Coarsened by Design: Measurement Error Bias in Grouped Regressors and Interaction Terms

Ramses Llobet — Department of Political Science, University of Washington — rllobet@uw.edu

Abstract

Many **continuous variables** in survey and administrative data are observed only through **ordered bins**, yet the inferences drawn from them are about the latent quantity itself.

- *What follows when a latent variable with a real metric is replaced by a proxy built out of its own bins?*

Standard practice takes the **bin midpoint**, closing the open top bracket with a Pareto tail (Hout 2004). This project answers with a theory of **coarsening-induced measurement error** (CIME): coarsened proxies of skewed, heavy-tailed latents carry **differential** error, biasing the treatment effect and, far more severely, its interaction with any moderator correlated with the latent.

2. The research setting

(1)	$Y = \alpha + \beta_Z Z + \varepsilon$	**structural: the true model**
(2)	$R = k \iff \tau_{k-1} \le Z < \tau_k$	**coding: fixed by the survey**
(3)	$W = f(R), \quad U = Z - W$	**decoding: chosen by the analyst**
(4)	$Y = \alpha + \beta_Z W + e_W, \;\; e_W = \varepsilon + \beta_Z U$	**feasible: the regression actually run**

- Y is the outcome and ε its ordinary regression error;
- Z is the true income; its effect β_Z is the estimand, and it is *never observed*;
- R is all the survey delivers: which bracket Z fell into, between *known* cutpoints τ, the top one **open at the upper end**.

Imputing a *score* $W = f(R)$ maps the bracket back onto the interval scale and introduces measurement error, $U = Z - W$. Substituting $Z = W + U$ into (1) yields (4): the leftover error e_W is **no longer** ε but $\varepsilon + \beta_Z U$, and U is itself a function of W.

Regressor and error are linked by construction.

3. The Berkson property

(a) All of the bias is one number, and it has two readings.

$$\operatorname{plim} \widehat{\beta}_W = \beta_Z \, s_W$$

$$s_W = \underbrace{\operatorname{Corr}(Z, W)}_{\textbf{validity}} \times \underbrace{\frac{\operatorname{sd}(Z)}{\operatorname{sd}(W)}}_{\textbf{scale ratio}} \qquad \text{the } \textbf{channel} \text{ reading}$$

$$\underbrace{s_W - 1}_{\textbf{relative bias}} = \frac{\operatorname{Cov}(W, U)}{\operatorname{Var}(W)} \qquad \text{the } \textbf{error} \text{ reading}$$

A proxy biases the slope through exactly two channels: **validity**, how closely W correlates with Z, and the **scale ratio**. A coarsened value *always* has validity below one, which introduces a ceiling factor. Standardizing the variable cancels the scale channel, so part of the bias is **a scale artifact**, but this is a necessity if we want to make inferences on the latent variable units.

(b) When is the error U nondifferential? With $\mu_k = \mathbb{E}[Z \mid R = k]$ the mean *inside* interval k satisfies the Berkson property, while any *miscentering*, $d_k = \mu_k - f(k)$, amplifies it:

$$U = \underbrace{(Z - \mu_R)}_{\textbf{scatter}} + \underbrace{d_R}_{\textbf{miscentering}} \implies \operatorname{Cov}(W, U) = \operatorname{Cov}\big(f(R),\, d_R\big).$$

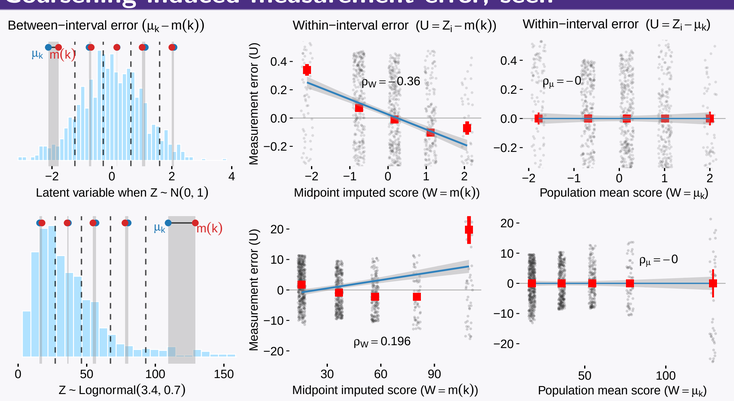
Under strong latent skewness and a heavy tail, miscentering tends to move with the score. **All of the bias is that co-movement.**

(c) Exactly one score has no miscentering.

midpoint, $f = m$ $\quad d_k \neq 0 \implies s_W \neq 1$ (bias)
bracket mean, $f = \mu_R$ $\quad d_k \equiv 0 \implies s_W = 1$ (no bias)

The conditional mean removes the part that is a **function of the bracket**'s scale. When $\mathbb{E}[U \mid W] = 0$, the **Berkson property**, and Berkson error does not bias a slope.

4. Coarsening-induced measurement error, seen



Left: the density and its cutpoints, the midpoint against the bracket mean; the gap is d_k. Middle: the error against the midpoint score slopes (-0.36 normal, $+0.196$ lognormal). Right: against the bracket-mean score it is flat at zero. The midpoint's bias cannot be signed in advance: the slope flips with the shape of the density. The conditional mean is flat regardless.

5. The fix: a calibrated proxy, and honest uncertainty

Two equations, solved in sequence. Fit a parametric distribution on the brackets by maximum likelihood, using only the upper and lower limits the survey certifies: each individual's income lies *somewhere inside* the reported bin interval. Then put its conditional mean into the outcome regression in income's place:

$$\ell(\theta) = \sum_i \log\Big[\, F(\tau_{R_i}; \theta(X_i)) - F(\tau_{R_i - 1}; \theta(X_i)) \,\Big], \qquad W_i = \mathbb{E}\big[Z \mid R_i, X_i; \widehat{\theta}\,\big].$$

The proxy must condition on *everything the outcome model conditions on*, or the bias merely moves onto the covariates.

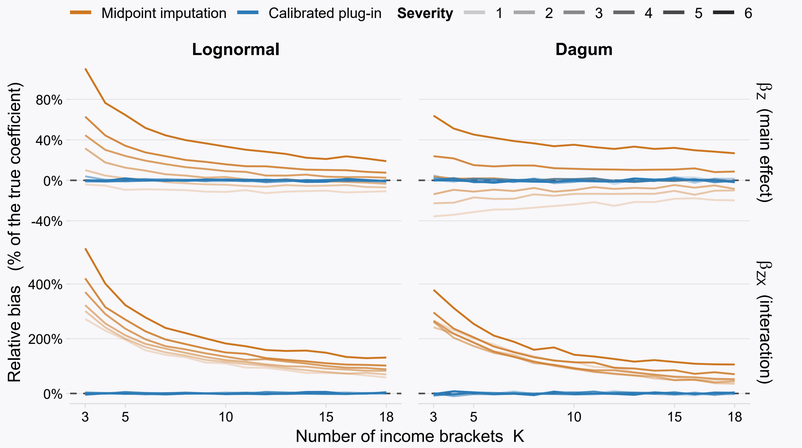
Heavy tails need a regularized anchor. μ_K is governed by the tail index ξ, which bracket counts barely identify. A *reach-gated* empirical-Bayes penalty anchors $\log \xi$ with a robust Student-t prior of precision $\lambda = \lambda_0(1 - \text{reach}^2)$: **the penalty stands down exactly to the extent the data already reach into the tail.**

The proxy is a generated regressor. A naive standard error treats W as known and *under-covers*. I address **measurement uncertainty** by drawing $\theta^{(m)}$ from its asymptotic distribution, rebuild the proxy, refit, and pool by Rubin's rules:

$$V_{\text{cd}} = V_{\text{within}} + \Big(1 + \tfrac{1}{M}\Big) B.$$

6. Monte Carlo Experiment: 300 simulations per cell

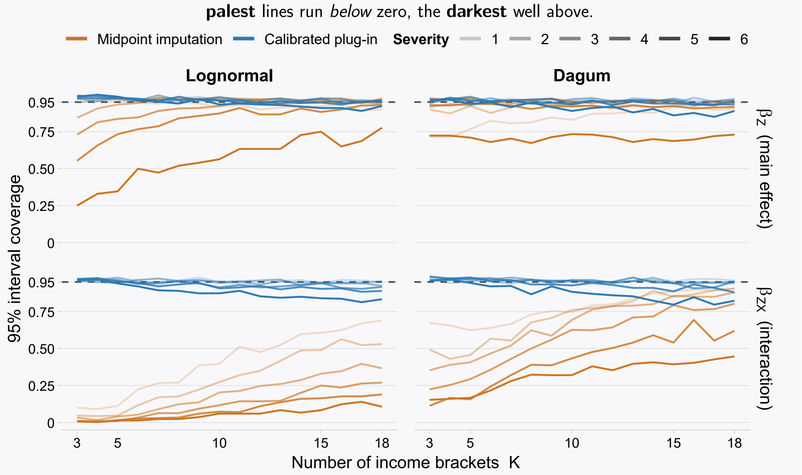
Design. Twelve DGPs (six **lognormal**, $\sigma = 0.6$ to 1.2; six Dagum, tail index $a = 5$ down to 2, the goal is to assess bias as we increase **skewness** and **heavy tail** severity). Within each DGP, we increase bins from $K = 3$ to 18. We test **coarsened treatment** measurement error and its **interaction** with a binary moderator that shifts income by δ, so the groups' within- bracket means differ. $N = 1000$. Truth: $\beta_Z = 1$, $\beta_{ZX} = 0.5$. **Two proxy comparisons:** the **midpoint imputation** against the **calibrated plug-in**.



One scale, both rows. At five brackets the midpoint inflates the interaction by $+196\%$ to $+323\%$ depending on tail severity (lognormal): a true 0.5 read as 1.5 to 2.1. At eighteen it is still $+59\%$ to $+131\%$. More brackets do not remove it.

The **plug-in** stays within 6% of the truth at every $K \geq 5$, and never worse than 10% even at three brackets.

Top row: the midpoint's main-effect bias *changes sign* with severity, so it cannot be signed a priori: the **palest** lines run *below* zero, the **darkest** well above.



Nominal 95% intervals. At five brackets the midpoint's interaction interval covers the truth 1% to 11% of the time (lognormal); at eighteen, 11% to 69%. The plug-in sits at 0.94–0.98 at five brackets and never falls below 0.80 anywhere.

Both panels: $\delta = 0.8$, the strongest coupling. **Colour is the estimator; shade is severity** — pale = mildest, dark = most severe (lognormal $\sigma = 0.6 \to 1.2$; Dagum $a = 5 \to 2$), six lines per arm. These are *not* confidence intervals: Monte Carlo error at 300 simulations is about ± 0.03 and is far too small to see.

7. Beramendi, Cansunar & Kwon (2025): information and redistribution

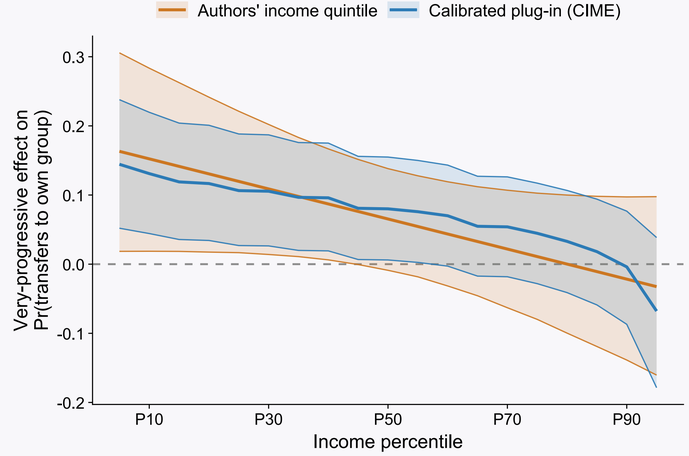
What they fit. A survey experiment in Korea ($N = 1000$). Respondents are told the fiscal system is more progressive than they believed. Outcome: supports higher transfers *to their own income group*. Weighted LPM, HC1.

$$Y_i = \beta_0 + \beta_1 T_i^V + \beta_2 W_i + \underbrace{\beta_3 \left(T_i^V \times W_i\right)}_{\textbf{the hypothesis}} + \mathbf{C}_i^\top \boldsymbol{\gamma} + \varepsilon_i, \qquad W_i = \text{income quintile}.$$

Hypothesis. The *poor* should demand more redistribution and the *rich* should not, so the treatment effect must **shrink as income rises**: $\beta_3 < 0$.

Income measure	$\widehat{\beta}_3$	SE	p
Authors' quintile	-0.060	0.036	0.095
Calibrated plug-in	$\mathbf{-0.098}$	0.032	**0.002**

Standardized per SD of each arm's own income measure. Same N, same model: *only the income variable changes.*



The quintile cannot rule out zero: P10→P90 change -0.174, CI $[-0.379, 0.030]$. The plug-in can: -0.135, CI $[-0.223, -0.048]$. The moderation BCK hypothesized is real. Their own income variable could not resolve it.

8. Bartolini et al. (2023): a social cure?

What they fit. ESS round 9, 29 countries ($N = 38{,}597$). Outcome: life satisfaction, 0–10. OLS, country fixed effects, HC1.

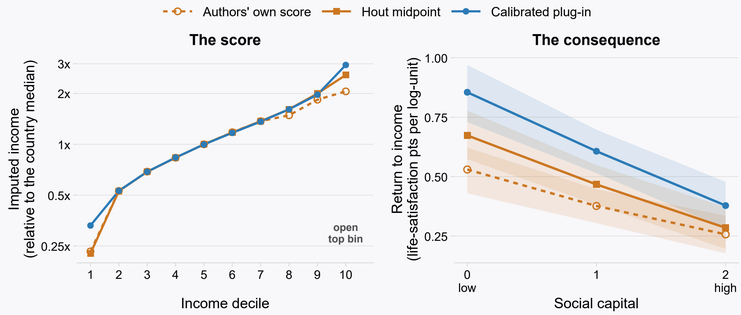
$$Y_i = \beta_0 + \beta_1 \log W_i + \sum_{k=1}^{2} \Big[\, \delta_k \, \mathrm{SC}_{ik} + \underbrace{\gamma_k \big(\mathrm{SC}_{ik} \times \log W_i\big)}_{\textbf{the hypothesis}} \Big] + \mathbf{C}_i^\top \boldsymbol{\theta} + \mu_{c(i)} + \varepsilon_i,$$

with $\mathrm{SC} \in \{0, 1, 2\}$ social capital and W_i = bracket-imputed household income.

Hypothesis. Social capital is a **social cure** for the race for income: the return to income should **fall** as social capital rises, so $\gamma_k < 0$. They report the income effect roughly halving.

Income score	$\widehat{\gamma}_2$	SE	p
Authors' own score	-0.274	0.054	< 0.001
Hout midpoint	-0.391	0.062	< 0.001
Calibrated plug-in	$\mathbf{-0.476}$	0.067	$< \mathbf{0.001}$

Identical brackets, identical model, identical log transform: *only the within-bracket score changes.* The lower two rows are on one real scale (annual 2015 international dollars, recovered from the showcards' own currency and period). The **plug-in** lands between the **midpoint** and the full-information EM (-0.589): **the likelihood ladder, on real data**.



Left: the score. The three rules are indistinguishable through every closed bracket and part company in the open top bin alone (2.06, 2.58, 2.95× the country median). Right: the consequence. Every score finds the buffering Bartolini predicts, but the coarse ones understate it: doubling income buys $+0.59$ life-satisfaction points at low social capital and $+0.26$ at high under the plug-in, against $+0.37 \to +0.18$ under the authors' score. The social cure is real. Their own measure could not see how strong it was, and the damage entered through one bin.

A score imputed from a bracket is not a noisy income. It is an *exact* estimator of a target the analyst never chose, and its bias cannot be signed in advance.

The within-bracket conditional mean is the unique score whose error the bracket cannot see. Not a better guess: the *only* unbiased one.

Damage concentrates in interactions, and *more brackets do not fix it.* In two published papers, correcting only the income variable **recovers a hypothesis that was there all along**.

Replicated: Beramendi, Cansunar & Kwon (2025), "Cheap Egalitarians? Incidence, Information and Demand for Redistribution," working paper · Bartolini, Piekalkiewicz, Sarracino & Slater (2023), "The moderation effect of social capital…," *PLOS ONE* 18(12): e0288455 · `intreg` (R) · authors' coarse measure · calibrated plug-in