-one-out cross-validation

Model	ELPD LOO[1]	SE of ELPD	ELPD Difference[2]	SE of Difference
Cumulative Probit	−15,263.0	47.0	−84.8	14.7
Cumulative Probit w/Unequal Variances	−15,262.8	47.0	−84.7	14.7
Categorical Logit	−15,178.2	47.9	0.0	0.0
Linear	−17,182.0	56.6	−2,003.8	50.4

[1] ELPD LOO: Expected Log Pointwise Predictive Density

[2] Difference compared to category-specific logit model

Summary of expected log pointwise predictive densities (ELPD), providing a quantitative measure of the relative fit of various models predicting latent fear of police. Higher ELPD values indicate better model fit. The linear model's ELPD (-17,182.0) is considerably lower than the CP MLM's (-15,263.0). Additional supplementary models (see online supplement) shows a Categorical Logit MLM model improves model fit even more (-15,178.2), though at the cost of a more complex model with a larger number of estimated parameters

Acknowledgements We wish to extend our heartfelt appreciation to Pickett, Graham, and Cullen for practicing open science and making the reanalysis possible. We also thank the editors and anonymous reviewers for helpful comments on earlier drafts.

Funding No funding was received for conducting this study.

Data Availability A primary online supplement containing all R code, results, and supplementary output for analyses presented in this paper, and a general-purpose tutorial for implementing ordinal modeling techniques in R (Appendix 1) is available online at https://reluctantcriminologists.com/supp/bd-supp/ and https://osf.io/n52kc/.

Declarations

Competing Interest The authors have no financial or proprietary interests in any material discussed in this article.

References

Agresti A (1989) Tutorial on modeling ordered categorical response data. Psychol Bull 105(2):290–301. https://doi.org/10.1037/0033-2909.105.2.290

Agresti A (2010) *Analysis of Ordinal Categorical Data* (Second). Hoboken, NJ: Wiley. Retrieved from https://onlinelibrary.wiley.com/doi/epub/https://doi.org/10.1002/9780470594001

Allison PD (1999) Comparing logit and probit coefficients across groups. Sociol Methods Res 28(2):186–208

Angrist JD, Pischke JS (2009) Mostly harmless econometrics: An empiricist's companion. Princeton University Press
Ball Will (2020) Visualising two ordinal variables using R. Retrieved March 14, 2025, from https://will-ball.github.io/ordinalplots/
Baranauskas AJ, Drakulich KM (2018) Media construction of crime revisited: media types, consumer contexts, and frames of crime and justice. Criminology 56(4):679–714. https://doi.org/10.1111/1745-9125.12189
Berkson J (1944) Application of the logistic function to bio-assay. J Am Stat Assoc 39(227):357–365. https://doi.org/10.1080/01621459.1944.10500699
Bliss CI (1934) The method of probits. Science 79(2037):38–39
Brauer, J. R., & Day, J. C. (2025). Money in the bank: progressive problem shifts in criminological science. *Theory and Society*. https://doi.org/10.1007/s11186-025-09645-z
Brauer JR (2025) Loose derivation chains and invisible anomalies in theory testing: evidence from self-control research. https://doi.org/10.31235/osf.io/n4xuf_v3
Britt CL, Rocque M, Zimmerman GM (2018) The analysis of bounded count data in criminology. J Quant Criminol 34(2):591–607. https://doi.org/10.1007/s10940-017-9346-9
Bürkner P-C (2017) Brms: an R package for Bayesian multilevel models using Stan. J Stat Softw 80:1–28. https://doi.org/10.18637/jss.v080.i01
Bürkner P-C, Vuorre M (2019) Ordinal regression models in psychology: a tutorial. Adv Methods Pract Psychol Sci 2(1):77–101. https://doi.org/10.1177/2515245918823199
Cameron AC, Trivedi PK (2013) *Regression Analysis of Count Data* (2nd ed.). Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781139013567
Carifio J, Perla R (2008) Resolving the 50-year debate around using and misusing Likert scales. Med Educ 42(12):1150–1152. https://doi.org/10.1111/j.1365-2923.2008.03172.x
Eronen MI, Bringmann LF (2021) The theory crisis in psychology: how to move forward. Perspect Psychol Sci 16(4):779–788
Fechner GT (1860) Elemente *der Psychophysik*. Leipzig, Germany: Breitkopf und Härtel
Fried EI, Nesse RM (2015) Depression sum-scores don't add up: why analyzing specific depression symptoms is essential. BMC Med 13:1–11
Fried EI, van Borkulo CD, Epskamp S, Schoevers RA, Tuerlinckx F, Borsboom D (2016) Measuring depression over time... or not? Lack of unidimensionality and longitudinal measurement invariance in four common rating scales of depression. Psychol Assess 28(11):1354
Fried EI, Flake JK, Robinaugh DJ (2022) Revisiting the theoretical and methodological foundations of depression measurement. Nat Rev Psychol 1(6):358–368
Garcia GD (2024) *Bayesian* estimation *in multiple comparisons*
Gelman A, Hill J, Yajima M (2012) Why we (usually) don't have to worry about multiple comparisons. J Res Educ Eff 5(2):189–211
Goodrich B, Gabry J, Ali I, Brilleman S (2020) *Rstanarm: bayesian applied regression modeling via stan*. Retrieved from https://mc-stan.org/rstanarm/
Greenberg DF (2016) Criminal careers: discrete or continuous? J Dev Life-Course Criminol 2(1):5–44. https://doi.org/10.1007/s40865-016-0029-2
Greene WH, Hensher DA (2010) Modeling Ordered Choices: A Primer. Cambridge University Press, Cambridge. https://doi.org/10.1017/CBO9780511845062
Hannan KR, Cullen FT, Graham A, Jonson CL, Pickett JT, Haner M, Sloan MM (2023) Public support for second look sentencing: is there a Shawshank redemption effect? Criminol Public Policy 22(2):263–292. https://doi.org/10.1111/1745-9133.12616
Herman S, Pogarsky G (2024) Moral salience, situational moral evaluations, and criminal choice. Crime Delinq. https://doi.org/10.1177/00111287241259474
Jones AM, Vaughan AD, Roche SP, Hewitt AN (2022) Policing persons in behavioral crises: an experimental test of bystander perceptions of procedural justice. J Exp Criminol 18(3):581–605. https://doi.org/10.1007/s11292-021-09462-1
Jung RC, Kukuk M, Liesenfeld R (2006) Time series of count data: modeling, estimation and diagnostics. Comput Stat Data Anal 51(4):2350–2364. https://doi.org/10.1016/j.csda.2006.08.001
Liu Q, Shepherd BE, Li C, Harrell FE Jr (2017) Modeling continuous response variables using ordinal regression. Stat Med 36(27):4316–4335. https://doi.org/10.1002/sim.7433
Lundberg I, Johnson R, Stewart BM (2021) What is your estimand? Defining the target quantity connects statistical evidence to theory. Am Sociol Rev 86(3):532–565. https://doi.org/10.1177/00031224211004187
Madon NS, Murphy K, Williamson H (2023) Justice is in the eye of the beholder: a vignette study linking procedural justice and stigma to Muslims' trust in police. J Exp Criminol 19(3):761–783. https://doi.org/10.1007/s11292-022-09510-4
McNeish D, Somers JA, Savord A (2024) Dynamic structural equation models with binary and ordinal outcomes in M plus. Behav Res Methods 56(3):1506–1532

Mehta PD, Neale MC (2005) People are variables too: multilevel structural equations modeling. Psychol Methods 10(3):259
Metcalfe C, Pickett JT (2022) Public fear of protesters and support for protest policing: an experimental test of two theoretical models*. Criminology 60(1):60–89. https://doi.org/10.1111/1745-9125.12291
Micceri T (1989) The unicorn, the normal curve, and other improbable creatures. Psychol Bull 105(1):156
Mummolo J (2018) Militarization fails to enhance police safety or reduce crime but may harm police reputation. Proc Natl Acad Sci U S A 115(37):9181–9186. https://doi.org/10.1073/pnas.1805161115
Munksgaard R, Ferris JA, Winstock A, Maier LJ, Barratt MJ (2023) Better bang for the buck? Generalizing trust in online drug markets. Br J Criminol 63(4):906–928. https://doi.org/10.1093/bjc/azac070
Nguyen TQ, Webb-Vargas Y, Koning IM, Stuart EA (2016) Causal mediation analysis with a binary outcome and multiple continuous or ordinal mediators: simulations and application to an alcohol intervention. Struct Equ Modeling 23(3):368–383
Ogle K, Peltier D, Fell M, Guo J, Kropp H, Barber J (2019) Should we be concerned about multiple comparisons in hierarchical Bayesian models? Methods Ecol Evol 10(4):553–564
Osgood DW, Finken LL, McMorris BJ (2002) Analyzing multiple-item measures of crime and deviance II: tobit regression analysis of transformed scores. J Quant Criminol 18(4):319–347. https://doi.org/10.1023/A:1021198509929
Peyton K, Sierra-Arévalo M, Rand DG (2019) A field experiment on community policing and police legitimacy. Proc Natl Acad Sci U S A 116(40):19894–19898. https://doi.org/10.1073/pnas.1910157116
Pickett JT, Graham A, Cullen FT (2022) The American racial divide in fear of the police. Criminology 60(2):291–320. https://doi.org/10.1111/1745-9125.12298
Pickett JT, Graham A, Nix J, Cullen FT (2024) Officer diversity may reduce Black Americans' fear of the police. Criminology 62(1):35–63. https://doi.org/10.1111/1745-9125.12360
Revelle W (2024) The seductive beauty of latent variable models: or why I don't believe in the Easter Bunny. Pers Individ Differ 221:112552
Rhemtulla M, van Bork R, Borsboom D (2020) Worse than measurement error: consequences of inappropriate latent variable measurement models. Psychol Methods 25(1):30
Ripley B, Venables B, Bates DM, Hornik K, Gebhardt A, Firth D, Ripley MB (2013) Package 'mass.' Cran r 538(113–120):822
Ropovik I (2015) A cautionary note on testing latent variable models. Front Psychol 6:1715
Scheel AM, Tiokhin L, Isager PM, Lakens D (2021) Why hypothesis testers should spend less time testing hypotheses. Perspect Psychol Sci 16(4):744–755. https://doi.org/10.1177/1745691620966795
Silver JR (2020) Moral motives, police legitimacy and acceptance of force. Policing 43(5):799–815. https://doi.org/10.1108/PIJPSM-04-2020-0056
Šimkovic M, Träuble B (2019) Robustness of statistical methods when measure is affected by ceiling and/or floor effect. PLoS ONE 14(8):e0220889
Smith P, Cullen FT, Pickett JT, Jonson CL (2025) Coerced motherhood behind bars: Public support for abortion access for incarcerated women. Criminol Public Policy 24(1):3–31. https://doi.org/10.1111/1745-9133.12675
Svensson R, Oberwittler D (2021) Changing routine activities and the decline of youth crime: a repeated cross-sectional analysis of self-reported delinquency in Sweden, 1999–2017. Criminology 59(2):351–386. https://doi.org/10.1111/1745-9125.12273
Torres CE, D'Alessio SJ, Stolzenberg L (2024) Marginal deterrence: the effect of illicit incentive on robbery escalation. J Econ Criminol 4:100062. https://doi.org/10.1016/j.jeconc.2024.100062
Tukey J (1977) *Exploratory Data* Analysis (1st edition). Reading, Mass. Pearson
Valle D, Ben Toh K, Laporta GZ, Zhao Q (2019) Ordinal regression models for zero-inflated and/or overdispersed count data. Sci Rep 9(1):3046. https://doi.org/10.1038/s41598-019-39377-x
Vehtari A, Gelman A, Gabry J (2017) Practical bayesian model evaluation using leave-one-out cross-validation and WAIC. Stat Comput 27(5):1413–1432. https://doi.org/10.1007/s11222-016-9696-4
Weisburd D, Wilson DB, Wooditch A, Britt C (2022) *Advanced Statistics in Criminology and Criminal Justice* (Fifth Edition 2022). Springer, Cham
Wheeler AP, Silver JR, Worden RE, Mclean SJ (2020) Mapping attitudes towards the police at micro places. J Quant Criminol 36(4):877–906. https://doi.org/10.1007/s10940-019-09435-8
Williams R (2009) Using heterogeneous choice models to compare logit and probit coefficients across groups. Sociol Methods Res 37(4):531–559
Winship C, Mare RD (1984) Regression models with ordinal variables. Am Sociol Rev 49(4):512. https://doi.org/10.2307/2095465
Wozniak KH, Pickett JT, Brown EK (2022) Judging hardworking robbers and lazy thieves: an experimental test of act- vs. person-centered punitiveness and perceived redeemability. Justice Q 39(7):1565–1591. https://doi.org/10.1080/07418825.2022.2111326

Liddell, T. M., & Kruschke, J. K. (2018). Analyzing ordinal data with metric models: what could possibly go wrong?. *Journal of Experimental Social Psychology*, *79*, 328-348.

Bauer, D. J., & Sterba, S. K. (2011). Fitting multilevel models with ordinal outcomes: performance of alternative specifications and methods of estimation. *Psychological methods*, *16*(4), 373.

Kołczyńska, M., Bürkner, P. C., Kennedy, L., & Vehtari, A. (2024). Modeling public opinion over time and space: trust in state institutions in Europe, 1989-2019. In *Survey Research Methods* (Vol. 18, No. 1, pp. 1-19).

Donnellan, E., Usami, S., & Murayama, K. (2023). Random item slope regression: an alternative measurement model that accounts for both similarities and differences in association with individual items. *Psychological Methods 30*(4), 744-769.

Flora, D. B., & Curran, P. J. (2004). An empirical evaluation of alternative methods of estimation for confirmatory factor analysis with ordinal data. *Psychological methods*, *9*(4), 466-491.

DiStefano, C., & Morgan, G. B. (2014). A comparison of diagonal weighted least squares robust estimation techniques for ordinal data. *Structural Equation Modeling: a multidisciplinary journal*, *21*(3), 425-438.

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

Lee, H., Pickett, J. T., Graham, A., Cullen, F. T., Jonson, C. L., Haner, M., & Sloan, M. M. (2024). Punitiveness toward social distancing deviance in the COVID-19 pandemic: findings from two national experiments. Journal of Experimental Criminology. https://doi.org/10.1007/s11292-024-09610-3

Flake, J. K., & Fried, E. I. (2020). Measurement schmeasurement: questionable measurement practices and how to avoid them. *Advances in methods and practices in psychological science*, *3*(4), 456-465.

Authors and Affiliations

Jonathan R. Brauer[1] · Jacob C. Day[2]

✉ Jonathan R. Brauer
jrbrauer@iu.edu

[1] Criminology & Criminal Justice, Indiana University Bloomington, Sycamore Hall Room 307. 1033 E 3Rd St, Bloomington, IN, USA

[2] Sociology & Criminology, University of North Carolina Wilmington, Wilmington, NC, USA