Identifying causal effects with subjective ordinal outcome data

Leonard Goff*

This version: September 2, 2025

Abstract

This paper studies the use of responses on ordered scales as an outcome in causal inference, accounting for variation between individuals in the interpretation of the response categories. When treatment variables are statistically independent of both potential outcomes and individuals' interpretations, the conditional mean function of response category number given the treatments identifies the signs of particular convex averages of causal effects. For example, nonparametric regressions with continuous treatments capture mean effects among individuals on the margin between successive response categories. While magnitudes for one continuous treatment variable are alone not quantitatively meaningful, comparisons between two such treatment variables are.

*Department of Economics, University of Calgary. For useful conversations, I thank Christopher Barrington-Leigh, Carol Caetano, Andrew Clark, Ben Crost, Junlong Feng, John Helliwell, Peter Hull, Caspar Kaiser, Sylvia Klosin, Louise Laage, Jean-William Laliberté, Simon Lee, Erzo Luttmer, Guy Mayraz, Max Norton, Bernard Salanié, Adam Rosen, Kevin Song, Takuya Ura and Sam Viavant. I thank Erzo Luttmer and Social Science Research Services at the University of Wisconsin for help with data access. Any mistakes are my own.

1 Introduction

Many survey questions ask respondents to choose from a set of two or more ordered categories that lack clear definitions, leaving the interpretation of those categories to the respondent. Examples include self-reported health status (SRHS), product or service ratings, job satisfaction, and questions gauging satisfaction with life overall. Individuals' responses are then often used as an outcome variable in research, frequently as a proxy for some underlying latent variable of interest (e.g. true health in the case of SRHS).[1]

A key question for this practice is how "reporting functions"—the way that individuals map that latent variable into one of the available response categories—impact conclusions drawn from the data. Bond and Lang (2019) influentially show that even if individuals share a common reporting function (but it is not ex-ante known to the researcher), averages of the latent variable cannot be meaningfully compared between groups using their responses, absent strong restrictions on the latent variable's unobserved distribution. More fundamentally, if the response categories lack objective definitions, reporting functions might vary between individuals, potentially confounding any attempt to study relationships between explanatory variables and the latent variable.

This paper shows that the observed categorical responses can nevertheless be informative about causal relationships in which this latent variable is the outcome, despite the dual threats of reporting functions being both (i) unknown to the researcher and (ii) heterogeneous across respondents. I decompose differences in the observed distribution of responses between covariate values into causal effects of those covariates on the latent variable. To do so, I strengthen the familiar selection-on-observables assumption that one or more explanatory variables are statistically independent of potential outcomes, adding to it that those explanatory variables are also independent of heterogeneity in reporting functions.

Concretely, I consider a general model of ordered response taking the form:

$$\begin{aligned} R_i &= r_i(H_i) = r(H_i, V_i) \\ H_i &= h_i(X_i) = h(X_i, U_i) \end{aligned} \tag{1}$$

where $H_i \in \mathbb{R}^K$ reflects a set of unobserved latent variables, and R_i an observed response mapped to a real number in some set $\mathcal{R}$. For example, $\mathcal{R} = \{0, 1\}$ for a binary yes/no question, or $\mathcal{R} = \{0, 1, 2, 3, 4\}$ for a question with five ordered response categories. I focus in the main text on the case of a scalar latent variable, later generalizing to $K > 1$.

The function $h_i(x)$ in (1) denotes the potential outcomes of the latent variable for individual i, indicating the value of H that would occur if a vector of observed explanatory variables X took each counterfactual value x. The function $r_i(h)$ represents individual i's

[1] A broad class of this type of survey questions use so-called *Likert scales*: e.g. allowing responses such as "strongly agree", "agree" ... "strongly disagree" to indicate agreement with a given statement, or to categorize quantities such as frequencies ("often", "sometimes", ... "almost never"). Hamermesh (2004) discusses the use of such outcomes in economics.

reporting function, which I assume to be weakly increasing in h for each i. The random vectors U_i and V_i parameterize heterogeneity across individuals in potential outcomes and reporting functions, respectively. The main statistical assumption of the model is that $X_i \perp\!\!\!\perp (U_i, V_i)$, which I relax to conditional independence given control variables. The researcher's objective is to learn how $h_i(x)$ varies with x, observing only R_i and X_i.

One of my key results is that if X_{1i} and X_{2i} reflect two continuously distributed components of the vector X_i, and $\mathcal{R}$ is associated with a set of integers, then

$$\frac{\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i]]}{\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i]]} = \frac{\tilde{\beta}_2}{\tilde{\beta}_1} \tag{2}$$

where $\tilde{\beta}_j$ reflects a convex weighted average across individuals of the causal effect of a small change in the j^{th} component of X on H. In particular, $\tilde{\beta}_j$ averages the causal partial derivative $\partial_{x_j} h(X_i, U_i)$ over individuals i who are on the margin between two response categories in $\mathcal{R}$.[2] If the conditional expectation $\mathbb{E}[R_i|X_i]$ happens to be linear, then the average derivative quantity $\mathbb{E}[\partial_{x_j}\mathbb{E}[R_i|X_i]]$ on the LHS of (2) is simply the coefficient on X_j in a linear regression of R on X. In this case, Eq. (2) affords a causal interpretation to the ratio of OLS regression coefficients for two continuous treatments.

Despite a growing trend in papers that leverage natural experiments with subjective outcome data,[3] empiricists have lacked formal results such as Eq. (2) to interpret precisely what is estimated by regressions in which subjectively-defined ordinal responses R are used as the dependent variable. This paper helps to fill the gap by showing that when the selection-on-observables research design is extended to include reporting-function heterogeneity, derivatives of the conditional expectation function of integer category numbers on X reveal positive aggregations of the local causal effects of X on H.[4] The weights in this aggregation have an intuitive form but are importantly not under the researcher's control. This illuminates the underidentification of *unweighted* means of causal effects, which correspond to the parameter analyzed by Bond and Lang (2019), when effects are heterogeneous. My results show that mean regression can nonetheless remain a useful tool for analyzing more general weighted averages of effects, without assuming cardinality or interpersonal comparability of H.

Throughout the paper, I discuss results through an application to survey questions that ask respondents about their overall satisfaction with life, and for ease of exposition refer to the latent variable H as "happiness".[5] For example, the popular Cantril Ladder question asks individuals to describe their satisfaction with life on an eleven point scale

[2] $\partial_{x_j} h(X_i, U_i)$ denotes $\partial_{x_j} h(x, U_i)$ with x_j the j^{th} component of x, evaluated at $x = X_i$ (and similarly for $\partial_{x_j}\mathbb{E}[R_i|X_i]$)

[3] Some prominent examples include Card et al. (2012), Benjamin et al. (2014), Lindqvist et al. (2020), Perez-Truglia (2020), and Dwyer and Dunn (2022).

[4] I also show that when the researcher is interested in establishing correlations rather than causation, the same estimands capture changes to the conditional quantile function of the underlying latent variable, without causal assumptions.

[5] This simplified language ignores e.g. distinctions between hedonic, affective and evaluative notions of well-being (Deaton, 2018; Helliwell and Barrington-Leigh, 2010).

from 0 to 10.[6] Questions like this about general well-being motivate treating the latent variable H as an outcome of normative interest, drawing on the notion of cardinal utility as a measure of welfare (Fleming, 1952; Harsanyi, 1955). With this interpretation, the marginal rates of substitution between treatment variables are a key input for welfare analysis, suggesting trade-offs that would be welfare improving for individuals. However, my results are also applicable to other outcomes elicited on ordered scales, e.g. general or mental health status, job satisfaction, product or service ratings, and other settings in which ordered response models might be employed with individual-specific heterogeneity in the thresholds between response categories.

I apply my formal results to revisit the influential study of Luttmer (2005), who considers the effects of household income as well as the incomes of one's neighbors on satisfaction with life. Using a selection-on-observables strategy and linear regression adjustment, Luttmer (2005) finds a positive coefficient on a household's own income along with a negative coefficient on mean income among their neighbors, suggesting that relative income comparisons are important for subjective well-being. My nonparametric identification results corroborate this interpretation under the maintained exogeneity assumptions, but without assuming cardinality or interpersonal comparability of individuals' responses to the well-being question. Regressions of $\mathbb{1}(R_i \leq r)$ on X for each r further suggest that differences in regression coefficients across the response distribution are driven by the unknown distribution of the underlying latent variable in this application, underscoring the theoretical observation that coefficients must be *compared* between variables to be quantitatively meaningful. I cannot reject equality across r of the marginal rates of substitution between own and neighbor income among respondents who are on the margin of choosing category r, and I estimate these "marginal" respondents overall to be similar to inframarginal respondents in terms of gender and education.

The identification results for continuous treatments that enable us to interpret the application above can be seen as limiting cases of a more general result that applies to discrete treatments as well. When a treatment variable of interest is discrete, I find however that comparisons of magnitude become more complicated. First, I show that when one compares the mean of R between two fixed values x and x' of the vector X:

$$\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x] = \mathbb{E}\left[\bar{f}_{x,x'}(\Delta_i, V_i) \cdot \Delta_i\right], \tag{3}$$

where $\Delta_i = h(x', U_i) - h(x, U_i)$ is the treatment effect of changing X from x to x' on outcome H for individual i. The "weight" $\bar{f}_{x,x'}(\Delta_i, V_i)$ is unknown but positive for all i, and Eq. (3) thus implies that if the sign of the treatment effect Δ_i is the same for all individuals, then the sign of $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ will be the same as that of the causal effect. This recovers a causal analog to the key point made by Bond and

[6] A prominent version of the Cantril ladder question asks: *Please imagine a ladder with steps numbered from zero at the bottom to ten at the top. Suppose we say that the top of the ladder represents the best possible life for you and the bottom of the ladder represents the worst possible life for you. If the top step is 10 and the bottom step is 0, on which step of the ladder do you feel you personally stand at the present time?* (Gallup, 2021).

Lang (2019), who argue that the conditional distributions $R_i|X_i = x'$ and $R_i|X_i = x$ are generally uninformative about the sign of $\mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x]$. Even when the treatment variables are randomly assigned, the sign of the unweighted average treatment effect $\mathbb{E}[\Delta_i]$ is not identified without strong assumptions.

Furthermore, the magnitude of the overall weight $\mathbb{E}\left[\bar{f}_{x,x'}(\Delta_i, V_i)\right]$ appearing in Eq. (3) can in general depend on the values x and x' being compared, and quantitative comparisons of regression coefficients can therefore be misleading if one or more of the treatment variables being considered is discrete and treatment effects are not small.[7] I describe how one can obtain bounds on the ratio of the total weight that the conditional expectation function applies to causal effects when comparing continuous to discrete variation in X. These analytic results suggest that when there are many response categories and individual reporting functions are approximately linear, discrete contrasts will tend to *overstate* causal effects relative to regression derivatives, by a factor that is upper bounded by two. I assess this implication through simulations and only find evidence of appreciable distortion when treatment effects are made implausibly large in the DGP.

I draw from my analysis three implications of my results for regression analysis using subjective ordinal outcomes. First, the focus on finding natural experiments popular in modern applied work yields a previously unrecognized benefit for subjective outcomes: reporting functions may also become uncorrelated with treatment variables of interest X, allowing researchers to draw conclusions about the signs of certain population averages of causal effects, without assuming interpersonal comparability of responses at the individual level, or taking R_i to be cardinally meaningful. Second, researchers can move beyond interpretations of the sign of effects and consider magnitudes only when multiple valid treatment variables are available. Third, these comparisons of magnitude are most informative when the two variables being compared are continuous rather than discrete. An implication is that identification in experimental work with subjective outcome variables would benefit from randomizing the doses of multiple treatments, in small increments.

2 Identification in a simplified model

To build intuition for the main results, let us begin with a simplified model in which all individuals share a reporting function, as assumed by previous work. I move to the general model in Section 3. The researcher observes (R_i, X_i), where R_i is an ordered response which we represent as an integer $R_i = 0, 1, \dots \bar{R}$ for some $\bar{R}$, and X_i is one or more observed covariates.

The researcher would like to interpret R as a proxy for some underlying continuous variable H, i.e. $R_i = r(H_i)$, where the function $r(\cdot)$ maps intervals of the real line into the $\bar{R} + 1$ discrete categories. Concretely, $r(h) = r$ for a given $r \in \{0, 1, \dots \bar{R}\}$ when h lies between $\tau(r-1)$ and $\tau(r)$, where we let $\tau(-1) = -\infty$ and $\tau(\bar{R}) = \infty$. The remaining

[7] The function $\bar{f}$ is defined in Sec. 6, and no longer depends upon Δ as $x' \to x$ and the difference becomes a derivative.

thresholds $\tau(r)$ are unknown to the researcher.

2.1 The signs of generic mean comparisons are not identified

Consider a quantity of interest taking the form

$$\theta := \mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x]$$

for some fixed covariate values x and x'. As an example, θ might represent the mean difference in "happiness" H among individuals i in two countries, x and x'.

Bond and Lang (2019) influentially show that even the *sign* of θ is generally not identified from the distribution of (R_i, X_i). This implies in particular that the sign of θ does not need to match that of the mean difference in observed responses $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$, a phenomenon often referred to as a sign "reversal" (Schröder and Yitzhaki, 2017). Bond and Lang (2019) argue that regressions of R_i on X_i are therefore generally uninformative about how the mean of H_i varies across subgroups of the population.

A special case that is highlighted by both Bond and Lang (2019) and Schröder and Yitzhaki (2017) is when the conditional distribution of H in the $X = x'$ group stochastically dominates that of the $X = x$ group (or vice versa).[8] In this case, the fact that θ is then positive will be revealed from the observable implication that $\mathbb{E}[R_i|X_i = x'] \geq \mathbb{E}[R_i|X_i = x]$. That is, one can conclude in this case that there is no sign reversal, despite not knowing the reporting function (i.e. the thresholds τ). However, since stochastic dominance of H between the groups cannot be verified empirically, neither Schröder and Yitzhaki (2017) nor Bond and Lang (2019) considers this to be an important special case for applied work. Indeed, when x and x' represent groups of individuals that differ from one another along many dimensions, there is generally no reason to expect either conditional distribution of H to stochastically dominate the other.

Though not assumed for my formal results, it is worth noting that stochastic dominance becomes a more natural condition in the context of studying the causal effects of X on H (this point is discussed further in Appendix A). Suppose that X_i represents one or more treatment variables that are randomly assigned. Let $h_i(x)$ denote potential outcomes with respect to counterfactual values x of X. Then by randomization, the distribution of H_i given $X_i = x$ is equal to the unconditional distribution of $h_i(x)$, and similarly the distribution of H_i given $X_i = x'$ is equal to the unconditional distribution of $h_i(x')$. The possibility of sign reversals now arises solely from heterogeneity in the sign of the treatment effect between x and x'.

For example, if one is willing to assume that the effect of switching X_i from x to x' has the same sign for all individuals, the sign of that effect is identified by $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$. However the magnitude of $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ is not directly informative about the average effect quantitatively, beyond establishing its sign.

[8] The distribution $H|X = x'$ stochastically dominates $H|X = x$ when $P(H_i \leq h|X_i = x') \leq P(H_i \leq h|X_i = x)$ for all h.

2.2 Intuition for identification results: ratios of effects in a linear model

With continuous X, we can establish a much stronger result. Suppose that X_i represents a vector of two continuous variables X_1 and X_2 that are statistically independent of potential outcomes. Assuming that treatment effects are entirely homogeneous across individuals i, we can identify not only the signs of the effects of X_1 and X_2 on H, but also the relative *magnitudes* of the two variables' effects.

Suppose the linear causal model $h_i(x) = x^T\beta + U_i$ holds, where U_i captures individual heterogeneity in potential outcomes. This corresponds to a setting in which the causal effect β_j of changing a single component x_j of the vector x by one unit is homogenous across units i, and is also the same across values of x. In this model, heterogeneity between individuals occurs through the additively separable scalar U_i. This model will serve merely to establish intuition before the more general analysis in which both causal effects and reporting functions are allowed to vary arbitrarily between individuals.

The natural parameters of interest in the linear model are the components β_j of β. Since we have made no assumptions that fix the scale of the random variable H, it will not be possible to identify β_j beyond an overall scale normalization. Therefore, we will instead aim to capture *ratios* of the parameters, e.g. β_2/β_1. The ratio β_2/β_1 captures the marginal rate of substitution of X_1 for X_2, if H is interpreted as a measure of utility.

Using the identity that $R_i = \sum_{r=0}^{\bar{R}-1} \mathbb{1}(r < R_i)$, it follows that the conditional mean is

$$\mathbb{E}[R_i|X_i = x] = \bar{R} - \sum_{r=0}^{\bar{R}-1} P(R_i \leq r|X_i = x).$$

Therefore, the regression derivative of R_i with respect to a small change in x_j is:

$$\begin{aligned}
\partial_{x_j}\mathbb{E}[R_i|X_i = x] &= -\sum_{r=0}^{\bar{R}-1} \partial_{x_j} P(R_i \leq r|X_i = x) = -\sum_{r=0}^{\bar{R}-1} \partial_{x_j} P(H_i \leq \tau(r)|X_i = x) \\
&= -\sum_{r=0}^{\bar{R}-1} \partial_{x_j} P(U_i \leq \tau(r) - x^T\beta|X_i = x) = -\sum_{r=0}^{\bar{R}-1} \partial_{x_j} P(U_i \leq \tau(r) - x^T\beta) \\
&= -\beta_j \cdot \left\{ \sum_{r=0}^{\bar{R}-1} f_U(\tau(r) - x^T\beta) \right\} \qquad (4)
\end{aligned}$$

where I have assumed that U is continuously distributed, and used that $U_i \perp\!\!\!\perp X_i$ by assumption. Notice that the sum in brackets does not depend on j in any way. Thus, provided that the regression derivative in the denominator is non-zero, we have that

$$\frac{\partial_{x_2}\mathbb{E}[R_i|X_i = x]}{\partial_{x_1}\mathbb{E}[R_i|X_i = x]} = \frac{\beta_2 \cdot \{\sum_{r=0}^{\bar{R}-1} f_U(\tau(r) - x^T\beta)\}}{\beta_1 \cdot \{\sum_{r=0}^{\bar{R}-1} f_U(\tau(r) - x^T\beta)\}} = \frac{\beta_2}{\beta_1}$$

In the special case of binary response ($\bar{R} = 1$) and with $U_i \sim \mathcal{N}(0, \sigma^2)$, note that Eq. (4) recovers the formula for marginal effects in the probit model: $\partial_{x_j} P(R_i = 1|X_i = x) =$

$\sigma^{-1}\phi(x^T\beta/\sigma)\cdot\beta_j$, where ϕ is the standard normal probability density function. While the scale of β cannot be identified in the probit model without fixing the value of σ, it cancels out when considering ratios of marginal effects, allowing β_2/β_1 to be identified.

Averaging back over the distribution of X_i, one can estimate β_2/β_1 nonparametrically, while making use of data at all values of X_i:

$$\frac{\mathbb{E}\left[\partial_{x_2}\mathbb{E}[R_i|X_i]\right]}{\mathbb{E}\left[\partial_{x_1}\mathbb{E}[R_i|X_i]\right]} = \frac{\beta_2}{\beta_1} \tag{5}$$

One of the main results of this paper is to show that Eq. (5) is not particular to a model with homogeneous and linear treatment effects, or to there being a common reporting function. Indeed, an analog of Eq. (5) holds as Eq. (13), where the coefficients β_j are replaced with convex weighted averages of the effect of a small change in x_j on H, among individuals who are marginal between any two response categories.[9]

3 A general model of ordered response

Let us continue to suppose that there is a meaningful latent value H_i which the researcher is ultimately interested in as an outcome. In the body of this paper I take H_i to be a (continuous) scalar, but Appendix C.3 extends results to the vector case.

3.1 Model setup

The researcher observes responses R_i generated as:

$$R_i = r_i(H_i) = r(H_i, V_i) \tag{6}$$

$$H_i = h_i(X_i) = h(X_i, U_i) \tag{7}$$

where $r_i(h)$ is in individual-specific function mapping happiness h to the space of possible responses $\mathcal{R}$, and X_i are observed covariates. Unless otherwise specified, I continue to let $\mathcal{R} = \{0, 1, \dots \bar{R}\}$ represent an ordered set of integers starting from zero or one, in line with common empirical practice. Results generalize if alternative numeric values are associated with the elements of $\mathcal{R}$, for example as described following Corollary 1.

The above model indexes heterogeneity in $r_i(\cdot)$ by a heterogeneity parameter $V_i \in \mathcal{V} \subseteq \mathbb{R}^{d_V}$. Since no constraints are placed on d_V, this is without loss of generality and the model is compatible with each individual having their own reporting function $r_i(h)$.

For each individual there is a function $h_i(\cdot)$ mapping values of a vector of J explanatory variables X into a value of H via (7), where heterogeneity in the function $h_i(\cdot)$ is represented by parameter $U_i \in \mathcal{U} \subseteq \mathbb{R}^{d_U}$. The primary interpretation of the function $h_i(x)$ is that it denotes potential outcomes for individual i as a function of counterfactual

[9] In the case of a common reporting function and U_i that enters additively, i.e. $h_i(x) = g(x) + U_i$, the simple argument establishing Eq. (4) already generalizes with $\partial_{x_j} g(x)$ replacing β_j. Treating the case with nonseparable U_i is more involved.

values of x, in some set of possible treatments $\mathcal{X} \subseteq \mathbb{R}^J$.[10] Since the dimension d_U is again left unrestricted and U_i may enter the function h nonseparably, the above model places no restriction on heterogeneity in potential outcomes or causal effects across individuals.

Note that (6)-(7) embed an exclusion restriction: X does not directly enter the equation for R, and only affects reports R through H. This is key for drawing inferences about the relationship between H and X from the observable joint distribution of R and X. In an earlier version of this paper (Goff, 2025), I show how the model can be generalized slightly to allow reporting behavior to depend directly on observables in a restricted way.

3.2 Identifying assumptions

The main assumptions I impose on the model are that reporting functions $r_i(\cdot)$ are weakly *increasing* in H_i, and that there exists variation in X that is conditionally independent of the latent individual heterogeneity U_i and V_i. I formalize these assumptions as follows.

Assumption MONO (weakly increasing reporting functions). *$r(h, v)$ is weakly increasing and left-continuous in h for all $v \in \mathcal{V}$*

Assumption EXOG (conditionally exogenous components of X). *For some set of observed variables W_i, we have i) $\{X_i \perp\!\!\!\perp V_i\} \mid W_i$; and ii) $\{X_i \perp\!\!\!\perp U_i\} \mid (W_i, V_i)$*

Appendix B shows how the general model with assumptions MONO and EXOG nests ordered response models from the literature, and how it relates to nonseparable outcome models from the instrumental variables literature.

The first part of Assumption MONO rules out cases in which individuals would report a *lower* value of R if H were *increased*. The additional assumption of left-continuity amounts to a simple normalization, since any weakly increasing function of bounded variation is continuous except at isolated points within its support.[11]

Lemma 1 shows that Assumption MONO is equivalent to the familiar notion of a set of "thresholds" $\tau_v(r)$ that separate the ordered categories in $\mathcal{R}$. Generalizing the model of Section 2 to allow reporting function heterogeneity, the thresholds now depend on v:

Lemma 1. *MONO holds iff for all $v \in \mathcal{V}, r \in \mathcal{R}$ and $h \in \mathcal{H}$:*

$$r(h, v) \leq r \iff h \leq \tau_v(r) \tag{8}$$

where $\tau_v(r) = \sup\{h \in \mathcal{H} : r(h, v) \leq r\}$ or $\tau_v(r) := \infty$ if the supremum does not exist.

All proofs are given in Appendix G. Lemma 1 implies that with $\mathcal{R} = \{0, 1, \ldots \bar{R}\}$, any

[10] An alternative interpretation of $h(x, u)$ is always also available and requires no causal assumptions, which is that h represents the conditional quantile function of H_i given X_i, with $U_i \in [0, 1]$ a scalar indicating i's rank in a distribution of their peers. This representation is helpful when causal effects are not the target, and the researcher is instead only interested in uncovering statistical features of the joint distribution between H_i and X_i. Details are given in Goff (2025).

[11] Hence a reporting function that is, say, right continuous rather than left continuous could be made left continuous by modifying the function on a set of Lebesque measure zero.

given reporting function $r(h, v)$ can be written as:

$$r(h,v) = \begin{cases} 0 & \text{if } h \leq \tau_v(0) \\ 1 & \text{if } \tau_v(0) < h \leq \tau_v(1) \\ 2 & \text{if } \tau_v(1) < h \leq \tau_v(2) \\ \vdots & \\ \bar{R} & \text{if } h > \tau_v(\bar{R}-1), \end{cases} \tag{9}$$

however Lemma 1 does not require $\mathcal{R}$ to be a set of consecutive integers.

The standard treatment of ordered response in which the researcher assumes a common set of thresholds represents the special case in which V_i is a degenerate random variable, corresponding to a single reporting function for all individuals.

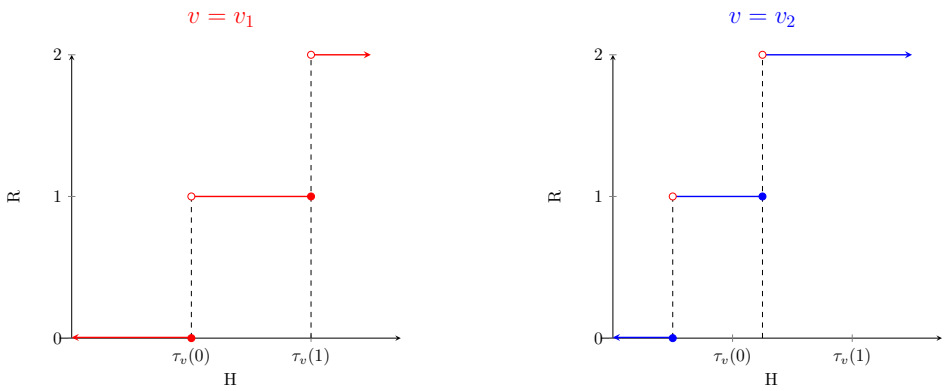


Figure 1: Examples of two different reporting functions, in a case with three categories: $\mathcal{R} = \{0, 1, 2\}$. The reporting function depicted in the right panel is more "optimistic" than the one in the left panel, as the threshold value of H for $R = 1$ and $R = 2$ are both lower than for the reporting function on the left (see Lemma 1).

It is worth emphasizing that Assumption MONO is compatible with individuals having direct preferences over the response categories r. For example, consider a utility maximization model in which $r(h, v) = \text{argmax}_{r \in \mathcal{R}}\, u(r, h, v)$, with utility u depending not only on happiness h, but also directly on the response category r. Let us further assume that the utility function takes the form $u(r, h, v) = \phi_v(r) - \alpha_v \cdot |h_v^*(r) - h|$ where individuals of type v obtain utility $\phi_v(r)$ from giving a response of r, but also value giving an answer close to a value $h_v^*(r)$ they perceive to correspond to response r. If the scale of ϕ is large enough relative to α_v , individuals' responses may be totally unaffected by changes in h (e.g. respondents that always choose 5 on a scale from 0 to 10). Provided that $h_v^*(r)$ is strictly increasing in r (i.e. higher categories are subjectively associated with higher values of happiness) and $\alpha_v > 0$, then u satisfies the property of *increasing differences* (cf. Milgrom and Shannon 1994) in (r, h), which in turn implies MONO.[12]

We now turn to the second identifying assumption, EXOG. A sufficient condition for EXOG is that there is (conditional) random variation in X, in the sense of being

[12]MONO is fully compatible with there being some individuals with preferences that *only* depend on r, giving the same response regardless of their H_i. Such individuals will not contribute to mean differences in R given X, under EXOG.

statistically independent of the latent heterogeneity (U_i, V_i) between individuals:

$$\{X_i \perp\!\!\!\perp (U_i, V_i)\} \mid W_i \tag{10}$$

Here $W_i \in \mathcal{W} \subseteq \mathbb{R}^{d_W}$ represents some vector of additional observed variables to be used as control variables in the analysis. The W_i can be thought of as components of U_i that are observed, and potentially correlated with the treatment variables of interest X_i. This allows the W_i to have a direct effect on happiness, though the W_i need not necessarily be "manipulable" in the typical causal sense (e.g. race). The W_i can also represent variables that affect both X_i and reporting function heterogeneity V_i. In settings with stratified experimental randomization, W_i isolates the strata.

Eq. (10) is technically stronger than the results require, but provides a natural foundation for EXOG and is simple to motivate. Even if an outcome H_i is directly observed, an assumption like $\{X \perp\!\!\!\perp U\}|W$ is frequently appealed to for the identification of causal effects, where some kind of experiment or natural experiment provides exogenous variation in X. Eq. (10) then simply requires this natural experiment to render X (conditionally) independent also of V. Note that under EXOG, U and V may be arbitrarily correlated with one another (e.g. if happier individuals have more optimistic reporting functions). In Appendix C.1, I relax EXOG to consider identification using instrumental variables.

3.3 Parameters of interest and preview of identification results

The function $h(x, u)$ is our main object of interest: how it varies holding u fixed yields the causal effect of a change in x on H. For example, $h(x', U_i) - h(x, U_i)$ is the "treatment effect" for unit i of moving between two counterfactual values x and x' of the vector X.

For most of the main results, I consider small changes in one or more components of x that are continuously distributed. The function $\partial_{x_j} h(x, U_i)$ for a given individual i characterizes the effect of a small change in the j^{th} component of X on H when $X = x$, which I refer to as the *marginal effect* of x_j on H when $X = x$. An average $\beta_j := \mathbb{E}[\partial_{x_j} h(X_i, U_i)]$ across all individuals i provides a population summary of this marginal effect. More generally, such averages can employ weights ρ_i that depend on the individual-level observables (X_i, W_i) and unobserved heterogeneity parameters (U_i, V_i). For example, for a given function $\rho(u, v, x, w) \geq 0$, we might consider a weighted average of the form:

$$\tilde{\beta}_j = \mathbb{E}[\rho_i \cdot \partial_{x_j} h(X_i, U_i)] \tag{11}$$

where $\rho_i := \rho(U_i, V_i, X_i, W_i)$ is positive with probability one and satisfies $\mathbb{E}[\rho_i] = 1$. I also use the notation $\tilde{\beta}_j$ for parameters that represent limits of $\tilde{\beta}_j$ for a sequence of such weighting functions $\rho(\cdot)$.[13] Section 4 shows that the sign of a parameter taking the form $\tilde{\beta}_j$ is identified under Assumptions MONO and EXOG, along with regularity conditions.

[13] For example, the average derivative $\mathbb{E}[\partial_{x_j} h(x, U_i) | h(x, U_i) = h]$ that conditions on a single value h for $h_i(x)$ represents the limit of $\tilde{\beta}_j$ for the function $\rho(U_i, V_i) = \frac{\mathbb{1}(h(x,U_i) \in [h, h+\epsilon])}{\mathbb{E}[\mathbb{1}(h(x,U_i) \in [h, h+\epsilon])]}$, as $\epsilon \to 0$. See also discussion after proof of Lemma 2.

Importantly, the weights ρ_i are not under the control of the researcher, but they are convex. As discussed above, sign *and magnitude* of a parameter taking the form $\tilde{\beta}_2/\tilde{\beta}_1$ is identified with two such X, where the same weighting ρ appears in both the numerator and the denominator.

Parameter	Identified?	Support restrictions	Sufficient conditions for identification
$sgn(\tilde{\beta}_1)$	Yes	One continuous X	
$\tilde{\beta}_2/\tilde{\beta}_1$	Yes	Two continuous X	
β_2/β_1	No	Two continuous X	Uniformly distributed reporting functions
$\widetilde{MRS}$	No	Two continuous X	Locally uncorrelated MRS or quasilinearity
$\mathbb{E}[MRS_i(x)]$	No	Two continuous X	Weakly separability
$sgn(\tilde{\Delta})$	Yes		
$sgn(\Delta)$	No		Stochastic dominance/common-sign of Δ_i
$\tilde{\Delta}/\tilde{\beta}_1$	No	One continuous X	Partial identification considered in section 6.2

Table 1: Summary of identification results in the general model (Sections 4 and 6).

If we interpret H as a measure of "utility", then $h_i(\cdot) = h_i(\cdot, U_i)$ can be thought of as i's utility function, and $H_i = h_i(X_i)$ as their realized utility (evaluated at i's actual X_i). Under this interpretation the *ratio* of two derivatives of $h_i(x)$ represents a local marginal rate of substitution of X_1 for X_2 when $X = x$, for individual i,[14] e.g.

$$MRS_i(x) := \frac{\partial_{x_2} h(x, U_i)}{\partial_{x_1} h(x, U_i)}$$

Weighted averages across individuals take the form $\widetilde{MRS} := \mathbb{E}\left[\rho_i \cdot MRS_i(X_i)\right]$ for ρ_i defined as following Eq. (11), or a limit of $\widetilde{MRS}$ for a sequence of such functions. Section 4.5 finds that such parameters are generally not identified under MONO and EXOG alone, if U_i is not a scalar. However, I give sufficient conditions to identify either MRS and $\widetilde{MRS}$ that involve functional assumptions on h.

Finally, this paper considers weighted averages of discrete *treatment effects* between two fixed values of X, i.e. $\Delta_i := h(x', U_i) - h(x, U_i)$ for some $x, x' \in \mathcal{X}$. Weighted averages of treatment effects take the form:

$$\widetilde{\Delta} := \mathbb{E}\left[\rho(U_i, W_i, X_i, W_i) \cdot \Delta_i\right]$$

with $\rho_i := \rho(U_i, V_i, X_i, W_i)$ as above, or the limit of $\widetilde{\Delta}$ for a sequence of such functions ρ. The overall average treatment effect $\Delta = \mathbb{E}[\Delta_i]$ (for a fixed x, x') is the causal analog of the parameter considered by Bond and Lang (2019). In Section 6 I show that while the sign of Δ is generally unidentified without the strong assumption of stochastic dominance between $H|X = x$ and $H|X = x'$, the sign of a parameter taking the form $\tilde{\Delta}$ is. Again, the weights ρ are convex but not under the control of the researcher.

Table 1 summarizes the main identification results of this paper regarding the parameters introduced in this section. The column labelled "Identified?" indicates whether

[14] Note that this interpretation only requires $h_i(\cdot)$ to represent utility in an ordinal sense: $MRS_i(x)$ yields the slope of the indifference curve for i that passes through the point x.

the parameter is point identified under the basic model (EXOG, MONO and regularity conditions). If the answer is No, then the rightmost column of Table 1 indicates additional restrictions that are sufficient to obtain identification. In the case of $\widetilde{MRS}$, these additional restrictions are fairly mild. For unweighted parameters such as $\mathbb{E}[MRS_i(x)]$ and Δ_i, the restrictions are more substantive.

4 Identification from continuous variation in X

Given the model outlined in the last section, let us consider what can be identified by looking at responses given continuous variation in X. Denote by $f_H(h|x,v,w)$ the density of H_i at h, conditional on $X_i = x$, $V_i = v$ and $W_i = w$. In this section, I suppose that at least one component of X is continuously distributed, and assume the following:

Assumption REG$_j$ (regularity conditions for X_j). *The following hold for given j: i) X_{ji} is continuously distributed; ii) $f_{H|XVW}(h|x,v,w)$ exists; iii) $\partial_{x_j} Q_{H|XVW}(\alpha|h,v,w) \leq M < \infty$ for all $\alpha \in [0,1]$, $h \in \mathcal{H}$, where $Q_{H|XVW}$ is the conditional quantile function of H given X,V,W; iv) for each x,w and h, $f_{H,\partial_{x_j}h(x,U)|XVW}(h,h'|x,v,w)$ exists and is upper bounded by some $c(h')$ where $\int c(h')|h'|dh' < \infty$, for all $v \in \mathcal{V}$.*

Assumption REG$_j$ reflects fairly standard regularity conditions, as described in Hoderlein and Mammen (2007). The only substantive modification added above is that I take the conditions to hold conditional on each reporting function type $V_i = v$.

4.1 Derivatives of the response distribution in terms of causal responses

Let $P(R_i \leq r|x,w) := P(R_i \leq r|X_i = x, W_i = w)$ denote the observed distribution of responses R_i given values x of treatments X_i and w of the control variables W_i. For brevity, I will often use this type of shorthand in long expressions.

The following Lemma is central to my analysis of identification with a continuous X:

Lemma 2. *Assume MONO and EXOG, and REG$_j$ for some $j \in \{1,\ldots,J\}$. Then:*

$$\partial_{x_j} P(R_i \leq r|x,w) = -\mathbb{E}\left\{ f_H(\tau_{V_i}(r)|x,V_i,w) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|H_i = \tau_{V_i}(r),x,V_i,w\right] \middle| W_i = w \right\}$$

Lemma 2 shows that the derivative of $P(R_i \leq r|X_i = x, W_i = w)$ with respect to changes in x_j provides a positively-weighted linear combination across individuals of the causal response in H to a small change in X_j, i.e. "marginal effects" as defined in Section 3.3.[15] In particular, Lemma 2 shows that $\partial_{x_j} P(R_i \leq r|x,w)$ can be interpreted in terms of average causal effects among individuals i who are themselves *marginal* between

[15] The inner expectation in Lemma 2 (square brackets [ ]) is over heterogeneity in causal effects U_i, while the outer expectation (curly brackets { }) is over heterogeneity V_i in reporting functions. Expanding the outer expectation, we have:

$$\partial_{x_j} P(R_i \leq r|x,w) = -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x,v,w) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|H_i = \tau_v(r),x,v,w\right] \tag{12}$$

The weights $dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x,v,w)$ that multiply the conditional expectation do not necessarily integrate to one—indeed all that we can say about $\mathbb{E}[f_H(\tau_{V_i}(r)|x,V_i,w)|W_i = w] = \int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x,v,w)$ is that it is positive.

response categories: that is $H_i = \tau_{V_i}(r)$ and a small increase in H_i would push them from category r to category $r+1$.[16] These local derivatives at a given x, w are identified by a nonparametric regression of $\mathbb{1}(R_i \leq r)$ on X_i and W_i, respectively.

Building on this result, we can show that if one averages these local regression derivatives across the observable distribution of X_i, W_i, one obtains an average causal effect that remains "local" to individuals on the margin between response categories, but is no longer specific to individuals having a particular value of X_i and W_i:

Theorem 1. *Under the assumptions of Lemma 2:*

$$\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i, W_i)] = -f_{H-\tau_V(r)}(0) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i, U_i) | H_i = \tau_{V_i}(r)\right]$$

The density $f_{H-\tau_V(r)}(0)$ is not identified by the data, but it does not depend on j. This unidentified density thus cancels out in ratios, so $\frac{\mathbb{E}[\partial_{x_2} P(R_i \leq r|X_i,W_i)]}{\mathbb{E}[\partial_{x_1} P(R_i \leq r|X_i,W_i)]} = \frac{\mathbb{E}\left[\partial_{x_2} h(X_i,U_i)|H_i=\tau_{V_i}(r)\right]}{\mathbb{E}\left[\partial_{x_1} h(X_i,U_i)|H_i=\tau_{V_i}(r)\right]}$.

4.2 Derivatives of the conditional expectation function

Beyond the case of binary survey questions, researchers do not typically estimate regressions of the response CDF evaluated at a fixed category r, as considered by Lemma 2. However, the result allows us to study the more common practice of modeling the conditional mean of R_i given X_i. Let $\tau_v := \{\tau_v(r)\}_{r \in \mathcal{R}}$ denote the set of all thresholds for individuals with reporting function v.

Using the identity $R_i = \sum_{r=0}^{\bar{R}-1} \mathbb{1}(r < R_i)$ as in Section 2, we have the following corollary to Theorem 1 that considers the conditional mean function, and generalizes Eq. (5):

Corollary 1. *If MONO and EXOG hold, and REG_j holds for $j = 1, 2$ then:*

$$\frac{\mathbb{E}[\partial_{x_2} \mathbb{E}[R_i | X_i, W_i]]}{\mathbb{E}[\partial_{x_1} \mathbb{E}[R_i | X_i, W_i]]} = \frac{\tilde{\beta}_2}{\tilde{\beta}_1} \tag{13}$$

where $\tilde{\beta}_j = \mathbb{E}[\partial_{x_2} h(X_i, U_i) | H_i \in \tau_{V_i}] := \sum_{r=0}^{\bar{R}-1} \omega_r \cdot \mathbb{E}[\partial_{x_j} h(X_i, U_i) | H_i = \tau_{V_i}(r)]$, and $\omega_r := \frac{f_{H-\tau_V(r)}(0)}{\sum_{r'=0}^{\bar{R}-1} f_{H-\tau_V(r')}(0)}$ are positive weights that sum to one.

In Corollary 1, response thresholds r that are more "populated" in the sense that $f_{H-\tau_V(r)}(0)$ is larger, receive higher weight, in such a way that $\tilde{\beta}_j = \mathbb{E}[\partial_{x_j} h(X_i, U_i) | H_i \in \tau_{V_i}]$. In the case with no control variables W_i, we then obtain Eq. (2) stated in the introduction.

If instead of a set of consecutive integers representing category numbers, the researcher associates alternative numerical values r_j with the ordered responses $\mathcal{R}$, where $r_0 < r_1 < \cdots < r_R$, then we have $R_i = r_0 + \sum_{j=0}^{R-1} (r_{j+1} - r_j) \cdot \mathbb{1}(r_j < R_i)$. The above results thus generalize with $f_H(\tau_v(r_j)|x, v, w)$ upweighted by the positive factor $(r_{j+1} - r_j)$. This implies that different labeling schemes could be used in estimation to achieve different weightings over local causal effects, though the maximum information would

[16] I thus use the term "marginal" in two distinct senses: i) a marginal effect as a derivative of $h(x, u)$; and ii) a marginal respondent who is at the threshold (i.e. "on the margin") between two response categories given their reporting function.

come from studying each threshold individually. When considering mean regression, using integer category labels is natural in that it weighs each threshold in proportion to its occupancy, delivering the simple and interpretable conditioning event $H_i \in \tau_{V_i}$ in Corollary 5. Numbering the categories $\mathcal{R}$ by consecutive integers is in this sense "special", among all other possibly non-linear numbering schemes $\{r_j\}$ that one might consider.

If the conditional mean function $\mathbb{E}[R_i|X_i = x, W_i = w]$ happens to be linear in x and w, then the quantity $\mathbb{E}[\partial_{x_j}\mathbb{E}[R_i|X_i, W_i]]$ on the LHS of Eq. (13) and (15) is simply the coefficient γ_j from the OLS regression

$$R_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_J X_{Ji} + \lambda^T W_i + \epsilon_i \tag{14}$$

where the vector of control variables W includes a constant. While specification 14 is common in empirical practice, Goff (2025) discusses the implications when this functional form is misspecified, i.e. when $\mathbb{E}[R_i|X_i, W_i]$ is not actually linear but the researcher proceeds in estimating (14) anyways. However, this issue is not specific to the use of subjective ordinal outcome variables, as it concerns a conditional mean function that relates observed variables only. Capturing the functional form of $\mathbb{E}[R_i|X_i, W_i]$ is a concern generally when selection-on-observables identification arguments are implemented via linear regression. In the empirical application, I compare OLS results with a flexible estimator.

4.3 Intuition for Lemma 2

The proof of Lemma 2 relates the derivative of the conditional CDF of R to a mixture of (infeasible) quantile regressions that condition on response type V_i, and then makes use of a connection between quantile regressions and local average structural derivatives (Hoderlein and Mammen, 2007; Sasaki, 2015).

By Eq. (12), the "weight" in the observable $\partial_{x_j} P(R_i \leq r|x, w)$ placed on an individual with happiness close to $\tau_v(r)$ is positive and proportional to $dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v, w)$. Figure 2 provides intuition for this particular weighting.

Suppose for simplicity there are no controls w. By the law of iterated expectations, we can write $\partial_{x_j} P(R_i \leq r|X_i = x)$ as a weighted average of $\partial_{x_j} P(R_i \leq r|X_i = x, V_i = v)$ across the various reporting functions v in the population. For a given v, $\partial_{x_j} P(R_i \leq r|X_i = x, V_i = v)$ captures the "flow" of individuals over the threshold $\tau_v(r)$ due to a small change in x_j, in one direction or the other. Some of these individuals could have negative effects: $\partial_{x_j} h(x, U_i) < 0$, denoted by arrows to the left in Figure 2. Others could have positive effects $\partial_{x_j} h(x, U_i) > 0$, indicated by rightward arrows in Figure 2. The net effect captured by $\partial_{x_j} P(R_i \leq r|X_i = x, V_i = v)$ depends on the average derivative $\mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = \tau_v(r), x, v\right]$ local to the threshold. Since the derivative ∂_{x_j} considers an infinitesimal change in X, any such "flow" over the threshold requires a positive density there: $f_H(\tau_v(r)|x, v) > 0$.[17]

[17] The quantity $f_H(h|x, v) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = h, x, v\right]$ at a given h is sometimes referred to as a "flow density", and appears in Kasy, 2022, Goff (2022) and in the physics of fluids, where it arises from the conservation of mass.

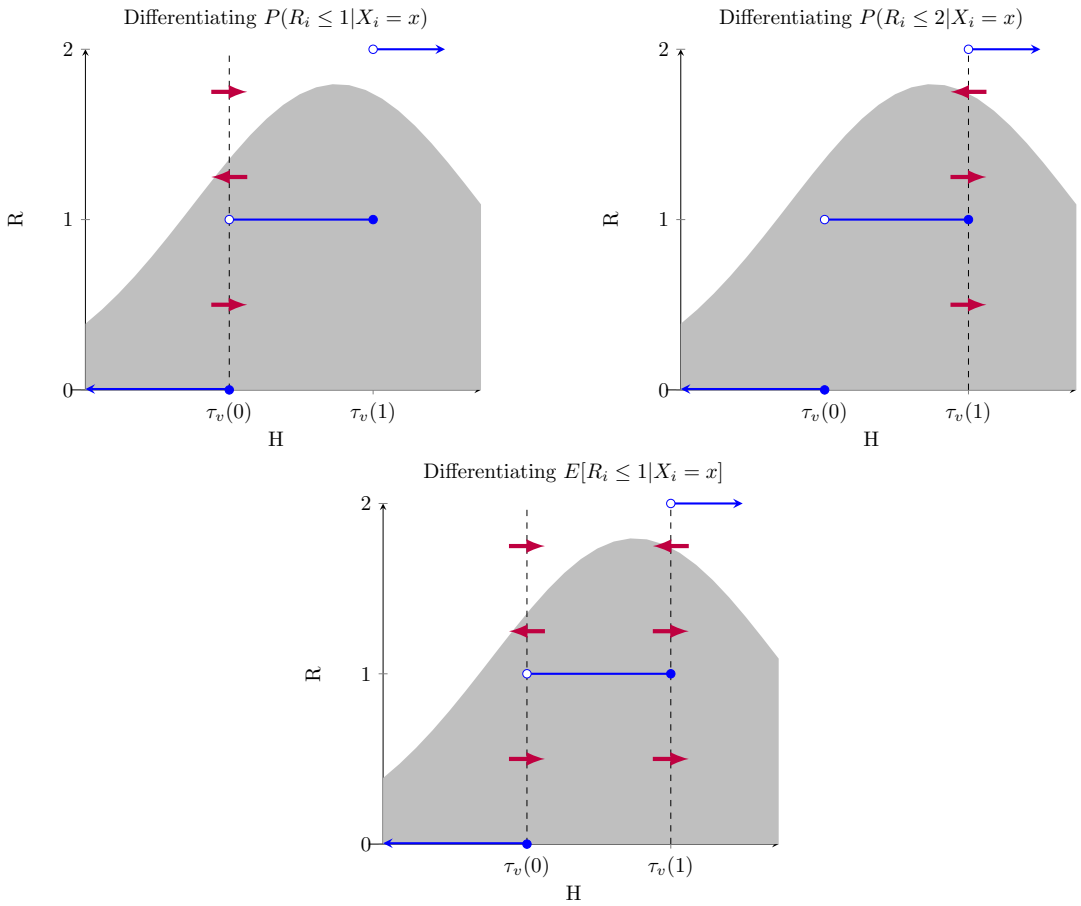


Figure 2: Intuition for Lemma 2 in a setting with $\mathcal{R} = \{0, 1, 2\}$: the derivative of $P(R_i \leq r|X_i = x)$ with respect to x_j captures the "flow" of individuals over threshold $\tau_v(r)$ due to a small change in x_j. Left: $\partial_{x_j} P(R_i \leq 1|X_i = x)$ captures flows over $\tau_v(0)$. Center: $\partial_{x_j} P(R_i \leq 2|X_i = x)$ captures flows over $\tau_v(1)$. Right: the derivative of $\mathbb{E}[R_i \leq r|X_i = x]$ with respect to x_j captures the "flow" of individuals over either threshold $\tau_v(0)$ or $\tau_v(1)$ due to a small change in x_j. The gray shaded curve in the background depicts the density of H_i.

I refer to individuals with $H_i = \tau_{V_i}(r)$ for some r as "marginal", or "indifferent" between response categories. Lemma 2 shows that local derivatives of the distribution of R_i conditional on X_i and R_i only average causal effects among these marginal respondents. These marginal respondents averaged over in the RHS of Lemma 2 cannot be individually identified, since neither H_i nor $\tau_{V_i}(r)$ are observed for a given i. However, I show in Appendix D.3 that if the sign of causal effects is assumed to be common across individuals, average characteristics of the marginal respondents can be identified (Section 5 provides an implementation).

4.4 Maximum reporting function heterogeneity would in fact be helpful

Ex-ante, it would seem that allowing for heterogeneity in reporting functions $r(\cdot, V_i)$ across individuals should make inferences about causal effects on H_i only harder, as compared with assuming a common mapping as in typical ordered response models. After all, heterogeneity in V_i precludes interpersonal comparisons of H_i between two individuals on the basis of their observed R_i (as illustrated in Figure 1). It is perhaps counter-intuitive,

then, that reporting function heterogeneity can in fact be *helpful* in drawing inferences about overall population averages of causal effects on H_i, when causal effects are also heterogeneous.

To see this, consider first the extreme case in which there is no reporting function heterogeneity, i.e. $\tau_{V_i}(r)$ is degenerate for each r. By Lemma 2, the observable derivative $\partial_{x_j} P(R_i \leq r|X_i = x)$ (omitting controls W, for simplicity) identifies an average causal effect among a very specific sub-population: those with $h(x, U_i) = \tau(r)$. The causal response among this specific group might be far from representative of the population mean of $\partial_{x_j} h(x, U_i)$, and may depend heavily on $\tau(r)$. As a concrete example, suppose that $h(x, u) = x_1 + e^{x_2+u}$, and that $X_{1i} \perp\!\!\!\perp X_{2i}$ with $\mathbb{E}[X_{1i}] = 0$, $\mathbb{E}[e^{X_{2i}}] = \mathbb{E}[e^{U_i}] = 1$. Then $\mathbb{E}[\partial_{x_2} h(X_i, U_i)] = \mathbb{E}[\partial_{x_1} h(X_i, U_i)] = 1$; the average marginal effect of either X_1 or X_2 on H is equal to unity. Furthermore, $\mathbb{E}[\partial_{x_2} h(X_i, U_i)/\partial_{x_1} h(X_i, U_i)] = 1$; the average marginal rate of substitution is also unity.[18]. Let observable responses be binary with $R_i = \mathbb{1}(H_i \geq \tau)$. Then $\mathbb{E}\left[\frac{\partial_{x_2} P(R_i=1|X_i)}{\partial_{x_1} P(R_i=1|X_i)}\right] = \tau$.[19] This may be puzzling, given that the value of τ is a property of the reporting function, and not of the causal response function h. However, the reporting function (captured by the value of τ) determines *for whom* causal effects are revealed by observed responses. The above example is engineered so that causal effects for the marginal sub-population happen to depend entirely on τ.

In the other extreme, "maximum" heterogeneity in reporting functions would occur if the response thresholds $\tau_{V_i}(r)$ for a given r were distributed uniformly across the real line (or a subset of it that contains all values of H_i in the population), rather than having a degenerate distribution at a single point. Corollary 2 below shows that if reporting function heterogeneity is furthermore independent of potential outcomes, then in this extreme $\partial_{x_j} P(R_i \leq r|X_i = x)$ in fact identifies the overall *unconditional* causal effect $\mathbb{E}[\partial_{x_j} h(x, U_i)|X_i = x]$ at a given x—rather than the average simply among individuals whose U_i and V_i make them marginal for response category r—up to a scale factor that does not depend on j. The ratio of unconditional effects $\beta_2/\beta_1 = \mathbb{E}[\partial_{x_2} h(X_i, U_i)]/\mathbb{E}[\partial_{x_1} h(X_i, U_i)]$ is then identified (and would be even in the pathological example described above).

Corollary 2. *Suppose that in addition to the assumptions of Lemma 2, i) $U_i \perp\!\!\!\perp V_i|W_i, X_i$ and ii) $\tau_{V_i}(r)|W_i$ is uniformly distributed on $[\ell_r, \mu_r]$ with supp$\{h(x, U_i)\} \subseteq [\mu_r, \ell_r]$. Then*

$$\mathbb{E}[\partial_{x_j} P(R_i \leq r|X_i, W_i)] = \mathbb{E}\left[\partial_{x_j} h(X_i, U_i)\right] \cdot \sum_r \frac{1}{\mu_r - \ell_r}$$

and thus $\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i, W_i]]/\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i, W_i]] = \mathbb{E}\left[\partial_{x_2} h(X_i, U_i)\right]/\mathbb{E}\left[\partial_{x_1} h(X_i, U_i)\right]$.[20]

Intuitively, Corollary 2 exploits that for each combination of U_i, X_i, W_i, there are individuals who are marginal between categories r and $r+1$, and that these marginal individuals

[18] To see this: $\partial_{x_2} h(x, u) = 1$ and $\partial_{x_2} h(x, u) = e^{x_2+u} = h(x, u) - x_1$, where $\mathbb{E}[e^{X_{2i}+U_i}] = \mathbb{E}[e^{X_{2i}}] \cdot \mathbb{E}[e^{U_i}] = 1$ by EXOG.

[19] By Lemma 2: $\frac{\partial_{x_2} P(R_i=1|X_i=x)}{\partial_{x_1} P(R_i=1|X_i=x)} = \frac{\mathbb{E}[h(X_i,U_i)-X_{1i}|h(X_i,U_i)=\tau, X_i=x]}{1} = \tau - x_1$, then use that $\mathbb{E}[X_{1i}] = 0$.

[20] Corollary 2 can be generalized to allow $\tau_{V_i}(r)$ to be correlated with W_i provided that it remains uniformly distributed conditional on W_i, with $\mathbb{E}\left[\partial_{x_2} h(X_i, U_i)\right]/\mathbb{E}\left[\partial_{x_1} h(X_i, U_i)\right]$ still identified if $u_{W_i,r} - \ell_{W_i,r}$ is independent of U_i.

have the same distribution of $\partial_{x_j} h(x, U_i)$ among all of those with that X_i, W_i. This is guaranteed by independence of V_i and U_i, and uniformity of $\tau_{V_i}(r)$, conditional on X_i, W_i.

Since it is plausible to expect considerable heterogeneity in reporting functions, Corollary 2 is suggestive that the causal effects identified by observable responses may be somewhat representative of effects for the population at large. However, the conditions required for Corollary 2 to hold exactly are restrictive and cannot be verified empirically. In practice, it can be useful to gather suggestive evidence about the representativeness of the marginal respondents in terms of observable characteristics. Proposition 3 in Appendix D.3 shows that mean characteristics among marginal respondents can be identified from the data, under additional assumptions. I implement this in the empirical study of Section 5.

4.5 Marginal rates of substitution

Corollary 1 to Theorem 1 shows that a ratio of average regression derivatives identifies the ratio of a conditional average causal effect of X_2 on H to the same conditional average of the effect of X_1 on H. Luttmer (2005) and Di Tella et al. (2001) represent two prominent empirical studies in which the relative magnitude of regression coefficients (with subjective well-being as the dependent variable) is interpreted as yielding the implicit trade-off between two goods. I now investigate this interpretation under general heterogeneity in causal effects and reporting functions.

In general, a ratio of averages is not the same as an average of ratios, and thus Eq. (13) does not immediately yield an average marginal rate of substitution parameter $\widetilde{MRS}$ of the form introduced in Section 3.3. A sufficient condition however is that $MRS_i(X_i)$ be uncorrelated with $\partial_{x_1} h(X_i, U_i)$, among marginal respondents.

Example 1 (MRS uncorrelated with the base effect). *Suppose that*

$$Cov\left(MRS_i(X_i), \partial_{x_1} h(X_i, U_i) \middle| H_i \in \tau_{V_i}\right) = 0.$$

Then

$$\frac{\mathbb{E}[\partial_{x_2} \mathbb{E}[R_i|X_i, W_i]]}{\mathbb{E}[\partial_{x_1} \mathbb{E}[R_i|X_i, W_i]]} = \mathbb{E}\left[MRS_i(X_i) \middle| H_i \in \tau_{V_i}\right], \tag{15}$$

capturing the average marginal rate of substitution between X_1 and X_2, among respondents who are marginal at any threshold, i.e. $H_i = \tau_{V_i}(r)$ for some r. This covariance condition says that heterogeneity in $MRS_i(X_i)$ across individuals is uncorrelated with heterogeneity in the magnitude of the marginal effect of X_1 alone.

In Appendix D.2, Proposition 2 shows how a similar result to Eq. (15) holds using the estimand $\frac{\partial_{x_2} \mathbb{E}[R_i|X_i=x, W_i=w]}{\partial_{x_1} \mathbb{E}[R_i|X_i=x, W_i=w]}$ which fixes a value of x. In this case note that variation in $MRS_i(x)$ conditional on H_i and X_i comes from U_i alone. Thus if U_i is degenerate conditional on X_i and the value of H_i (e.g. if h is invertible in a scalar u), then the needed covariance condition holds automatically.

Proposition 2 of Appendix D.2 also shows how one can obtain a one-sided bound on the RHS of (15) by relaxing this to assume a known *sign* of the correlation between $MRS_i(X_i)$ and $\partial_{x_1} h(X_i, U_i)$.

The argument made in Example 1 above is fundamentally a statistical one, in the general case where U_i is multi-dimensional. An alternative approach is to impose some structure on the function $h(x, u)$ that is sufficient to interpret $\frac{\partial_{x_2} \mathbb{E}[R_i|X_i=x,W_i=w]}{\partial_{x_1} \mathbb{E}[R_i|X_i=x,W_i=w]}$ as a marginal rate of substitution, and the ratio of averages of such derivatives as an average of such marginal rates of substitution. I give two examples. The first is a weakly separable structural function:

Example 2 (weakly separable special case). *We say that the potential outcomes function $h(x, u)$ is weakly separable between x and u when*

$$h(x, u) = \mathtt{h}(g(x), u), \tag{16}$$

i.e. some function $g : \mathcal{X} \to \mathbb{R}$ aggregates over the treatments X into a scalar $g(x)$, which is then combined through $\mathtt{h}$ with heterogeneity u in a way that may or may not be additively separable.[21] *When (16) holds, Lemma 2 yields*

$$\frac{\partial_{x_2} \mathbb{E}[R_i|x, w]}{\partial_{x_1} \mathbb{E}[R_i|x, w]} = \frac{\partial_{x_2} g(x)}{\partial_{x_1} g(x)} = \frac{\partial_g h(g(x), U_i) \cdot \partial_{x_2} g(x)}{\partial_g h(g(x), U_i) \cdot \partial_{x_1} g(x)} = \frac{\partial_{x_2} h(x, U_i)}{\partial_{x_1} h(x, U_i)} = MRS_i(x) \tag{17}$$

i.e. $\partial_{x_2} \mathbb{E}[R_i|x, w] / \partial_{x_1} \mathbb{E}[R_i|x, w]$ recovers $MRS_i(x)$ (which is common among to all i), provided that it is not infinite. Eq. (17) simplifies greatly because the derivatives of $g(x)$ do not depend on v, and is derived in Section D.2.[22] *From (17) we have that* $\mathbb{E}[MRS_i(X_i)] = \mathbb{E}\left[\frac{\partial_{x_2} \mathbb{E}[R_i|X_i,W_i]}{\partial_{x_1} \mathbb{E}[R_i|X_i,W_i]}\right]$.

One particular variety of a weakly separable model is a model with an additive scalar error term $\mathtt{h}(g, u) = g + u$. This is equivalent to imposing that the causal effect of changing between any treatment values x and x' is the same for all individuals. If furthermore $g(x) = x^T \beta$, then we obtain the linear model of Section 2. To my knowledge, this common-effects structure has not explicitly been relaxed in existing work.

Another special case is when H represents utility and preferences are quasilinear:

Example 3 (quasilinear special case). *Suppose that h represents preferences and for each individual, these preferences are quasi-linear in X_1 such that $h(x, u) = x_1 + h(x_2, \dots x_J, u)$. Quasi-linear utility is widely used in economics to simplify welfare analysis (see e.g. Feng and Lee 2025).*[23] *When the elements of X are priced, quasilinearity*

[21] Weakly separable models for ordered response in which u is a scalar have been studied by Matzkin (1994). Appendix D.1 discusses how the above result, which does not require u to be a scalar, relates to that body of work.

[22] From (17), a testable implication of the weakly separable model is that $\partial_{x_2} \mathbb{E}[R_i|x, w] / \partial_{x_1} \mathbb{E}[R_i|x, w]$ does not depend on the value of the controls w. Another implication is that $\partial_{x_2} P(R_i \leq r|x, w) / \partial_{x_1} P(R_i \leq r|x, w)$ does not depend on r.

[23] Although quasi-linearity is a property of *preferences* and not a particular utility representation of them, $h(x, u) = x_1 + h(x_2, \dots x_J, u)$ yields a cardinalization of ordinal preferences in which a unit increase in x_1 has the same effect on mean utility regardless of which individual it accrues to. Under this normalization, $\mathbb{E}[h(x, U_i)]$ for example represents a utilitarian social welfare function whose value is unaffected by transfers of x_1 between individuals.

in X_1 can also deliver demand functions for the remaining goods that do not depend on income (see e.g. Nocke and Schutz 2017).

In this case the condition $Cov\left(MRS_i(X_i), \partial_{x_1} h(X_i, U_i) \middle| H_i \in \tau_{V_i}\right) = 0$ *is satisfied trivially, because* $\partial_1 h(x, U_i) = 1$ *with probability one. Thus we have that* $\frac{\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i,W_i]]}{\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i,W_i]]} = \mathbb{E}\left[MRS_i(X_i) \middle| H_i \in \tau_{V_i}\right]$. *Further* $\frac{\partial_{x_2}\mathbb{E}[R_i|X_i=x,W_i=w]}{\partial_{x_1}\mathbb{E}[R_i|X_i=x,W_i=w]} = \mathbb{E}\left[MRS_i(x) \middle| H_i \in \tau_{V_i}, X_i = x, W_i = w\right]$, *which can be derived as a special case of Proposition 2 given in Appendix D.2.*

5 Empirical illustration

In a prominent paper, Luttmer (2005) studies the effects of absolute and relative income on life satisfaction, investigating whether individuals draw on social comparisons in assessing their personal well-being. To do so, Luttmer (2005) merges data from the 1987 and 1992 waves of the U.S. National Survey of Families and Households (NSFH)—which contains a question on self-reported satisfaction with life along with self-reported socioeconomic data—to information on the local average earnings for a given household constructed from the Current Population Study and the 1990 Census.

Let i denote the primary respondent of an individual household in the NSFH. We consider two treatment variables $X_i = (X_{1i}, X_{2i})$, where X_{1i} denotes the log of household income for i's household (self-reported in the NSFH) and X_{2i} denotes average predicted log earnings in the Public Use Microdata Area (PUMA) in which i lives. The construction of this variable is described in detail in Luttmer (2005). R_i denotes i's response to the question "taking things all together, how would you say things are these days?", reported on a one to seven Likert-type scale in which a response of one indicates "very unhappy" and seven "very happy".[24] Finally, W_i represents a vector of control variables that includes home size/type/value, employment, education, gender, marriage, race religion, state fixed effects and PUMA characteristics.

I follow Luttmer (2005) and focus on households in which the main respondent was married in both waves of the NSFH. Details on my sample construction are provided in Appendix F. While I let i denote the main respondent for a household, the primary regression specification of Luttmer (2005) averages values of R_i and W_i between the main respondent and their spouse, finding very similar results. I focus on the individual-level specification for two reasons: i) the main respondent and their spouse may have different reporting functions, and the interpretation of a "marginal" respondent developed in Section 4 is more informative with a single reporting function; and ii) I do not focus solely on linear models, where averaging across observations within a household does not affect the functional form of the regression.

[24] The intermediate values 2-6 do not have associated descriptions in the survey (e.g. "somewhat happy"), and are labeled by integers only.

5.1 Basic result interpreted through the lens of Lemma 2

The main results of Luttmer, 2005 exploit a selection-on-observables strategy, estimating an OLS regression of R_i on X_i and W_i, i.e. Eq (14):

$$R_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \lambda^T W_i + \epsilon_i \tag{18}$$

and ascribing a causal interpretation to the coefficients γ_1 and γ_2. Luttmer uses fixed effects regressions as well as data on movers between PUMAs to argue that selection due to neighborhood choice is not a major concern in this context. Luttmer further argues that individuals' definitions of "very happy" or "very unhappy" are not affected by X, by replicating the qualitative results with other outcome variables that are expected to be less prone to this threat. I refer the reader to sections IV.B and IV.C of Luttmer (2005) for details. These arguments motivate Assumption EXOG in this context.

Luttmer finds that an increase in household earnings increases subjective well-being $\gamma_1 > 0$, while an increase in the earnings of one's neighbors decreases subjective well-being $\gamma_2 < 0$. This provides evidence that well-being is influenced not only by one's absolute income, but also one's relative income compared with the reference group of one's neighbors.[25] In OLS estimates of (18), the positive coefficient on own income has about half the magnitude as the negative coefficient on PUMA (neighbors') income.[26] That is, if one's PUMA were to go up by 1%, one's own income would need to go up by about 2% to leave the respondents' well-being unaffected.

	(1)	(2)	(3)
	Luttmer (2005) Table 1	OLS	DML
Own income	0.111***	0.0877***	0.0910***
	(0.0240)	(0.0232)	(0.0167)
PUMA income	-0.248**	-0.229**	-0.192**
	(0.0830)	(0.0840)	(0.0665)
Ratio PUMA/own	-2.234	-2.614	-2.110
se(ratio)	.	(1.160)	(0.904)
Controls	X	X	X
Clustering by PUMA	X	X	X
Sample size	8023	7939	7939

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 2: Replication of Luttmer (2005)'s results for the main respondent, and alternative de-biased machine learning estimator. For this column, the first two rows report average local derivatives, and "Ratio" measures the ratio of average derivatives. See Appendix Table 3 for estimates without controls (OLS and kernel regression).

I confirm this finding qualitatively in Column (1) of Table 2. Column (2) reports the numerical results published in Table 1 of Luttmer (2005) (main respondent column), in which $\hat{\gamma}_2/\hat{\gamma}_1 = -2.23$. In Column (2) I implement regression (18) on the publicly available NSFH data merged with the PUMA income variable from Luttmer (2005) (see

[25] This finding has since been replicated using experimental variation in beliefs about relative income (Rooij et al., 2024).

[26] Luttmer (2005) also reports estimates that instrument for own-income to overcome potential measurement error. I focus on magnitudes from the benchmark OLS regression (18).

Appendix F). I obtain similar results in both sign and magnitude, with $\hat{\gamma}_1 = 0.0877$ and $\hat{\gamma}_2 = -0.229$ for a ratio of $\hat{\gamma}_2/\hat{\gamma}_1 = -2.614$.[27] If $\mathbb{E}[R_i|X_i, W_i]$ is indeed a linear function of X_i and W_i, then the OLS coefficients γ_j from regression (18) capture $\mathbb{E}[\partial_{x_j}\mathbb{E}[R_i|X_i, W_i]]$, and by Corollary 1 the ratio γ_2/γ_1 captures a ratio of the average effect of a small increase in PUMA income over individuals whose happiness is exactly at the threshold between any two response categories, to the average effect of a small increase in own income for that same group of individuals.

If $\mathbb{E}[R_i|X_i, W_i]$ on the other hand is not linear in fact in X_i and W_i, then OLS estimates of Eq. (18) are not guaranteed to be interpretable in terms of causal effects, even if the assumptions of Corollary 1 do hold. However, since (R_i, X_i, W_i) are all observed one can instead employ a flexible estimator that does not require $\mathbb{E}[R_i|X_i, W_i]$ to be linear. Although $\mathbb{E}[R_i|X_i = x, W_i = w]$ is a nonparametric function, the average derivatives $\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i, W_i]]$ and $\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i, W_i]]$ can be estimated at the $\sqrt{n}$ rate.

Column (3) of Table (1) employs the debiased machine-learning (DML) average derivative estimator of Chernozhukov et al. (2022), as adapted by Klosin and Vilgalys (2023), to estimate the $\mathbb{E}[\partial_{x_j}\mathbb{E}[R_i|X_i, W_i]]$ by Lasso. In line with Luttmer (2005), I cluster standard errors at the PUMA level. To facilitate this, I use a nonparametric cluster bootstrap to estimate standard errors (based on 500 iterations). To ease the computational burden in the bootstrap, I impose some separability between some components of W_i in the regression function in implementation, but allow a flexible functional form between X_1 and X_2. Details are provided in Appendix F.2. For Column (3), the row labeled "Ratio PUMA/own" reports the DML estimates of $\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i, W_i]]/\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i, W_i]]$ in Column (3). Overall, the estimates of average derivatives and their ratio are numerically fairly similar to the $\hat{\gamma}_j$ reported by Luttmer (2005) from the OLS specification (18). For comparison, Appendix F.2 also reports estimates employing a fully nonparametric kernel-regression approach without control variables W_i. This also yields similar estimates for the average derivatives.

5.2 Decomposing mean effects by response category

While we know by Corollary 1 that the regression derivatives reported in Table 2 average over respondents who are on the margin between two adjacent response categories, we also know from Theorem 1 that we can isolate causal effects for respondents that are on a single such margin r and $r+1$, for some $r \in \{1, 2, \ldots 6\}$.

The top two panels of Figure 3 report coefficients from a linear probability model that takes, for a given r, the conditional expectation function $\mathbb{E}[\mathbb{1}(R_i \leq r)|X_i = x, W_i = w] = P(R_i \leq r|X_i = x, W_i = w)$ to be linear in X_i and W_i, with coefficients $(\gamma_{1r}, \gamma_{2r}, \lambda_r)$ specific to that response category r, i.e. $\mathbb{1}(R_i \leq r) = \gamma_{1r}X_{1i} + \gamma_{2r}X_{2i} + \lambda_r^T W_i + \epsilon_{ri}$ with $\mathbb{E}[\epsilon_{ri}|X_i, W_i] = 0$. The specification is otherwise identical to Column (2) in Table 2.

[27] That I am not able to match the numerical results exactly is likely explained by the many choices involved in how exactly to define some of the control variables, or possible updates to the underlying NSFH data over the last two decades.

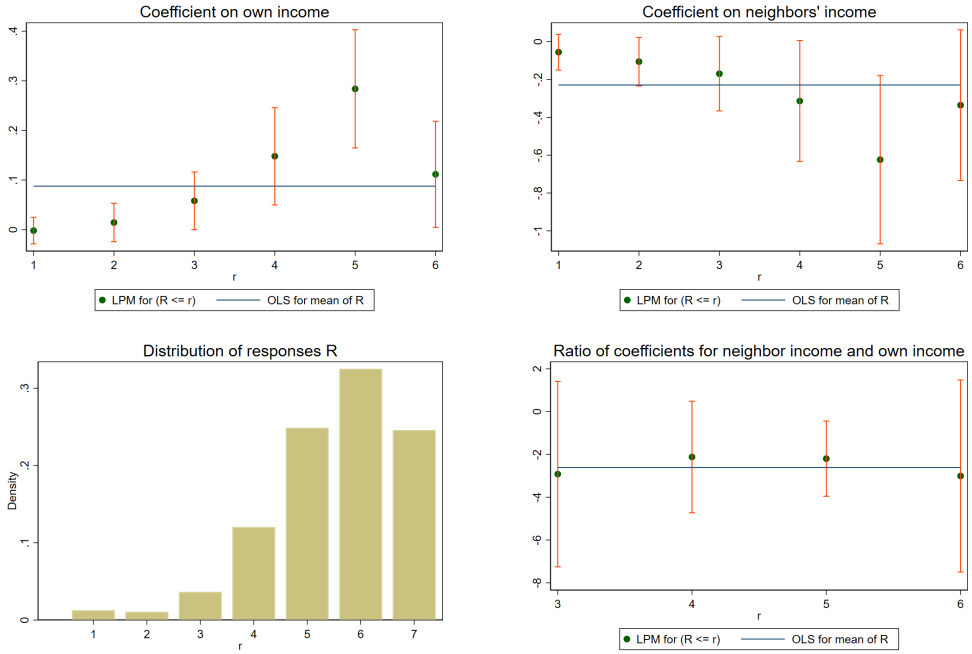


Figure 3: Visualization of the OLS estimates of γ_{1r} (top-left), γ_{2r} (top-right), the ratio γ_{2r}/γ_{1r} (bottom-right) from the regression $\mathbb{1}(R_i \leq r) = \gamma_{1r}X_{1i} + \gamma_{2r}X_{2i} + \lambda_r^T W_i + \epsilon_{ri}$ for $r \in \{1, 2, \dots 6\}$, along with a histogram of the response categories $r \in \{1, \dots 7\}$. The horizontal line in the upper panels and bottom right panel depicts the corresponding value from mean regression (see Table 2).

The plots reveal that the sign of $\hat{\gamma}_{1r}$ is positive for all r when it is statistically significant, the sign of $\hat{\gamma}_{2r}$ is consistently negative when it is statistically significant, and the ratio $\hat{\gamma}_{2r}/\hat{\gamma}_{1r}$ never differs from the "aggregate" value of -2.614 recovered by mean regression in a statistically significant way. An F-test of equality of γ_{2r}/γ_{1r} across all r fails to reject (p-value: 0.98). See Appendix F for estimates in table form.

By Theorem 1, the hump-shaped patterns observed in the top two panels of Figure 3 could be explained by either of two factors, provided that the linear probability model is correctly specified: (i) heterogeneity in the mean causal effect among the individuals at each of the thresholds r; and (ii) by differences in the density of individuals at each threshold, i.e. $f_{H-\tau_V(r)}(0)$. Appendix F.4 shows that a histogram of R_i can approximate how these densities vary across the thresholds. The bottom-left panel of Figure 3 depicts $P(R_i = r)$ and reveals that it does indeed capture the same basic pattern as γ_{1r} and mirrored by γ_{2r}. Indeed, we see in the bottom-right panel that the pattern cancels out nearly exactly, so that γ_{2r}/γ_{1r} is roughly constant across r (categories 1 and 2 are omitted because they are very imprecisely estimated).

Overall, the strong similarity in the shapes of the first three panels of Figure 3 are suggestive that the differences in $\hat{\gamma}_{1r}$ and $\hat{\gamma}_{2r}$ are driven by the underlying latent density of happiness, rather than by heterogeneity in causal effects across the happiness distribution. This is consistent, for example, with a simple constant-effects model in which $\gamma_{2r}/\gamma_{1r} = \beta_2/\beta_1$, or more generally by a weakly-separable model of the form $h(x, u) = \mathtt{h}(g(x), u)$.

5.3 Who are the marginal respondents?

Figure 4 compares the gender balance and education of respondents that on the margin between categories r and $r+1$, for each r, with that of the population as a whole. These comparisons are based on Proposition 3 in Appendix D.3, which leverages additional assumptions to identify averages of an attribute A_i among marginal respondents.

The upper panels of Figure 4 report estimates of $\mathbb{E}[A_i|H_i = \tau_{V_i}(r)]$, under an assumption that $\{X_i \perp\!\!\!\perp (A_i, U_i, V_i)\}|W_i$ and imposing the additional restriction that the sign of the effect of household income on happiness is the same for all units (not that this assumption is *not* imposed for the main results). The implementation takes the conditional expectation of $A_i \cdot \mathbb{1}(R_i \leq r)$ to be linear in x and w and assumes a linear probability model for $\mathbb{1}(R_i \leq r)$ (see Appendix D.3 for details).

In particular, the top left panel displays 95% confidence intervals for $\mathbb{E}[A_i|H_i = \tau_{V_i}(r)]$ versus $r \in \{2, 3 \dots 6\}$ when A_i is taken to be an indicator for the main respondent i attending college.[28] The horizontal line (orange) depicts the overall sample mean (an estimate of $\mathbb{E}[A_i]$). For none of the margins r can we reject the null hypothesis that the average rate of college among marginal respondents for that category is the same as the overall population mean. A similar result appears in the top-right panel, in which this calculation is repeated with A_i equal to i's years of education. There is some evidence that individuals on the margin on categories four and five out of seven have fewer years of education than the average. This is consistent with the finding of Barrington-Leigh (2024) that lower-education individuals are more likely to "bunch" at focal points in the response space $\mathcal{R}$, for example the midpoint (which is indeed 4 on the 1 to 7 scale).

The bottom panels of Figure 4 exploit the identification of the relative odds for a binary A_i, comparing marginal respondents to the population as a whole–see Eq. (33) in Appendix D.3 for an explicit expression. This result makes use of a weaker assumption in Proposition 3 that only assumes that A_i would represent a valid control variable to add to W_i. In this case all that is required is to implement regressions of $\mathbb{1}(R_i \leq r)$ on x and w separately by subsample defined by A_i.

The bottom panels report 95% confidence intervals for this ratio of odds, with the horizontal line (orange) depicting unity (equal odds in both populations). The bottom-right panel sets A_i to be an indicator for the main respondent being female, and compares the relative odds of being female among marginal respondents to the population overall. None are statistically different from unity. For clarity, the confidence interval for $r = 3$, which is very large, is not shown.

Overall, the results of this section indicate there is mild evidence that marginal respondents have somewhat less education than the overall population, for the central category in the response space. No differences are detected across gender. The rightmost confidence interval in each panel of Figure 4, labeled "Avg", replaces indicators for $R_i \leq r$

[28] Confidence intervals for $r = 1$ are dropped in all panels of Figure 4 for visibility, as the standard error is much larger than for other r.

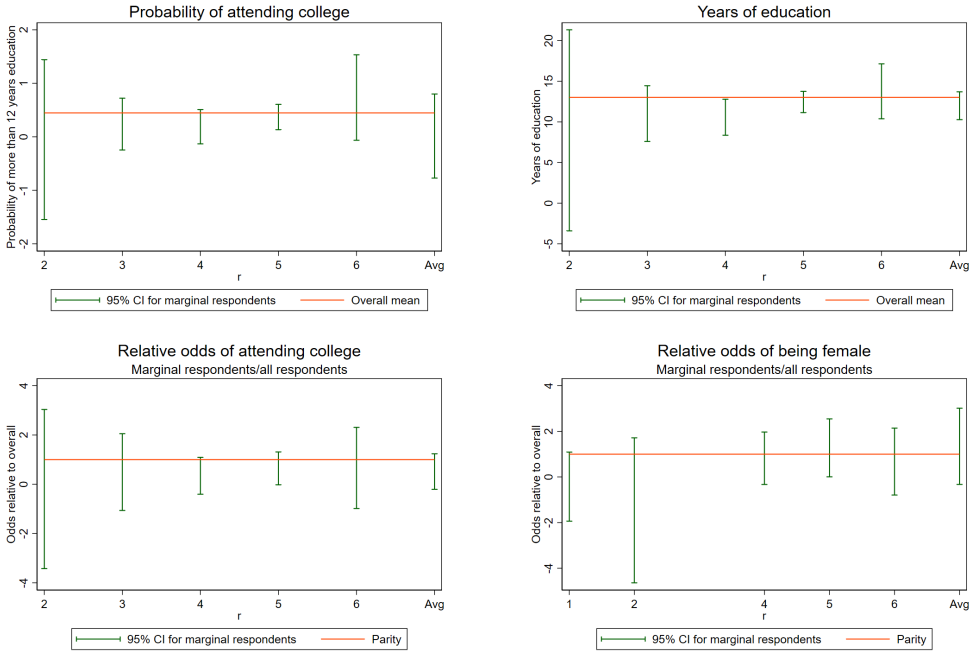


Figure 4: Attributes of marginal respondents, main respondent data. See text for details.

with R_i to approximate an "average" comparison considering all of the response categories at once. In all cases, we do not find any evidence that the marginal respondents overall differ from the infra-marginal respondents in education or gender.

6 Identification from discrete variation in X

The analysis thus far has considered what is identified by examining how the conditional distribution of R changes over infinitesimal differences in X. This section now considers taking discrete differences in treatment values (nesting the results thus far in the limit of small changes). I find that differences in the distribution of R over discrete changes in X can again be interpreted causally, and identify the sign of causal effects if those effects have the same sign across individuals. However, unlike the case with continuous treatments, magnitudes cannot be quantitatively compared between regressors, absent further assumptions. Discrete treatment variables are prevalent in practice, so this highlights a limitation of what is learned from subjective outcomes even in e.g. experiments with two treatment arms.

6.1 Identifying the signs of convex averages of treatment effects

Consider any two fixed values x and x', and define $\Delta_i := h(x', U_i) - h(x, U_i)$ to be the treatment effect of moving from $X_i = x$ to $X_i = x'$ for unit i. Further, let $f_H(y|\Delta, x, v, w)$ denote the density of H_i conditional on $\Delta_i = \Delta$, $X_i = x$,$V_i = v$ and $W_i = w$. As before, let $P(R_i \leq r|x, w)$ denote a shorthand for $P(R_i \leq r|X_i = x, W_i = w)$. The following expression shows what is identified from the conditional distribution of R_i across this

discrete change between values x and x':

Theorem 2. *Under MONO and EXOG:*

$$P(R_i \leq r|x', w) - P(R_i \leq r|x, w) = -\mathbb{E}[\bar{f}_H(\tau_{V_i}(r)|\Delta_i, x, V_i, w) \cdot \Delta_i | W_i = w]$$

where $\bar{f}_H(y|\Delta, x, v, w) := \frac{1}{\Delta}\int_{y-\Delta}^{y} f_H(h|\Delta, x, v, w) \cdot dh$ *is the average density of* H *between* $y - \Delta$ *and* y*, among individuals with reporting function* v*, treatment effect* Δ*, and* $(X_i, W_i) = (x, w)$.[29]

Theorem 2 generalizes Eq. (3) from the introduction, which before introducing control variables w used the notation $\bar{f}_{x,x'}(\Delta, v)$ for $\bar{f}_H(y|\Delta, x, v, w)$. To simplify notation, I now leave the dependence of $\bar{f}_H(y|\Delta, x, v, w)$ on x' (through the definition of Δ) implicit.

Similar to Lemma 2, Theorem 2 shows that the change in $P(R_i \leq r|W_i = w, X_i = x)$ over discrete changes in x can be written as a positive linear combination of the causal effect of that variation in X on H. The weighting factor $\bar{f}_H(\tau_{V_i}(r)|\Delta_i, x, V_i, w)$ is positive for each i, and the sign of a parameter of the form $\tilde{\Delta}$ introduced in Section 3.3 is thus identified. However $\bar{f}_H(\tau_{V_i}(r)|\Delta_i, x, V_i, w)$ is unknown to the researcher, determined in part by individuals' reporting functions and the underlying distribution of H_i.

Intuitively, respondents with treatment effect value Δ are "counted" in the above average if there exists a positive mass of such individuals with $(X_i, W_i) = (x, w)$ and happiness H_i in the range $\tau_{V_i}(r) - \Delta$ to $\tau_{V_i}(r)$. Note that Theorem 2 exhausts all implications of the observable data (R_i, X_i) under MONO and EXOG regarding variation in the potential outcome functions $h(x, u)$ with respect to x (for a fixed value of the controls W_i).[30]

Figure 5 illustrates an example of Theorem 2. Suppose there are two response categories $\mathcal{R} = \{0, 1\}$ with a common reporting function $r(h) = \mathbb{1}(h \geq \tau)$. By iterating expectations over Δ_i, we can consider a single value Δ of Δ_i at a time. Thus we aim to show that $\mathbb{E}[R_i|x', \Delta] - \mathbb{E}[R_i|x, \Delta] = \bar{f}_H(\tau|x, \Delta) \cdot \Delta$, using that $P(R_i \leq 0|x', \Delta) = 1 - \mathbb{E}[R_i|x', \Delta]$. In Figure 5, I make the conditioning on $\Delta_i = \Delta$ implicit to simplify notation, taking an example in which X_i is an indicator for marriage with $x' = 1$, $x = 0$.

Lemma 2 as a limiting case of Theorem 2: A similar expression to that of Theorem 2 shows up in the "bunching design", which leverages bunching at kinks in decision-makers' choice sets for identification of behavioral elasticities. Since the kink compares just two distinct slopes, an identification problem emerges for elasticity parameters (Blomquist et al., 2021). An assumption sometimes used sidestep this issue is that the kink is "small" (e.g. Saez 2010; Kleven 2016, see Goff 2022 for a discussion). An analogous assumption in the context of Theorem 2 would be that Δ_i is always small so that for each $\Delta \in supp\{\Delta_i\}$, the density $f_H(h|\Delta, x, v, w)$ is approximately constant for all h between $\tau_v(r) - \Delta$ and

[29] By "between $y - \Delta$ and y" I mean in the interval $[\min\{y - \Delta, y\}, \max\{y - \Delta, y\}]$, regardless of the sign of Δ. Note that $\bar{f}_H(y|\Delta, x, v)$ is positive even if $\Delta < 0$, in which case it is equal to the average density between y and $y + |\Delta_i|$.

[30] Given any such fixed w, once $P(R_i \leq r|X_i = x, W_i = w)$ is known for all r for some fixed reference value x of the explanatory variables, along with the distribution of $X_i|W_i = w$, the only remaining information available from the data takes the form of differences $P(R_i \leq r|x', w) - P(R_i \leq r|x, w)$ for various values of x' and r.

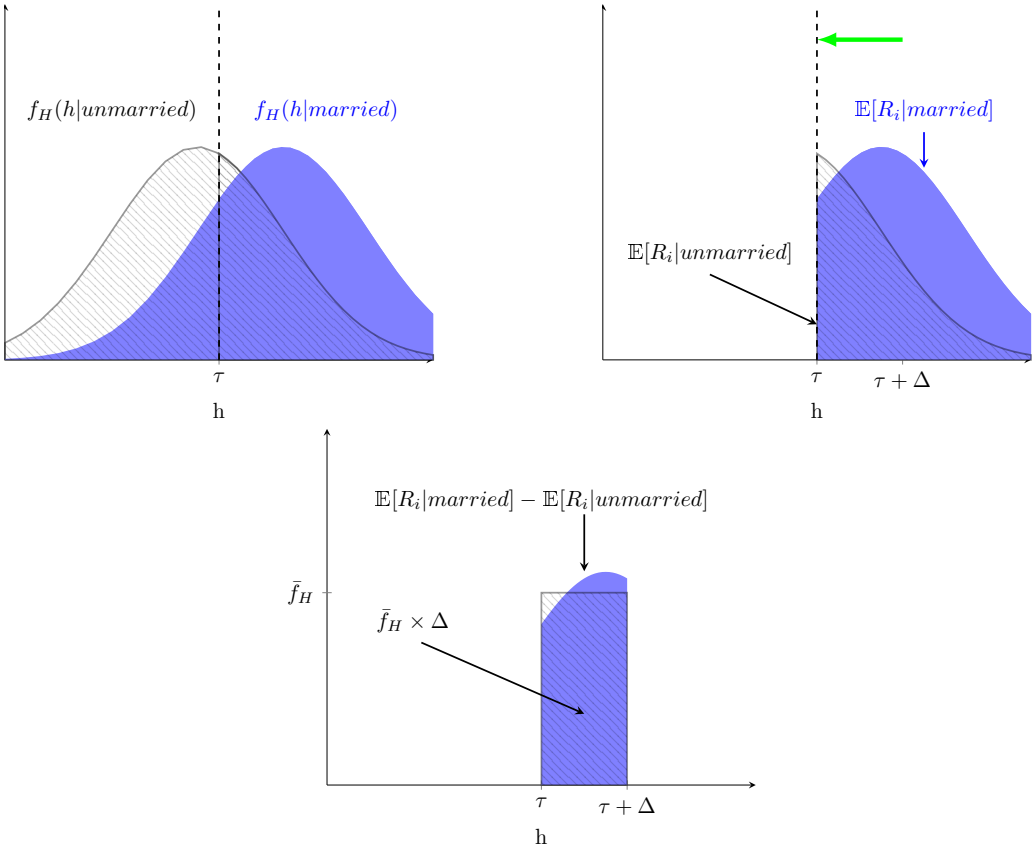


Figure 5: Visualization of Theorem 2. Conditional on $\Delta_i := h(married, U_i) - h(unmarried, U_i) = \Delta$, $f_H(\cdot|married) = f_{h(married,U)}(\cdot)$ is a rightward shift of $f_H(\cdot|unmarried) = f_{h(unmarried,U)}(\cdot)$, by Δ. Thus $\mathbb{E}[R_i|married, \Delta] - \mathbb{E}[R_i|unmarried, \Delta]$ is the area under $f_H(\cdot|married)$ between τ and $\tau + \Delta$, which is in turn equal to a rectangle of width Δ and height $\bar{f}_H$, where $\bar{f}_H$ is the average of $f_H(\cdot|married)$ across this interval.

$\tau_v(r)$. Under this assumption, Theorem 2 would simplify to:

$$P(R_i \leq r|x', w) - P(R_i \leq r|x, w) = -\int dF_{V|W}(v|w) \cdot f_H(\tau_{V_i}(r)|\Delta_i, x, V_i, w) \cdot \Delta_i$$
$$= -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v, w) \cdot \mathbb{E}[\Delta_i|H_i = \tau_v(r), X_i = x, V_i = v, W_i = w] \quad (19)$$

Eq. (19) exactly recovers the weighting over individuals achieved by Lemma 2 using continuous variation in x. In particular, the quantity $\mathbb{E}[\Delta_i|\tau_v(r), x, v, w]$ appears above with the same weight $-dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v, w)$ as $\mathbb{E}[\partial_{x_j} h(x, U_i)|\tau_v(r), x, v, w]$ does in Eq. (12). However, the constant density assumption used to obtain (19) is quite hard to justify *except* in the limit that Δ_i is very small with probability one.[31] Section 6.2 thus explores this issue further when Δ_i is not small, in the context of mean regression.

Mean regression: As our main focus is regressions capturing the conditional mean of R_i with $\mathcal{R}$ an integer response scale, one can again aggregate Theorem 2 across the response

[31] If we consider the limit $x' \to x$ with the two differing only in component j, this approximation becomes exact and Eq. (19) applied to $(P(R_i \leq r|x', w) - P(R_i \leq r|x, w))/(x'_j - x_j)$ reduces to Lemma 2. See Lemma SMALL in Goff (2022).

categories r to obtain:[32]

$$\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w] = \mathbb{E}\left[\sum_{r=0}^{\bar{R}-1} \bar{f}_H(\tau_v(r)|\Delta, x, v, w) \cdot \Delta_i \middle| W_i = w\right] \tag{20}$$

Recall from Lemma 2 that derivatives of the conditional distribution of R yield causal effects $\nabla_x h(x, U_i)$ with weights proportional to $\sum_r f_H(\tau_v(r)|x, v)$. By contrast, (20) shows that discrete differences in X recover treatment effects $\Delta_i = h(x', U_i) - h(x, U_i)$ with "weights" that themselves depend upon Δ_i through $\sum_r \bar{f}_H(\Delta_i, x, v, w)$. Since this quantity depends not only on the density of H at response thresholds $\tau_v(r)$ but also the density at points within Δ of such thresholds through $\bar{f}$, the two weighting schemes do not lead to estimands that can obviously be directly compared.

Note that the sign of $\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w]$ does *not* reflect the sign of the unweighted average treatment effect $\mathbb{E}[\Delta_i]$. Rather, the sign of the difference in means depends on how positive and negative treatment effects are aggregated over by the weights $\sum_r \bar{f}_H(\tau_v(r)|\Delta, x, v, w)$. If the CDF functions of $h(x, U_i)$ and $h(x', U_i)$ cross, then there must be some individuals with $\Delta_i < 0$ while others with $\Delta_i > 0$.[33]. This connects Theorem 2 to the result of Bond and Lang (2019), discussed further in Appendix A.

6.2 Comparing discrete and continuous regressors

Eq. (19) and Theorem 2 together imply that regression coefficients between discrete and continuous treatment variables can be meaningfully compared quantitatively in terms of causal effects in the limit that treatment effects for the discrete treatment are very small, given a linear conditional mean function of R on X.

More generally, a researcher who is interested in comparing a local regression derivative to the mean difference across two discrete groups can construct ratios of the form:

$$\frac{\mathbb{E}[R_i|X_i = x', W_i = w] - \mathbb{E}[R_i|X_i = x, W_i = w]}{\partial_{x_1}\mathbb{E}[R_i|X_i = x'', W_i = w]} \tag{21}$$

for some x,x', and x''. For example, if $X = (income, marriage)$ with $x' = (y, married)$ and $x = (y, unmarried)$ for any income y and $x'' = (y, m)$ for $m \in \{married, unmarried\}$, then (21) would yield a comparison of regression contrasts involving income to those involving marriage. If $\mathbb{E}[R_i|X_i, W_i]$ were fully linear, then the numerator of (21) would be the regression coefficient on marriage and the denominator would be the regression coefficient on income.

In Appendix E, I show that complications in the interpretation of (21) arise from two effects: discreteness of the response scale, and "non-linearity" in the spacing of the thresholds $\tau_v(r)$ for a given individual. There I define a formal notion of the response categories being "dense" in the space of latent H_i, for each reporting function type v.

[32] To obtain the notation of Eq. (3) in the introduction from (20), define $\bar{f}_H(\Delta, x, v, w) := \sum_{r=0}^{\bar{R}-1} \bar{f}_H(\tau_v(r)|\Delta, x, v, w)$.

[33] Specifically, then $P(\Delta_i < 0) \geq \sup_t \left\{F_{h(x',U_i)}(t) - F_{h(x,U_i)}(t)\right\}$ and $P(\Delta_i > 0) \geq \sup_t \left\{F_{h(x,U_i)}(t) - F_{h(x',U_i)}(t)\right\}$; see e.g. Fan and Park (2010)

This *dense response limit* allows us to conceptualize there as being an infinite number of response categories, while remaining contained between 0 and a fixed $\bar{R}$. The dense response limit delivers a tractable approximation for deriving analytical results, which may be reasonable to apply in instances in which the survey question offers many response categories between a lower and upper limit (e.g. integers from 0 to 100).

In Appendix E, Proposition 6 shows that bounds on the ratio of total weights in Equation (34) can be obtained in the dense response limit when each individual spaces out the thresholds $\tau_v(r)$ at roughly equal intervals—yielding reporting functions that are individually linear. Proposition 6 gives two sets of bounds, one derived under weaker assumptions than the other. First, a more general bound suggests that discrete contrasts will tend to *overstate* causal effects relative to regression derivatives, by a factor that is upper bounded by two. A second bound further assumes that the "sensitivity" of individual reporting functions is not too heterogeneous, and suggests that the inflation factor can also be upper bounded by the reciprocal of the fraction of the population that do not bunch at the endpoints of the response scale. This bound is close to unity when there are few such bunchers, which can be verified empirically.

In this setting, and if $\mathbb{E}[R_i|X_i, W_i]$ is close to linear, the two bounds can be intersected and we can conlude that discrete contrasts will overstate causal effects relative to regression derivatives, but not by much. This obtains a special case in which regression coefficients from a mix of discrete and continuous regressors coefficients from an OLS regression can be meaningfully compared quantitatively.

To assess the performance of the theoretical bounds described above, Appendix E.5 simulates several data-generating-processes (DGPs) for H_i and for the response functions $r(\cdot, V_i)$. The simulations generally provide an optimistic picture that quantities of the form $\{\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w]\} / \partial_{x_j} \mathbb{E}[R_i|x, w]$ can be interpreted as close to a ratio of weighted averages of causal effects, in these DGPs. In general, results do not seem to differ substantially whether the number of response categories is small, or whether there are few or many different reporting functions present in the population. When treatment effects become very *large* relative to the dispersion of happiness in the population, non-linearity in the density of the conditional distribution of happiness becomes important and the simulations make apparent that comparisons of magnitude can become misleading.

7 Conclusion

This paper investigates the identification of causal effects when using subjective responses as an outcome variable. Such reports typically ask individuals to choose a response from an ordered set of categories, and how individuals use those categories can be expected to differ by individual i. Nevertheless, researchers may be willing to suppose that individual responses reflect the value of a well-defined latent variable H_i.

Without observing H_i and without assuming it is possible to rank individuals by H_i

on the basis of their responses R_i, we have seen that the conditional distribution of R_i given exogenous covariates X_i can still be informative about the causal effects of X on H. While this allows one to observe the sign of causal effects under the assumption that this sign is common across individuals, and to compare magnitudes between two continuous treatments, continuous treatment variables typically only identify effects among individuals who are on the threshold between two response categories. Meanwhile with discrete treatments, mean comparisons can impose different total weightings over individuals in the population. The results suggest that researchers should be mindful of the variation being employed when comparing the magnitude of regression effects across explanatory variables, even when those variables are as good as randomly assigned.

In particular, I draw from the results three practical suggestions for the use of regression analysis for causal inference with subjective ordinal outcomes. First, the critique that such responses are only ordinarily and not quantitatively meaningful is most pronounced when researchers compare two populations that differ among many dimensions in an uncontrolled observational setting. However, when the researcher properly isolates exogenous variation in treatment variables by estimating $\mathbb{E}[R_i|X_i]$ flexibly under as-good-as-random assignment of the treatment variable(s) of interest, regression differences and derivatives do identify the sign of a convex average of causal effects. Second, to make such regression derivatives quantitatively meaningful, researchers should focus on *comparing* across treatment variables when more than one is available. Third, researchers should still exercise some caution when comparing the magnitudes of two discrete treatment effects or between a discrete treatment effect and the slope for a continuous treatment. The relative magnitudes of convex averages of causal effects can still be partially identified in such settings with further assumptions, and weakening these assumptions represents a possible avenue for future research.

Appendices

Appendix A provides an extended discussion of how my results relate to Bond and Lang (2019). Appendix B relates my general model of ordered response to ones previously considered in the literature. Appendix C considers several extensions to my baseline model, such as using instrumental variables rather than selection-on-observables for identification, or allowing for a multivariate latent variable. Appendices D and E develop some supporting theoretical results for the paper, and Appendix F provides additional material related to the empirical application.

A Reconciling my results with Bond and Lang (2019)

Bond and Lang (2019) (BL) do not focus on causal effects, but show that certain statements about even the statistical relationship between H and X cannot be answered by the data (R, X). This is natural since H itself is not directly observed. To connect my results to BL, I first develop results echoing theirs in the notation of this paper while also focusing on statistical, rather than causal relationships. I then impose Assumption EXOG to discuss how my results can coexist with BL's, in the language of causal effects.

To prove their main result about the non-identification of $sgn(\theta)$, where $\theta := \mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x]$, Bond and Lang (2019) draw on results from Manski and Tamer (2002) for regressions with interval-valued outcome data. An alternative way to see the problem directly is to write θ as:

$$\theta = \int_0^1 \left\{Q_{H|X=x'}(u) - Q_{H|X=x}(u)\right\} \cdot du \tag{22}$$

where $Q_{H|X}$ is the conditional quantile function of H_i given X_i.[34] Meanwhile, the mean difference in R_i instead identifies, in the case of a common reporting function:

$$\mathbb{E}[R|X = x'] - \mathbb{E}[R_i|X_i = x] = \int_0^1 \bar{r}'_{x',x}(u) \cdot \left\{Q_{H|X=x'}(u) - Q_{H|X=x}(u)\right\} \cdot du, \tag{23}$$

where $\bar{r}'_{x',x}(u) := \frac{r(Q_{H|X=x'}(u)) - r(Q_{H|X=x}(u))}{Q_{H|X=x'}(u) - Q_{H|X=x}(u)}$ is the "average rate of change" in the common reporting function $r(\cdot)$ between $Q_{H|X=x}(u)$ and $Q_{H|X=x'}(u)$. Eq. (23) can be derived by noting that $Q_{r(H)|X=x}(u) = r(Q_{H|X=x}(u))$ (Hosseini, 2010).

Eq. (23) thus represents a re-weighting of the quantile differences $Q_{H|X=x'}(u) - Q_{H|X=x}(u)$ that appear in (23) with uniform weight under the integral over all $u \in [0, 1]$ in Eq. (22). The quantity $\bar{r}'_{x',x}(u)$ is weakly positive for any (x, x', u) since r is weakly increasing, and will exhibit discrete jumps or falls at the u for which $Q_{H|X=x'}(u)$ and

[34] Eq. (22) follows from the identity $\mathbb{E}[A] = \int_0^1 Q_A(u) \cdot du$ for any random variable A.

$Q_{H|X=x}(u)$ lie on opposite sides of a response threshold $\tau(r)$.[35] Where exactly the weight $\bar{r}'_{x',x}(u)$ is smaller or larger depends on the distribution of latent happiness and the spacing of the response thresholds $\tau(r)$, which are both unknown.

Another special case in which the sign of $\mathbb{E}[R|X = x'] - \mathbb{E}[R_i|X_i = x]$ does identify the sign of θ is when the conditional distribution $H|X = x'$ stochastically dominates the conditional distribution $H|X = x$ (or vice versa), as mentioned in the introduction. In this case the sign of $Q_{H|X=x'}(u) - Q_{H|X=x}(u)$ is positive for all $u \in [0, 1]$, implying that θ and $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ will both be positive. If instead $Q_{H|X=x'}(u) < Q_{H|X=x}(u)$ for some u, while $Q_{H|X=x'}(u) > Q_{H|X=x}(u)$ for other u (i.e. the conditional quantile functions *cross*), then it will generally be possible to reverse the ordering of $\mathbb{E}[R_i|X_i = x']$ and $\mathbb{E}[R_i|X_i = x]$ for a given θ depending on where the unknown function $r(\cdot)$ increases the fastest (see Schröder and Yitzhaki (2017) for a version of this argument).

A.1 Convex averages of causal effects are identified, and interesting

Since the observable data are not dispositive on their own, one can of course always couple the data with stronger assumptions to identify the sign of θ. Suppose that $X_i = x'$ indicates the i is a resident of the United States and $X_i = x$ that i is a resident of Japan. If one is willing to assume that the country with higher mean R has a higher happiness at every quantile level $u \in [0, 1]$—whichever country that is—then the sign of θ is identified. But since life differs in many ways between the US and Japan which may matter in different ways for different individuals, it is hard to make this argument compellingly. Indeed, the underidentification problem for the sign of θ is most acute when comparing means of R_i between two distinct populations that differ from one another along multiple dimensions, and each of which is quite heterogeneous on its own.

The above problem appears in a much less pronounced way when X_i represents a vector of treatments that are as-good-as-randomly assigned, as in Theorems 1 and 2 of this paper. In particular, if the treatment effect $\Delta_i = h_i(x') - h_i(x)$ has the same sign for all units i, then $H_i|X_i = x'$ *necessarily* stochastically dominates $H_i|X_i = x'$. As an example, consider a linear potential outcomes model in which $h_i(x) = h_i(x, U_i) = x^T\beta + U_i$. The treatment effect Δ_i is then $\Delta := (x' - x)^T\beta$, the same for all i. Given randomization $U_i \perp\!\!\!\perp X_i$, $Q_{H|X=x}(u) = Q_{h(X)}(u) = x^T\beta + Q_U(u)$ and the quantile difference $Q_{H|X=x'}(u) - Q_{H|X=x}(u) = \Delta$, numerically the same for all $u \in [0, 1]$.[36] The quantile

[35] Interestingly, when individuals differ in their reporting functions, it is conceivable that $\mathbb{E}[R|X = x'] - \mathbb{E}[R_i|X_i = x]$ achieve a uniform weighting over the quantiles $u \in [0, 1]$ (cf. Section 4.4). If $V_i \perp\!\!\!\perp X_i$, then I show in Proposition 5 of Goff (2025) that (23) generalizes to $\mathbb{E}[R_i|x'] - \mathbb{E}[R_i|x] = \int_0^1 \bar{r}'_{x',x}(u) \cdot \{Q_{H|X}(u|x') - Q_{H|W}(u|x)\}\, du$, where $\bar{r}'_{x',x}(u) := \int dF_V(v) \cdot \frac{r(Q_{H|X}(u|x'),v) - r(Q_{H|X}(u|x),v)}{Q_{H|X}(u|x') - Q_{H|X}(u|x)}$. With a continuum of V_i, $\bar{r}'_{x',x}(u)$ can be smooth and possibly constant over u.

[36] In fact without loss of generality we can normalize U_i to be uniform on $[0, 1]$, and $h(x, u) = Q_{H|X=x}(u)$. To see this, suppose instead that U_i has CDF F_U, but given randomization we have that $U_i \perp\!\!\!\perp X_i$. Note that with probability one, $h_i(x) = Q_{h(x)|X_i}(T_i)$ where $T_i := F_{h(x)|X}(h_i(x)|X_i)$. This is a general property of conditional distributions, see e.g. Lemma 3 of Goff et al. (2024) for a proof. Observe that since $h_i(x) = x^T\beta + U_i$ with $U_i \perp\!\!\!\perp X_i$, $T_i = F_U(U_i)$. Define $\tilde{h}_i(x) := Q_{H(x)|X_i}(T_i)$. We can similarly work out $\tilde{h}_i(x)$ to be $\tilde{h}_i(x) = Q_{x^T\beta + U_i|X}(T_i|X_i) = x'\beta + Q_U(T_i)$ using $U_i \perp\!\!\!\perp X_i$. Putting this all together, we have that with probability one $h_i(x) = \tilde{h}_i(x) = x'\beta + \tilde{U}_i$, where we define $\tilde{U}_i := Q_U(F_U(U_i))$. Note that $Q_U(F_U(U_i)) \sim Unif[0, 1]$ and is independent of X_i (if U_i is not continuously distributed, $\tilde{U}_i$ can be suitably redefined so that it remains uniform, see Lemma 4 of Goff et al. (2024)).

functions never cross.

Assuming a linear causal model with homogeneous treatment effects would be very restrictive, but the above example illustrates a broader point. As Theorem 2 shows, differences in the distribution of R_i between two distinct points $X_i = x'$ and $X_i = x$ reveal, under random assignment, positive aggregations of treatment effects Δ_i, among units whose response value would change given a counterfactual shift from x to x'. In the limit that $x' \to x$, the local derivative of $P(R_i \leq r|x)$ at x yields the sign of the unweighted average marginal causal effect of changing x among individuals at the threshold between response categories r and $r+1$, as shown in Lemma 2. Whether this local average effect among marginal respondents is informative about the overall average effect of changing x to x' depends on how heterogeneous casual effects are in the population. With random assignment, the overall average treatment effect (ATE) between treatment values x and x':

$$\mathbb{E}[\Delta_i] = \mathbb{E}[h(x', U_i)] - \mathbb{E}[h(x, U_i)] = \mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x]$$

corresponds exactly to the parameter θ considered by BL.

This type of reasoning echoes the analysis of instrumental variables with heterogeneous treatment effects. In the LATE model of Imbens and Angrist (1994), a binary instrument reveals the average effect of a binary treatment among compliers. Whether this local average is informative about the overall ATE depends on how different treatment effects are between the compliers and other groups in the population. Unlike in the LATE context, the "marginal respondents" in our setting that are averaged over in the causal effects revealed by the data constitute a measure-zero subset of the population given that for each reporting function v they represent a single value of the continuous variable H. Furthermore, the magnitudes of regression derivatives or differences reflect not only magnitudes of causal effects, but the density of happiness values near the thresholds between response categories. This underscores the value of *comparing* the magnitudes across treatment variables, rather than interpreting the magnitudes individually.

A.2 Targeting ratios of effects rather than the effect of one treatment

Indeed, recall from Eq. (2) that the ratio of local regression derivatives identifies the ratio of convex combinations of the causal effects of the two continuous treatment variables. In a model $h(x, u) = \mathtt{h}(g(x), u)$ that is weakly separable between x and u, Section 4.5 showed that this ratio in turn identifies both the sign and magnitude of marginal rates of substitution between the treatments. For example, if $g(x) = x^T\beta$, we identify β_2/β_1.

The weakly separable class of functions is quite broad, and includes cases in which we may not even be able to identify the sign of parameters like β_1 or β_2 individually, due to the very problem highlighted by Bond and Lang (2019). Yet, we can identify both the sign and the exact magnitude of their *ratio*. This is a counter-intuitive result, so I illustrate it below with a simple example.

Consider the model $h(x,u) = (\beta_1 x_1 + \beta_2 x_2) \cdot (u - 1/3) + 1/3$ with $\beta_1, \beta_2 > 0$. This is a weakly separable model with $g(x) = x^T\beta$ and $\mathtt{h}(g,u) = g \cdot (u - 1/3) + 1/3$. Note that in this model the sign of the effect of a small increase in x depends on U_i, if $U_i > 1/3$ then $\partial_{x_1}(x, U_i) = \beta_1 \cdot (U_i - 1/3)$ is positive. If $U_i < 1/3$, then $\partial_{x_1}(x, U_i)$ is negative. The same considerations apply to X_2. If for example $U_i \sim Unif[0,1]$, then the average effect of a small increase in either treatment ends up being positive, since then $\mathbb{E}[U_i - 1/3] = 1/6$. However if instead $U_i \sim Unif[0, 1/2]$, then the average marginal effect is negative. The distribution of U_i is not known by the researcher, and the sign of $\mathbb{E}[\partial_{x_j} h(x, U_i)]$ is not identified from the data for either $j \in \{1, 2\}$ and any x.

However, the sign and the magnitude of $\mathbb{E}[\partial_{x_2} h(x, U_i)]/\mathbb{E}[\partial_{x_1} h(x, U_i)]$ *is* identified. The reason is that the sign of each variable's individual effect cancels out in the ratio: the sign of the ratio is identified, despite the sign of each variable's own mean effect being unknowable. If we let $\mathtt{h}'$ denote the partial derivative of $\mathtt{h}$ with respect to its first argument, then $\mathtt{h}'(g,u) = u - 1/3$ and:

$$\frac{\partial_2 \mathbb{E}[R_i | X_i = x]}{\partial_1 \mathbb{E}[R_i | X_i = x]} = \frac{\cancel{\mathbb{E}[\mathtt{h}'(x^T\beta, U_i)]} \cdot \beta_2}{\cancel{\mathbb{E}[\mathtt{h}'(x^T\beta, U_i)]} \cdot \beta_1} = \frac{\cancel{\mathbb{E}[U_i - 1/3]} \cdot \beta_2}{\cancel{\mathbb{E}[U_i - 1/3]]} \cdot \beta_1} = \frac{\beta_2}{\beta_1}$$

by Equation (17). Although the sign of $\mathtt{h}'(g,u)$ varies with u and $\mathbb{E}[\mathtt{h}'(x^T\beta, U_i)]$ is not identified by the data, it appears in both the numerator and the denominator and cancels.

We can see this phenomenon manifest with discrete differences in X as well. Suppose that $\beta_2 = 10$ and $\beta_1 = 1$, and $U_i \sim Unif[0,1]$. Consider $x = (1,0)'$ and $x' = (1+\epsilon, 0)'$ for $\epsilon > 0$, so that $x^T\beta = 1$ and $(x')^T\beta = 1 + \epsilon$. We then have $Q_{H|X=x'}(u) - Q_{H|X=x}(u) = \epsilon \cdot (u - 1/3)$, so the conditional quantile functions always cross at $u = 1/3$. Accordingly, the sign of $\theta = \mathbb{E}[H_i | X_i = x'] - \mathbb{E}[H_i | X_i = x]$ is not identified, as argued by Bond and Lang (2019). This holds for any ϵ, even as it becomes very close to zero. Accordingly, the sign of the overall average marginal effect $\mathbb{E}[\partial_{x_1} h(x, U_i)]$ with $x = (1,0)'$ remains unidentified as we take $\epsilon \downarrow 0$. However, as we saw above, the ratio $\mathbb{E}[\partial_{x_2} h(x, U_i)]/\mathbb{E}[\partial_{x_1} h(x, U_i)]$ *is* identified, in this weakly separable model of potential outcomes.

B Relationship to existing econometric models

The framework of this paper outlined in Section 3 is primarily related to two strands of econometric literature: i) models of ordered response; and ii) nonseparable outcome models with possible endogeneity and instrumental variables. This section describes the relationship to both of these literatures.

B.1 Ordered response models

The model outlined in Section 3 nests familiar econometric models of ordered response, that typically make parametric assumptions about the functions h, r and the distribution of unobservables, while entirely eliminating heterogeneity in v.

For example, the probit model treats the case in which $\mathcal{R} = \{0, 1\}$, and lets $R_i = \mathbb{1}(X_i'\beta + U_i \geq 0)$ where $U_i|X_i \sim N(0, \sigma^2)$ where σ is often normalized to 1. This fits into the general model above with V_i taken to be degenerate (all units share a value v), $\tau(0) = 0$, U_i a scalar and $h(x, u) = x^T\beta + u$ for some $\beta \in \mathbb{R}^{d_x}$. The assumption that U is independent of X_i then implies EXOG. In the probit model, the effect on H of a switch from $X_i = x$ to $X_i = x'$ is common across units, given by $(x' - x)^T\beta$. The ordered probit model maintains this same structure but with a larger set of categories $\mathcal{R} = \{0, 1, \ldots \bar{R}\}$, with corresponding thresholds $\tau(0), \tau(1), \ldots, \tau(\bar{R} - 1)$ common across individuals.

Despite the popularity of (ordered) probit and logit models, it is not necessary to impose a parametric structure on $h(x, u)$ or the distribution of U to obtain identification in binary and ordered choice settings. Matzkin (1992) shows that h can be identified up to scale under fairly general conditions if u is a scalar and h admits a separable structure: $h(x, u) = g(x) + u$ for some function g. This model allows for individual-specific reporting functions in a trivial sense, since owing to the additive separability the distinction between thresholds $\tau_v(r)$ and the error u is simply a matter of definition.[37] However, a separable model like $h(x, u) = g(x) + u$ for potential outcomes, like the probit model, imposes the strong restriction that treatment effects are the same for all individuals. My results allow for treatment effect heterogeneity, and nests a leading case of Matzkin (1992) when the treatment variables are all continuous (see Appendix D.1).

B.2 Nonseparable outcome models with or without endogeneity

Suppose for the moment that H were observed. Then Equation (7) along with Assumption EXOG would yield a nonseparable model for the outcome H with a set of exogenous regressors X, with no restrictions on the dimension of heterogeneity U or functional restrictions like monotonicity in X or U. In this general setting, Hoderlein and Mammen (2007) and Sasaki (2015) show that with continuous X quantile regressions reveals outcome-conditioned average treatment effect parameters (this terminology is due to Hoderlein and Sasaki 2013). Kasy (2022) provides similar results for a multi-dimensional set of outcome variables, and Chernozhukov et al. (2015) extend to panel data settings. Blundell et al. (2017) use invertibility assumptions to afford identification of an entire structural function with multi-dimensional outcomes.

However in my setting only R is observed, and not H. This leads to the model of Section 3 in which R, H and X are related through Equations (6) and (7). This structure resembles triangular instrumental variables (IV) models, where my X plays the role of the instrument(s) and Eq. (7) represents the "first stage" relationship between the instrument(s) and endogeneous regressor. Reporting functions play the role of the outcome equation in an IV setup, and "endogeneity" arises if $U_i \not\perp\!\!\!\perp V_i$, explicitly allowed

[37] Indeed, fixing any r and defining $Y_i^r = \mathbb{1}(R_i \leq r)$ we may write $Y_i^r = \mathbb{1}(g(X_i) + \eta_i^r \leq 0)$ where $\eta_i^r = U_i - \tau_{V_i}(r)$. Under conditions given by Matzkin (1992), the function g and the distribution of η^r can be identified (up to a scale normalization). See also Cunha et al. (2007). Since this can be done for each value r, the function g is in fact overidentified with more than two categories. Matzkin (1994) establishes conditions for identification of g in a weakly separable model $Y_i = r(\mathtt{h}(g(X_i), \eta_i))$, but requires η_i to be scalar.

in my model. However unlike IV settings, one cannot observe the "endogenous variable" H_i, which renders the analysis of identification very different in my setting.[38] In the literature thus far that has assumed H is observed, it has been found that monotonicity assumptions can be helpful in securing identification when structural functions are taken to be nonseparable as they are in my model (Imbens and Newey, 2009; D'Haultfœuille and Février, 2015; Torgovitsky, 2015; Hoderlein et al., 2016).

The result of Hoderlein and Mammen (2007) for nonseparable models with exogeneity has previously been used to study identification from discrete choice probabilities in Chernozhukov et al. (2019). Matzkin (2019) also analyzes some nonseparable models of discrete choice. To my knowledge the present paper is the first to leverage results on the link between quantile regressions and conditional average causal effects to address the concerns highlighted by Bond and Lang (2019) regarding the use of ordinal scales.

Finally, I note that this paper is related to the literature on measurement error and misclassification, in that one might view R as a imperfect measure of H contaminated by the reporting function. However, I let latent happiness H and responses R exist on entirely different scales (e.g. H in $\mathbb{R}$ and R in a set of integers), in the tradition of ordered response models and in common with Bond and Lang (2019). This feature also distinguishes the approach of this paper from models of rounding (Hoderlein et al., 2015), measurement error (Schennach and Hu, 2013), and discrete misclassification (Hu, 2008; Oparina and Srisuma, 2022).

C Extensions of the basic model

C.1 Using instrumental variables for identification

Suppose for that rather than making Assumption EXOG, we instead have a set of observed variables Z_i to use as instruments for X_i. We assume each X_j for $j = 1 \dots J$ is continuously distributed, and Z_i contains a continuously distributed instrument corresponding to each X_j, i.e.

$$X_{1i} = x_1(Z_i, W_i, \eta_{1i}), \quad X_{2i} = x_2(Z_i, W_i, \eta_{2i}) \quad \dots \quad X_{Ji} = x_J(Z_i, W_i, \eta_{Ji})$$

Finally, for each $j = 1 \dots J$, suppose that $x_j(z, w, \eta_j)$ is strictly increasing in η_j. Let $\eta_i = (\eta_{1i}, \eta_{2i}, \dots \eta_{Ji})^T$. We now assume that Z_i, rather than X_i, is (conditionally) independent of all other heterogeneity across individuals i:

Assumption INSTRUMENT (conditional independence of instruments).

$$\{Z_i \perp\!\!\!\perp (\eta_i, U_i, V_i)\} \,|W_i$$

[38] Indeed, the IV analogy yields some intuition for my results: although variation in X induces exogeneous variation in H and in R through H, we cannot re-scale the "reduced form" relationship between X and R by the "first stage" relationship between H and X, since H is unobserved.

The following result adapted from Imbens and Newey (2009) implies that under INSTRUMENT we can use η_i as a control variable in W_i, in the sense that

Lemma. *Under INSTRUMENT and the IV model above:* $\{X_i \perp\!\!\!\perp (U_i, V_i)\}|(\eta_i, W_i)$

Proof. Note that INSTRUMENT implies that $\{Z_i \perp\!\!\!\perp (U_i, V_i)\}|W_i, \eta_i$. Furthermore, conditional on η_{ji} and W_i, the only remaining variation in X_{ji} comes from Z_i. This is true for each j, so conditional on η_i and W_i, the only variation in X_i comes from variation in Z_i, i.e. X_i is simply a function of Z_i. The result then follows. □

Thus, if η_i is simply included in the vector of controls along, INSTRUMENT implies EXOG. "Controlling" for η_i is feasible, because given that each $x_j(z, w, \eta_j)$ is strictly increasing in η_j, we can without loss redefine $\eta_{ji} = F_{X_j|Z,W}(X_{ji}|Z_i, W_i)$ which can be estimated from the data for each j and individual i.[39] If no controls are needed for INSTRUMENT, then EXOG holds with $W_i := F_{X_j|Z}(X_{ji}|Z_i)$.

C.2 Subjectively-defined latent variables

In the main body of the paper, I assume that individuals use a reporting function $r_i(h)$ that is an increasing function of the variable h that the researcher is interested in. Given this, the model can accommodate arbitrary heterogeneity in $r_i(\cdot)$ (or equivalently: the locations of the thresholds that i uses), so long as this variation is independent of explanatory variables.

However in many applications, one might worry that not only are the definitions of the categories $\mathcal{R}$ subjective, but so is the definition of the quantity that individuals are asked to use in answering the survey question. For example, when answering a life-satisfaction question some individuals might think about their recent life experiences, while others may think about their whole life in aggregate. Some might spend a lot of time thinking about the question, while others might answer quickly and intuitively. Accordingly, let individual i use variable $\tilde{H}^i$ when they answer the survey question, where $\tilde{H}_i := \tilde{H}_i^i$ is i's value of this quantity that they define for themself. The key assumption that will allow us to extend the model to account for this kind of heterogeneity is that each $\tilde{H}_i$ is a weakly increasing function of some H, where H is an objectively-defined variable of ultimate interest to the researcher.

[39] Note that since $x_j(z, w, \eta_j)$ is strictly increasing in η_j, $F_{X_j|Z,W}(x_j|Z_i = z, W_i = w) = P(\eta_{ji} \leq x_j^{-1}(z, w, x_j)|W_i = w)$ where we have used INSTRUMENT. Define $\tilde{\eta}_{ji} := F_{X_j|Z,W}(X_{ji}|Z_i, W_i) = P(\eta_{ji} \leq x_j^{-1}(Z_i, W_i, X_{ji})|W_i) = F_{\eta_j|W}(\eta_{ji}|W_i)$. Observe from this that we can write $\tilde{\eta}_{ji}$ as a function of η_{ji}, conditional on W_i. Define $\tilde{\eta}_i = (\tilde{\eta}_{1i}, \tilde{\eta}_{2i}, \ldots \tilde{\eta}_{Ji})^T$ which is similarly a deterministic function of η_{ji} conditional on W_i. Since conditioning on $\tilde{\eta}_i$ and W_i is the same as conditioning on η_i and W_i, the random vector $\tilde{\eta}_i$ satisfies $\{Z_i \perp\!\!\!\perp (U_i, V_i)\}|W_i, \tilde{\eta}_i$. Note finally that $\tilde{\eta}_{ji} \sim Unif[0, 1]$ and with probability one $X_{ji} = \tilde{x}_j(Z_i, W_i, \tilde{\eta}_{ji})$ where $\tilde{x}_j(z, w, u) := Q_{X_j|W=w,Z=z}(u))$ for each $u \in [0, 1]$. Thus the Lemma holds after redefinition of η_i to be $\tilde{\eta}_i$ and each function x_j to be $\tilde{x}_j$.

I extend the model as follows: observables (R_i, X_i) are now related by

$$R_i = \tilde{r}_i(\tilde{H}_i) = \tilde{r}(\tilde{H}_i, S_i) \tag{24}$$

$$\tilde{H}_i = \tilde{h}_i(H_i) = \tilde{h}(H_i, T_i) \tag{25}$$

$$H_i = h_i(X_i) = h(X_i, U_i) \tag{26}$$

where *both* $\tilde{r}(\cdot, s)$ and $\tilde{h}(\cdot, t)$ are assumed to be weakly increasing and left-continuous. The new function, $\tilde{h}_i(h)$, can be defined in terms of counterfactuals: what would i's value of their subjectively-defined latent variable $\tilde{H}^i$ be if their objectively-defined happiness H_i were h? T_i can be of arbitrary dimension, allowing individual-specific mappings between H and $\tilde{H}$.

Now suppose that $\{X_{ji} \perp\!\!\!\perp (T_i, U_i, V_i)\} | \ W_i$. If we define $V_i = (S_i, T_i)$, then EXOG holds, and defining $r(\cdot, v) = \tilde{r}(\tilde{h}(\cdot, t), s)$ MONO now holds as well, allowing us to apply the main results of the paper. Note that EXOG is now stronger than it was in the baseline model: if we want to accommodate heterogeneity in what latent variable $\tilde{H}$ individuals use to answer the question, we must assume that heterogeneity to also be conditionally independent of X_j. In addition to the existing exclusion restriction that variation in X_j does not alter reporting functions $\tilde{r}_i$, we now have an additional implicit exclusion restriction that variation in X_j does not affect the subjective definitions T_i that individuals apply to generate $\tilde{H}_i$ in terms of H_i.

One nice feature of this extended version of the model is that the researcher may be more willing to make structural assumptions about the function $h(x, u)$ now that it is made explicit that H may differ from what individual's actually have in their mind when they answer the question. For example, if causal effects on some notion of objective life satisfaction H are assumed to be homogeneous (so that $h(x, u) = g(x) + u$), then marginal rates of substitution can be identified through Eq. (17), despite individuals using $\tilde{H}$ rather than H to answer the survey question.

As another example, suppose that ordinal preferences are quasilinear in X_1 and define $H_i = X_{1i} + \phi(X_{2i}, \ldots, X_{Ji})$ from the utility representation that counts a unit of X_1 the same for each individual i (see discussion in footnote 23). Then the extended model of this section allows different individuals to apply different cardinalizations of their preferences—captured by $\tilde{h}_i(\cdot)$—when they report R_i.

C.3 Multivariate latent variables

In some settings, it may be appealing to assume that subjective responses are driven by a vector of latent variables rather than a single one.

For example, Barreira et al. (2021) studies the mental health of economics graduate students in U.S. PhD programs, and include a question in which respondents are asked to agree or disagree with the statement "I have very good friends at my Economics Department". In such a case, respondents might consider both the quantity and quality

of friendships in their definition of "having good friends". The emphasis that respondents place on each may also vary by individual.

To model this case, we might replace Eq. (6) with

$$R_i = r(H_{1i}, H_{2i}, V_i)$$

where r is weakly increasing in both H_1 (number of friends) and H_2 ("average" quality of friendships). We further assume two separate structural functions $h_1(X, U)$ and $h_2(X, U)$ describing the effects of the X on quantity and quality of friendships, respectively.

For simplicity, let us first consider a case with a single reporting function $r(H_1, H_2)$, and a scalar x, and where no controls W_i are needed for EXOG. It will be useful to write

$$\frac{d}{dx_j} P(R_i \leq r | X_i = x) = \int \int_{T(r)} \frac{d}{dx} f_H(h_1, h_2|x) \cdot dh_1 dh_2 \tag{27}$$

where $T(r)$ is the set of (h_1, h_2) such that $r(h_1, h_2) \leq r$. In the above I have assumed dominated convergence so that one can interchange the integrals and derivative.

In the two-dimensional case, Eq. 4.1 of Hoderlein and Mammen (2008) shows that a quantity like $\frac{d}{dx} f_H(h_1, h_2|x)$ can be rewritten as:

$$\frac{d}{dx} f_H(h_1, h_2|x) = -\nabla \circ \begin{pmatrix} f_H(h_1, h_2|x) \cdot \mathbb{E}[\partial_x h_1(x, U) | H_{1i} = h_1, H_{2i} = h_2, X_i = x] \\ f_H(h_1, h_2|x) \cdot \mathbb{E}[\partial_x h_2(x, U) | H_{1i} = h_1, H_{2i} = h_2, X_i = x] \end{pmatrix}$$

where for a vector-valued function $\mathbb{h}(x)$, we let $\nabla \circ \mathbb{h}$ denote the divergence of $\mathbb{h}$. More generally, Kasy (2022) shows that for a vector $\mathbb{h} = (h_1, h_2, \ldots h_K)'$ of any finite dimension:

$$\frac{d}{dx} f_H(\mathbf{h}|x) = -\nabla \circ \{ f_H(\mathbf{h}|x) \cdot \mathbb{E}[\partial_x \mathbf{h}(x, U) | \mathbf{h}, x] \}$$

where we let $\mathbf{h}(x, U)$ be a vector of $(\mathbf{h}_1(x, U), \mathbf{h}_2(x, U) \ldots \mathbf{h}_K(x, U))'$.

In the general case with any $K \geq 1$ and again allowing reporting-function heterogeneity (satisfying EXOG), and multiple treatment variables, Eq. (27) becomes

$$\frac{d}{dx_j} P(R_i \leq r | X_i = x) = \int dF_{V|W}(v|w) \int_{T_v(r)} \frac{d}{dx_j} f_H(\mathbf{h}|x) \cdot d\mathbf{h} \tag{28}$$

where $T_v(r) := \{\mathbf{h} : r(\mathbf{h}, v) \leq r\}$.

An application of the divergence theorem allows us to rewrite Eq. (27) as an integral over *the boundary* $\partial T_v(r)$ of the set $T_v(r)$:

$$\frac{d}{dx_j} P(R_i \leq r | X_i = x) = \int dF_{V|W}(v|w) \int_{\partial T_v(r)} f_H(\mathbf{h}|x, v) \cdot \mathbb{E}[\partial_{x_j} \mathbf{h}(x, U) | \mathbf{h}, x, v] \circ \mathbf{n}_v(\ell) \cdot d\ell$$

where $\mathbf{n}_{x,v}(\ell)$ represents a normal vector perpendicular to $\partial_{T_v(r)}$ at a point indexed by ℓ. Figure 6 depicts this in the two-dimensional example. In that case, ℓ is a scalar index that parameterizes the path along the one-dimensional boundary of $T_v(r)$.

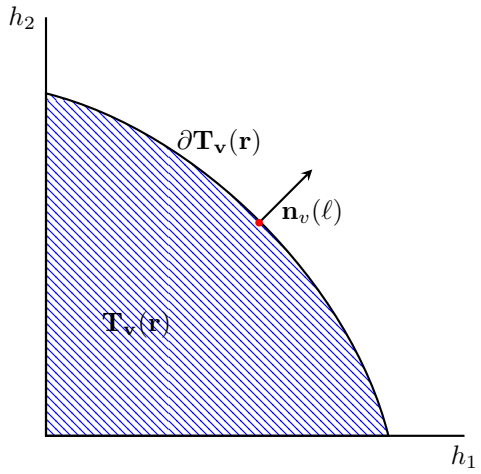


Figure 6: Components of $\hat{n}(\ell)$ are positive, by monotonicity of $h(h_1, h_2, v)$ w.r.t h_1 and h_2.

Provided that $r(\mathbf{h}, v)$ is weakly increasing in each component of $\mathbf{h}$ (for all reporting functions v), the components $n_{v,j}(\ell)$ of $\mathbf{n}_v(\ell)$ will be positive, as illustrated in Figure 6.

In the two-dimensional case for example, we have:

$$-\frac{d}{dx_j}P(R_i \leq r|X_i = x) = \int dF_{V|W}(v|w) \int_{\partial T_v(r)} f_H(\mathbf{h}|x,v) \cdot \left\{\hat{n}_{v,1}(\ell) \cdot \mathbb{E}[\partial_{x_j} h_1(x,U)|\mathbf{h},x,v] + \hat{n}_{v,2}(\ell) \cdot \mathbb{E}[\partial_{x_j} h_2(x,U)|\mathbf{h},x,v]\right\} \cdot d\ell$$

Suppose for example that $h_j(x,u) = x'\beta_k + u$ where β_{jk} represents the effect of treatment variable X_j on H_k. Then this becomes

$$\frac{d}{dx_j}P(R_i \leq r|X_i = x) = -\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{j1} \cdot \hat{n}_{v,1}(\ell) + \beta_{j2} \cdot \hat{n}_{v,2}(\ell)\} \cdot d\ell \middle| X_i = x\right]$$

where the expectation is over response functions V_i. Unless the boundary $\partial T_v(r)$ is linear in $\mathbf{h}$, the positive weights $\hat{n}_{v,2}(\ell)$ will generally vary with ℓ across the inner integral. However, the effects of two treatment variables can still be meaningfully compared. For example, suppose we have two continuous treatment variables of interest: X_1 and X_2, and that for each latent variable H_k, $\beta_{2k} = \gamma\beta_{1k}$. Then:

$$\frac{\partial_{x_2}P(R_i \leq r|X_i = x)}{\partial_{x_1}P(R_i \leq r|X_i = x)} = \frac{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{11} \cdot \hat{n}_{v,1}(\ell) + \beta_{12} \cdot \hat{n}_{v,2}(\ell)\} \cdot d\ell \middle| X_i = x\right]}{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{21} \cdot \hat{n}_{v,1}(\ell) + \beta_{22} \cdot \hat{n}_{v,2}(\ell)\} \cdot d\ell \middle| X_i = x\right]} = \gamma$$

D Additional identification results for continuous treatments

D.1 Additional results in the weakly separable case

This appendix continues the analysis of a weakly separable structural function $h(x,u) = \mathtt{h}(g(x),u)$ from Section 4.5 in the main text.

In the still simpler case of a partially linear h function, (17) leads to the following:

Corollary 3. *Suppose MONO and EXOG and REG_j for $j = \{1, 2\}$ hold, and that $h(x, u)$ takes the form $h(x, u) = x_1\beta_1 + x_2\beta_2 + g(x_3, \ldots x_J) + u$ (e.g. $h(x, u) = x^T\beta + u$) with $\beta_2 \neq 0$. Then, if EXOG holds with no control variables $\mathbb{E}[R_i|x]$ is also weakly separable, i.e. $\mathbb{E}[R_i|x] = \phi(\gamma_1 x_1 + \gamma_2 x_2, x_3 \ldots x_J)$ for some function ϕ, and $\gamma_2/\gamma_1 = \beta_2/\beta_1$. With controls, we instead have that $\mathbb{E}[R_i|x, w]$ is weakly separable in x for a fixed w, that is γ_1, γ_2 and function ϕ may all depend on w.*

Proof. Fix a w, and let $m(x) := \mathbb{E}[R_i|X_i = x, W_i = w]$. By (17), we have that $\partial_{x_2} m(x)/\partial_{x_1} m(x) = \beta_2/\beta_1$, for all x. This implies that m takes the form of ϕ above. □

As a final note, we can see how Lemma 2 recovers an identification result of Matzkin (1994) in the case of no controls W_i and REG_j holding for all components X_j of x. Note first that given the weakness of the assumptions made, we could only ever hope to identify $g(x)$ up to an increasing transformation. One functional restriction that removes this arbitrariness, considered by Matzkin (1994), is to suppose $g(x)$ is homogeneous of degree one. Matzkin (1994) also imposes that u be a scalar. In this case, Eq. (17) implies that g is identified up to scale:

Proposition 1. *Suppose MONO and EXOG hold, there are no controls W, and each of the $X_1 \ldots X_J$ are continuously distributed satisfying REG. Suppose further that $h(x, u) = \mathtt{h}(g(x), u)$, where g is homogeneous of degree one, continuously differentiable, and for some k: $\partial_{x_k} g(x) \neq 0$ for all $x \in \mathcal{X}$ with $\mathcal{X}$ a convex set in $\mathbb{R}^J$. Then $g(x)$ is identified up to an overall scale.*

The proof of Proposition 1 in Appendix G gives an explicit expression for $g(x)$. Note that Proposition 1 does not require U_i to be a scalar, in this regard generalizing a result of Matzkin (1994).

D.2 Details: marginal rates of substitution

By Lemma 2, a version of Eq. (13) also holds at a single value of X and W:

$$\frac{\partial_{x_2}\mathbb{E}[R_i|X_i = x, W_i = w]}{\partial_{x_1}\mathbb{E}[R_i|X_i = x, W_i = w]} = \frac{\tilde{\beta}_2(x, w)}{\tilde{\beta}_1(x, w)} \tag{29}$$

where $\tilde{\beta}_j(x, w) := \mathbb{E}\left[\frac{\rho(x,v,w)}{\mathbb{E}\left[\rho(x,V_i,w)|H_i \in \tau_{V_i}, X_i = x, w\right]} \cdot \partial_{x_j} h(x, U_i) \middle| H_i \in \tau_{V_i}, X_i = x, W_i = w\right]$ with $\rho(x, v, w) := \sum_r f_H(\tau_v(r)|x, v, w)$, and we assume that $\lim_{h\to\infty} f_H(h|x, v, w) = 0$ and that for each $v \in \mathcal{V}$, the $\tau_v(r)$ are all distinct.[40]

[40] Let A and B be random variables, where B is absolutely continuous and let $\mathcal{B}$ be a finite set of distinct values. Assume that $\mathbb{E}[A|B = b]$ is continuous in b, so we can then define $\mathbb{E}[A|B \in \mathcal{B}]$ simply as $\lim_{\epsilon\downarrow 0} \mathbb{E}[A| \min_{b\in\mathcal{B}} |B - b| < \epsilon]$ which works out to $\sum_{b\in\mathcal{B}} \frac{f_B(b)}{\sum_{b'\in\mathcal{B}} f_B(b')} \cdot \mathbb{E}[A|B = b]$.

In the weakly separable case (16), this simplifies:

$$\frac{\partial_{x_2}\mathbb{E}[R_i|x,w]}{\partial_{x_1}\mathbb{E}[R_i|x,w]} = \frac{\int dF_{V|W}(v|w)\cdot\rho(x,v,w)\cdot\partial_{x_2}g(x)\cdot\mathbb{E}\left[\partial_g \mathtt{h}(g(x),U_i)|H_i\in\tau_v,x,v,w\right]}{\int dF_{V|W}(v|w)\cdot\rho(x,v,w)\cdot\partial_{x_1}g(x)\cdot\mathbb{E}\left[\partial_g \mathtt{h}(g(x),U_i)|H_i\in\tau_v,x,v,w\right]}$$

which is equal to $\frac{\partial_{x_2}g(x)}{\partial_{x_1}g(x)}$ since the $\partial_{x_j}g(x)$ factor out and the terms in purple cancel.

A convenient feature of a weakly separable model is that since individual heterogeneity U affects the X variables after they are aggregated by g, ratios like $\partial_{x_1}g(x)/\partial_{x_2}g(x)$ captures the marginal rate of substitution between x_1 and x_2 for each unit. By contrast, (29) is not necessarily equal to a weighted average over marginal rates of substitution in the population, when they are heterogeneous between units. The following proposition gives a special case in which it does, without the strong condition of weak separability.

Proposition 2. *If in addition to the assumptions of Lemma 2 for $j=1,2$, we have*

- $Cov\left(\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)},\partial_{x_1}h(x,U_i)\middle|\, H_i\in\tau_{V_i},x,w\right)=0$
- $\{V_i \perp\!\!\!\perp U_i\}\mid(H_i\in\tau_{V_i},X_i,W_i)$

then

$$\mathbb{E}\left[\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)}\middle|\, H_i\in\tau_{V_i},X_i=x,W_i=w\right]=\frac{\partial_{x_2}\mathbb{E}[R_i|X_i=x,W_i=w]}{\partial_{x_1}\mathbb{E}[R_i|X_i=x,W_i=w]}$$

If $Cov\left(\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)},\partial_{x_1}h(x,U_i)\middle|\, H_i\in\tau_{V_i},x,w\right)\le 0$, *then* $\mathbb{E}\left[\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)}\middle|\, H_i\in\tau_{V_i},x,w\right]\ge\frac{\partial_{x_2}\mathbb{E}[R_i|X_i=x,W_i=w]}{\partial_{x_1}\mathbb{E}[R_i|X_i=x,W_i=w]}$ *and vice-versa if the inequality is reversed.*

Proof. From Lemma 2 and $R_i=\sum_{r=0}^{\bar{R}-1}\mathbb{1}(r<R_i)$, we have that

$$\partial_{x_j}\mathbb{E}[R_i|x,w]=\mathbb{E}\left[\rho(x,V_i,w)\cdot\partial_{x_j}h(x,U_i)|H_i\in\tau_{V_i},X_i=x,W_i=w\right] \tag{30}$$

This expression shows that $\partial_{x_j}\mathbb{E}[R_i|x,w]$ averages over all units having $X_i=x$ (and $W_i=w$), located at *any* of their individual-specific happiness thresholds, with (positive but not convex) weights $\rho(X_i,V_i,X_i)$.

Given $\{V_i \perp\!\!\!\perp U_i\}\mid(H_i\in\tau_{V_i},X_i,W_i)$ and (30), we have for $j\in\{1,2\}$

$$\partial_{x_j}\mathbb{E}[R_i|x,w]=\mathbb{E}\left[\rho(x,V_i,w)|H_i\in\tau_{V_i},X_i=x,W_i=w\right]\cdot\mathbb{E}\left[\partial_{x_j}h(x,U_i)|H_i\in\tau_{V_i},X_i=x,W_i=w\right]$$

So the RHS of (29) becomes: $\mathbb{E}\left[\partial_{x_2}h(x,U_i)|\,H_i\in\tau_{V_i},x,w\right]/\mathbb{E}\left[\partial_{x_1}h(x,U_i)|\,H_i\in\tau_{V_i},x,w\right]$. Now, using $Cov\left(\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)},\partial_{x_1}h(x,U_i)\middle|\, H_i\in\tau_{V_i},x,w\right)\le 0$,

$$\mathbb{E}\left[\partial_{x_1}h(x,U_i)|\,H_i\in\tau_{V_i},x,w\right]\le\mathbb{E}\left[\frac{\partial_{x_2}h(x,U_i)}{\partial_{x_1}h(x,U_i)}\middle|\, H_i\in\tau_{V_i},x,w\right]\cdot\mathbb{E}\left[\partial_{x_1}h(x,U_i)|\,H_i\in\tau_{V_i},x,w\right]$$

and analogously if $\le$ is replaced with $\ge$. □

Proposition 2 requires reporting heterogeneity V_i to be conditionally orthogonal to structural function heterogeneity U_i. Further, one must be able to at least sign the correlation

of marginal rates of substitution and heterogeneity in marginal effects with respect to x_2. This correlation might be negative, if for example, individuals with high returns to x_2 do not have returns to x_1 that are proportionally as high, on average.

D.3 Characterizing the marginal respondents

The following result gives conditions under which average characteristics of respondents on the margin between response category r and $r+1$, which drive the average causal effect identified by Lemma 2, can be identified from the data:

Proposition 3. *Let A_i be an individual characteristic such that EXOG holds conditionally on A_i, i.e. $\{X_i \perp\!\!\!\perp U_i\}|(A_i, W_i, V_i)$ and $\{X_{ji} \perp\!\!\!\perp V_i\}|(A_i, W_i)$. Suppose further that for treatment j the sign of $\partial_{x_j} h(x, U_i)$ is the same for all individuals i. Then (under REG_j and further regularity conditions described in the proof):*

$$\mathbb{E}[A_i|H_i = \tau_{V_i}(r), X_i = x, W_i = w] = \frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r|A_i, x, w)|x, w]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r|A_i, x, w)|x, w]} \tag{31}$$

Under the stronger independence condition that $\{X_i \perp\!\!\!\perp (A_i, U_i, V_i)\}|W_i$, this becomes

$$\mathbb{E}[A_i|h(x, U_i) = \tau_{V_i}(r), X_i = x, W_i = w] = \frac{\partial_{x_j} \mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r)|x, w]}{\partial_{x_j} P(R_i \leq r|x, w)} \tag{32}$$

The stronger assumption $\{X \perp\!\!\!\perp (A, U, V)\}|W$ in Proposition 3 leading to Eq. (32) is a natural one if the treatment(s) X are as-good-as-randomly assigned (conditional on W), and A represents a characteristic of individuals unaffected by the treatments X. In this case A will be independent of the treatments in the same sense that U and V are. As an example, one could in a study in which gender is observed estimate the proportion of respondents at each response margin r that are women. To do this, one only needs to supplement the regression contemplated by Lemma 2 with another than multiplies $\mathbb{1}(R_i \leq r)$ by characteristic A_i, and compute the ratio of regression derivatives.

The weaker condition leading to Eq. (31) would hold if A is a variable that *could* be added as a valid control variable in W, but does not *need to* be for EXOG to hold. This is perhaps harder to motivate, but it is certainly weaker than the above. Abadie (2003) similarly considers the identification of mean attributes of IV compliers, when those attributes represent valid control variables.[41] The result of Proposition 3 is also related to an intermediate result used in the proof of Theorem 1 in Hoderlein et al. (2016).

A particularly simple special case occurs when A_i is binary. Then (31) yields $P(A_i = 1|H_i = \tau_{V_i}(r), x, w)/P(A_i = 1|x, w) = \partial_{x_j} P(R_i \leq r|A_i = 1, x, w)/\mathbb{E}[\partial_{x_j} P(R_i \leq r|X_i =$

[41] In the case of complier characteristics, the LATE monotonicity assumption plays a role analogous to the assumption that $sgn(\partial_{x_j} h(x, U_i))$ is common across i in Proposition 3. Analogously, the compliers are not individually identified.

$x, A_i)|x, w]$. As a consequence, we then have that:

$$\frac{P(A_i = 1|H_i = \tau_{V_i}(r), x, w)/P(A_i = 0|H_i = \tau_{V_i}(r), x, w)}{P(A_i = 1|x, w)/P(A_i = 0|x, w)} = \frac{\partial_{x_j} P(R_i \leq r|A_i = 1, x, w)}{\partial_{x_j} P(R_i \leq r|A_i = 0, x, w)} \tag{33}$$

This says that, for example, the ratio of the local regression derivative for $R_i \leq r$ between the male and female subsamples reveals the odds (conditional on $X_i = x$) of being a woman for the marginal respondents of response category r, as compared to the odds of being a woman for *all* respondents (including the infra-marginal ones).

The simplest implementation of Eq. (32) would take the conditional expectation of $A_i \cdot \mathbb{1}(R_i \leq r)$ to be linear in x and w, in addition to assuming a linear probability model for $\mathbb{1}(R_i \leq r)$. Given this restriction, the identified quantity

$$\mathbb{E}[A_i|h(x, U_i) = \tau_{V_i}(r), X_i = x, W_i = w] = \frac{\partial_{x_j} \mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r)|x, w]}{\partial_{x_j} P(R_i \leq r|x, w)}$$

does not depend on x or w, and thus the ratio of the coefficient on X_{ji} in these two regressions identifies $\mathbb{E}[A_i|H_i = \tau_{V_i}(r)]$. The above is implemented in Figure 4, which takes X_{ji} to be i's household income. Proposition 3 could equally well have been applied using the regression coefficients for PUMA income instead under the same assumptions.

The bottom panels of Figure 4 similarly approximate the relevant regressions with linear probability models, which in turn implies that the local relative odds

$$\frac{P(A_i = 1|H_i = \tau_{V_i}(r), X_i = x, W_i = w)/P(A_i = 0|H_i = \tau_{V_i}(r), X_i = x, W_i = w)}{P(A_i = 1|X_i = x, W_i = w)/P(A_i = 0|X_i = x, W_i = w)}$$

do not depend on x or w for a given r.

E What would be identified with a smooth reporting function

This section considers the question considered in Section 6.2 from the main text: when can discrete differences in the conditional expectation of R given X be interpreted quantitatively, by comparing their magnitude to that of a regression derivative? This question motivates a comparison of the Lemma 2 result for a setting with discrete response categories to a hypothetical case in which the space of responses $\mathcal{R}$ were instead a continuum. Then I consider such a continuum as a limit of richer and richer response spaces, which is necessary to develop some of the formal results in Section 6.2 of the main paper.

E.1 Comparing discrete and continuous regressors

Beginning with (21) from the main text, our goal is to examine the causal interpretation of $\frac{\mathbb{E}[R_i|X_i=x', W_i=w] - \mathbb{E}[R_i|X_i=x, W_i=w]}{\partial_{x_1} \mathbb{E}[R_i|X_i=x'', W_i=w]}$ outside of the limit that Δ_i is very small (so that Eq. (19) holds). Under the assumptions of Lemma 2, if $\mathcal{R} = \{0, 1, \ldots, \bar{R}\}$ then for each j

that satisfies REG_j:

$$\partial_{x_j}\mathbb{E}[R_i|x,w] = \mathbb{E}\left\{ \sum_r f_H(\tau_{V_i}(r)|x,V_i,w)\cdot \mathbb{E}\left[\partial_{x_j}h(x,U_i)|H_i=\tau_{V_i}(r),x,V_i,w\right] \middle| W_i = w \right\}$$

Combining this with Eq. (20), we know that the ratio in Eq. (21) is equal to

$$\frac{\mathbb{E}\left[\sum_r \bar{f}_H(\tau_{V_i}(r)|\Delta_i,x,V_i,w)\cdot\Delta_i|W_i=w\right]}{\mathbb{E}\left\{\sum_r f_H(\tau_{V_i}(r)|x'',V_i,w)\cdot\mathbb{E}\left[\partial_{x_1}h(x'',U_i)|H_i=\tau_{V_i}(r),x'',V_i,w\right]|\,W_i=w\right\}} \tag{34}$$

To interpret this as informative about the relative magnitudes of Δ_i and $\partial_{x_1}h(x'',U_i)$, the relevant question is how similar the sum $\sum_r f_H(\tau_{V_i}(r)|x'',V_i,w)$ over densities at the thresholds is to the corresponding sum over mean densities: $\sum_r \bar{f}_H(\tau_{V_i}(r)|\Delta_i,x,V_i,w)$, at least on average. If these quantities tend to be close to one another in magnitude, then Eq. (21) uncovers something close to the ratio of two convex averages of causal effects. If they differ by an unknown amount, then interpreting (21) in terms of the relative magnitudes of causal effects is not possible.

Reasoning about the magnitudes involved in (34) is challenging in full generality, but it is possible to derive analytical results to guide our intuition by assuming that there are "many" response categories in $\mathcal{R}$. Given the definition of $\bar{f}$, notice that $\sum_r f_H(\tau_v(r)|x'',v,w)$ and $\sum_r \bar{f}_H(\tau_v(r)|\Delta,x,v,w)$ are similar for a given (Δ,v,w) if

$$\sum_r \frac{1}{\Delta}\int_{\tau_v(r)-\Delta}^{\tau_v(r)} f_H(y|\Delta,x,v,w)dy \approx \sum_r f_H(\tau_v(r)|x'',v,w) \tag{35}$$

Observe that the two sides of (35) can *only* differ because the summation occurs over H_i evaluated at the discrete thresholds $\tau_v(r)$. If instead the sums over r were replaced by integrals over all possible values of H_i, we would have $\int\left\{\frac{1}{\Delta}\int_{h-\Delta}^{h} f_H(y|\Delta,x,v,w)dy\right\}dh = \int f_H(h|x'',v,w)\cdot dh$, which holds trivially because both sides evaluate to unity for any Δ,v,x,w and x''.[42] Thus it would seem that we have a second "limit" in which discrete and continuous regression differences can be compared: when there are many response categories. However, I show below that discrete sums over the thresholds do not exactly correspond to equal-weighted integrals over h in the limit of a continuum of response categories. Rather, in this limit the integrals also involve the quantity $r'(h,v)$, which measures how responsive response function v is at h. Nevertheless, the intuition provided by the above logic suggests that looking at the limit of many categories may provide a tractable means of evaluating the quality of Eq. (35) as an approximation.

E.2 Continuous regressors with a continuum of responses

As a benchmark, this section imagines an intermediate situation in which respondents can select a response from some bounded continuum in $\mathcal{R}$. This allows us to separate

[42] This is immediate for the RHS, which integrates a density. To see it for the LHS, reverse the order of integrals to obtain $\int dy\cdot f_H(y|\Delta,x,v,w)\left\{\frac{1}{\Delta}\int_y^{y+\Delta}dh\right\}=1$.

the effect of reporting heterogeneity from that of information loss due to discretization of the latent variable H_i into categories.

Suppose $\mathcal{R}$ is a convex subset of $\mathbb{R}$, for simplicity $\mathcal{R} = [0, \bar{R}]$ for some maximum response value $\bar{R}$. Figure 7 depicts two examples of reporting functions on this continuum of responses.

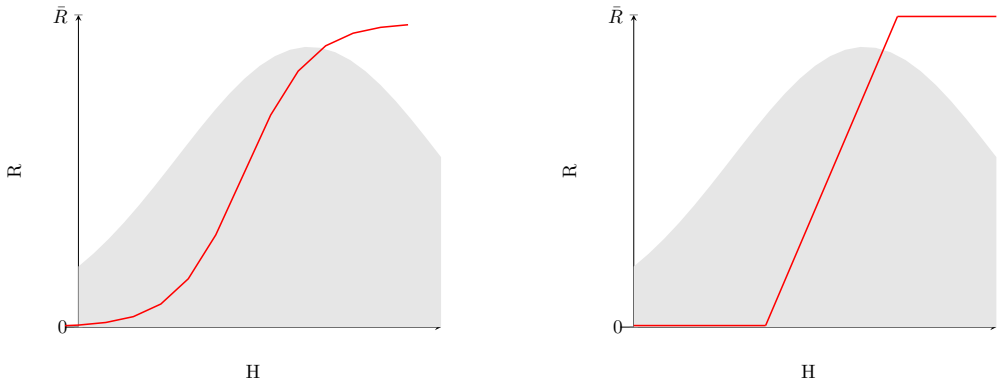


Figure 7: Example of two "continuous" reporting functions, with the density of H depicted in gray.

While the example on the left side of Figure 7 is a smooth sigmoid shape mapping $\mathbb{R}$ to the interval $[0, \bar{R}]$, the piecewise-linear reporting function on the right has kinks at $\tau_v(0)$ and $\tau_v(\bar{R})$ beyond which the function is flat. Nevertheless, we may define a derivative function $r'(h, v)$ of any given $r(h, v)$ with respect to h, which by virtue of MONO can only fail to exist only at isolated points in $\mathcal{H}$ for a given v.[43] Provided that H_i is continuously distributed, it therefore does not affect results to treat $r'(h, v)$ as defined for all h. With "smooth reporting", we have the following analog of Lemma 2:

Proposition 4. *Assume MONO, EXOG and REG for at least one j, with $\mathcal{R}$ a convex subset of $\mathbb{R}$. Then:*

$$\nabla_x \mathbb{E}[R_i|x, w] = \int dF_{V|W}(v|w) \int dh \cdot r'(h, v) \cdot f_H(h|x, v, w) \cdot \mathbb{E}\left[\nabla_x h(x, U_i)|h, x, v, w\right] \quad (36)$$

provided the "boundary condition": $\lim_{h \to \pm\infty} f_H(h|x, v, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = h, x, v.w\right] = 0$, *i.e. average partial effects do not explode for extreme values of H_i, any faster than the density of H_i falls off in h, for each v and j satisfying REG.*

The proof of Proposition 4 makes use of a result of Kasy (2022) that relates derivatives of the density of an outcome with respect to policy variables, to the rate of change of the "flow density" quantity introduced in the discussion of Lemma 2.

We can compare this expression to what would be recovered by the infeasible regression

[43] This is an application of "Lebesque's theorem" that monotone functions are differentiable almost everywhere.

of H_i on X_i and W_i (i.e. if H_i were observed):

$$\nabla_x \mathbb{E}[H_i|x,w] = \int dF_{V|W}(v|w) \int dh \cdot 1 \cdot f_H(h|x,v,w) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|H_i = h,x,v,w\right] \tag{37a}$$

And with integer categories $\mathcal{R}$, using Lemma 2:

$$\nabla_x \mathbb{E}[R_i|x,w] = \int dF_{V|W}(v|w) \sum_r f_H(\tau_v(r)|x,v,w) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|H_i = \tau_v(r),x,v,w\right] \tag{37b}$$

These three expressions differ only in what multiplies $f_H(h|x,v,w) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|h,x,v,w\right]$ for various values of h. Relative to (37a), (36) introduces the derivative $r'(h,v)$ of the reporting function. Intuitively, $r'(h,v)$ corresponds to how closely spaced the thresholds are near a given value of h. If this spacing varies across the support of h, causal effects will be up-weighted for the h where $r'(h,v)$ is largest, relative to the h where the derivative is smaller. Comparing (37b) to (36) shows that using subjective responses with discrete categories further involves information loss due to the discretization: the integral over all h is replaced by a sum over the thresholds $\tau_v(r)$.[44]

E.3 The "dense response limit" of many categories

In practice, survey questions do not typically allow individuals to give any real number (within a range) in response to subjective questions. However, results based on Proposition 4 provide a more tractable setting to derive analytical results. If $\mathcal{R}$ is sufficiently rich, then this will provide a useful approximation to the actual properties of that setting (e.g. Benjamin et al., 2014 elicits life-satisfaction data with 100 categories). Below, I give a formal definition of this "dense response limit" corresponding to an integer response space $\mathcal{R} = \{0,1,\ldots,\bar{R}\}$, which proves useful in the analysis of Section 6.2. Appeal to this limit is indicated by the symbol $\overset{R}{\to}$ in the results of Section 6.2.

To define the dense response limit for a fixed $\bar{R}$, consider a sequence of response spaces $\mathcal{R}_n = \{0, 1/n, 2/n, \ldots, (n\bar{R})/n\}$ where note that $n\bar{R}$ has $n\bar{R}+1$ categories ranging from 0 to $(n\bar{R})/n = \bar{R}$. For a fixed value of reporting heterogeneity v, consider a sequence of reporting functions $r_n(\cdot,v)$ indexed by n, and let $\tau_{v,n}(\cdot)$ be a function from $\mathcal{R}_n$ to $\mathbb{R}$ representing the thresholds corresponding to each function $r_n(\cdot,v)$ in the sequence.

Definition (dense response limit). *Fix a $v \in \mathcal{V}$. Consider a sequence of reporting functions $r_n(\cdot,v)$ for $n \to \infty$. We say that the sequence converges to response function*

[44] In the case of linear reporting functions with a continuous response space, Proposition 4 generalizes a result of Greene (2005) for marginal effects in the double-censored Tobit model. The Tobit model takes a linear structural model $h(x,u) = x^T\beta + u$. Greene shows that if the error term u has any continuous distribution, a marginal effect is equal to the true structural effect times the probability that an observation is not censored at either endpoint. (36) with no covariates w reduces to $\partial_{x_1}\mathbb{E}[R_i|x] = \beta_1 \cdot \int dF_V(v) \cdot \frac{\bar{R}}{\mu(v)-\ell(v)} \cdot P(0 < R_i < \bar{R}|x,v)$ using that $r'(h,v) = \frac{R}{\mu(v)-\ell(v)} \cdot \mathbb{1}(\ell(v) < h < \mu(v))$. The traditional Tobit model further treats V_i as degenerate with $\mu - \ell = R$, so the above recover's Greene's result that $\partial_{x_1}\mathbb{E}[R_i|X_i = x] = \beta_1 \cdot P(0 < R_i < R|X_i = x)$.

$r(\cdot, v)$ in the dense response limit, denoted as $r_n(\cdot, v) \xrightarrow{R} r(\cdot, v)$, if:

$$\lim_{n\to\infty} \tau_{v,n}(r_n) = \tau_v(r)$$

for any sequence of $\{r_n\}_{n=1}^{\infty}$ where $r_n \in \mathcal{R}_n$ for each n, such that $\lim_{n\to\infty} r_n = r$ for some $r \in [0, \bar{R}]$ (according to the Euclidean metric on the reals). For any functional of all response functions $\theta(\{r_n(\cdot, v)\}_{v\in\mathcal{V}})$, let $\theta(r_n) \xrightarrow{R} \Theta$ denote that Θ evaluates the functional θ at the limiting family of response functions: $\Theta = \theta(r)$.

Intuitively, if the actual response scale is the integers 0 to $\bar{R}$, the dense response limit instead approximates reports as taking on any real number in $[0, \bar{R}]$.

As a concrete example, consider linear response function $r(h, v) = \min\{\bar{R}, \max\{0, h\}\}$ ranging from 0 to $\bar{R}$ on the continuum $\mathcal{R} = [0, \bar{R}]$. Consider the sequence of reporting functions $r_n(h, v) = \max_{r\in\mathcal{R}_n : h \le \tau_{v,n}(r)} r$, where we let the thresholds be $\tau_{v,n}(r) = r$ for each $r \in \mathcal{R}_n, r < \bar{R}$ (recall that $\tau_{v,n}(r) = \infty$ when r is equal to it's highest value in the response space, in this case $\bar{R}$). The response function $r_n(h, v)$ then represents a "staircase" function that jumps from the r^{th} category $(\frac{r-1}{n})$ to the $(r+1)^{th}$ category $(\frac{r}{n})$ at $\tau_{v,n}((r-1)/n) = (r-1)/n$. In this case $r_n(\cdot, v) \xrightarrow{R} r(\cdot, v)$ in the dense response limit, because for any sequence $\{r_n\}_{n=1}^{\infty}$ such that $\lim_{n\to\infty} r_n = r \in [0, \bar{R}]$ (for example $r_n = \max_{r'\in\mathcal{R}_n : r' \le r} r'$) we have that $\lim_{n\to\infty} \tau_{v,n}(r_n) = \lim_{n\to\infty} r_n = r$.

In the dense response limit, discrete differences in the mean of R_i depends upon the average slope $r'(h, V_i)$ of the response function $r(\cdot, V_i)$ for h between H_i and $H_i + \Delta_i$:

Proposition 5. *Under MONO, EXOG, and REG, then in the dense response limit*

$$\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w] \xrightarrow{R} \bar{R} \cdot \mathbb{E}[\Delta_i \cdot \bar{r}'(H_i, \Delta_i, V_i)|X_i = x, W_i = w]$$

where $\bar{r}'(y, \Delta, v) := \frac{1}{\Delta} \int_y^{y+\Delta} r'(h, v) \cdot dh$.

Since $\bar{r}' \ge 0$, the weights on Δ_i in Proposition 5 are positive and aggregate to[45]

$$\Pi_{x,x'} := \bar{R} \cdot \mathbb{E}[\bar{r}'(H_i, \Delta_i, V_i)|X_i = x, W_i = w] \tag{38}$$

Proposition 4 in Appendix E derives an analogous result to Proposition 5 for regression derivatives in the case of a continuous component of x. That result shows that the total weight on causal effects in a derivative $\partial_{x_j}\mathbb{E}[R_i|x, w]$ is, by comparison:

$$\Pi_x := \bar{R} \cdot \mathbb{E}[r'(H_i, V_i)|X_i = x, W_i = w] \tag{39}$$

For ease of notation, I leave the dependence of quantities Π_x and $\Pi_{x,x'}$ on the value of the control variables W_i implicit.

A comparison of Π_x and $\Pi_{x,x'}$ allows us to interpret the relative magnitudes of discrete

[45]Note that if Δ_i and $\bar{r}'(H_i, \Delta_i, V_i)$ are uncorrelated conditional on $X_i = x, W_i = w$, then we can further write the RHS of Proposition 5 as $\mathbb{E}[\Delta_i|X_i = x] \cdot \bar{R} \cdot \mathbb{E}[\bar{r}'(H_i, \Delta_i, V_i)|X_i = x]$.

and continuous differences in $\mathbb{E}[R_i|X_i = x, W_i = w]$, as in Eq. (21). If we have, for example, a binary X_1 and continuous X_2, and we let $x' = (1, x_2)$ and $x = (0, x_2)$ for some $x_2 \in \mathbb{R}$, then:

$$\frac{\mathbb{E}[R_i|X_i = x', W_i = w] - \mathbb{E}[R_i|X_i = x, W_i = w]}{\partial_{x_1}\mathbb{E}[R_i|X_i = x, W_i = w]} \xrightarrow{R} \frac{\tilde{\beta}_2(x, x', w)}{\tilde{\beta}_1(x'', w)} \cdot \frac{\Pi_{x,x'}}{\Pi_x} \tag{40}$$

where $\tilde{\beta}_1(x'', w)$ is a convex weighted average over the (derivative) causal effect of X_1 on H and $\tilde{\beta}_2(x, x', w)$ is a convex weighted average over causal effects of X_2 on H. If the aggregate weights are close in magnitude, i.e. $\Pi_{x,x'}/\Pi_x \approx 1$, then we can identify the relative magnitudes of these causal averages to a good approximation.

E.4 Heterogeneous linear reporting in the dense response limit

To assess whether the approximation that $\Pi_{x,x'}/\Pi_x \approx 1$ is plausible, I impose a further simplification. Let us say that *heterogeneous linear reporting* holds with $\mathcal{R} = \{0, 1, \dots \bar{R}\}$ if each individual spaces out the thresholds $\tau_v(r)$ evenly within some individual-specific range, i.e. $\tau_v(r) = \ell(v) + r \cdot \frac{\mu(v)-\ell(v)}{\bar{R}}$ where $\ell(v) = \tau_v(r)$ is the threshold between the two lowest categories for an individual with $V_i = v$, and $\mu(v)$ is the threshold between the top two categories.[46]

Heterogeneous linear reporting captures the idea that response functions are "linear", while still allowing them to vary by individual. Heterogeneous linear reporting may be a reasonable assumption if individuals aim to maximize the informativeness of their responses by equally spreading out the response categories (van Praag, 1991), given their subjective definitions $\ell(v)$ and $\mu(v)$ of the minimum and maximum category thresholds.[47] Kaiser and Vendrik (2022) summarize empirical evidence in support of linearity, for example from asking individuals directly about their response thresholds, or asking about verifiable outcomes such as an individual's height.

With heterogeneous linear reporting, a partial identification result holds analytically in the dense response limit:

Proposition 6. *Suppose that the following hold in addition to MONO,EXOG,REG:*

1. $r(h, v) \xrightarrow{R} \ell(v) + \frac{h-\ell(v)}{\mu(v)-\ell(v)}$, *i.e. reporting is (heterogeneously) linear in the dense response limit; and*

2. *For each* Δ *in the support of* Δ_i, $f_H(h|\Delta, x, v, w)$ *is increasing on the interval* $[\ell(v) - |\Delta|, \ell(v) + |\Delta|]$, *and decreasing on the interval* $[\mu(v) - |\Delta|, \mu(v) + |\Delta|]$

[46] Note that in the limit of many categories $\bar{R}$, this can be well approximated by the linear reporting function $\lim_{\bar{R}\to\infty} \frac{r(h,v)}{\bar{R}} = \mathbb{1}(\ell(v) \le h \le \mu(v)) \cdot \frac{h-\ell(v)}{\mu(v)-\ell(v)}$.

[47] Many studies justify the use of regression based approaches to studying subjective data R_i by interpreting such data as a direct measurement of H_i. However, the function $r(\cdot, v)$ cannot literally be the identity function if $\mathcal{R}$ is a set of integers, unless we think that "true" happiness also only takes integer values. We might view the cardinality approach as instead supposing that $r(h)$ is homogeneous across individuals and that the thresholds $\tau(r)$ are equally spaced apart.

Then

$$\frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x + \Pi_{x'})} \in [1, 2],$$

Furthermore, suppose that the lengths of reporting intervals are not too variable across individuals relative to variability in bunching at the endpoints 0 *and* $\bar{R}$*, in the sense that*

$$Var\left[\frac{1}{\mu(V_i) - \ell(V_i)}\middle| x, w\right] \leq Var\left[\mathcal{B}_i | x, w\right] \cdot \mathbb{E}\left\{\frac{1}{\mu(V_i) - \ell(V_i)}\middle| x, w\right\}^2,$$

where $\mathcal{B}_i := P(R_i = 0 \text{ or } R_i = \bar{R}|X_i, V_i, W_i)$*, then*

$$\frac{1}{2} \leq \frac{\Pi_{x,x'}}{\Pi_x} \leq \frac{1}{(1 - \mathbb{E}[\mathcal{B}_i|x, w])^2},$$

Proposition 6 provides two sets of bounds on the ratio of the total weight on causal effects in $\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w]$, to the total weight on causal effects in a derivative $\partial_{x_j}\mathbb{E}[R_i|x, w]$. The first bound, $\frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x + \Pi_{x'})} \in [1, 2]$ implies that, in the setup of Eq. (40):

$$\frac{\mathbb{E}[R_i|X_i = x', W_i = w] - \mathbb{E}[R_i|X_i = x, W_i = w]}{\frac{1}{2}\partial_{x_1}\mathbb{E}[R_i|X_i = x', W_i = w] + \frac{1}{2}\partial_{x_1}\mathbb{E}[R_i|X_i = x, W_i = w]} \xrightarrow{R} \theta \cdot \frac{\tilde{\beta}_2(x, x', w)}{\tilde{\beta}_1(x, x', w)} \tag{41}$$

where θ is some number between 1 and 2, and $\tilde{\beta}_1(x, x', w)$ is a convex combination of $\partial_{x_1} h(X_i, U_i)$. This bound requires no assumptions on how variable the happiness scale lengths $\mu(V_i) - \ell(V_i)$ can be across individuals with different V_i.

By contrast, the second set of bounds requires us to assume that the coefficient of variation of $\frac{1}{\mu(V_i) - \ell(V_i)}$ is no greater than the standard deviation of $\mathcal{B}_i$, conditional on X_i and W_i. Assuming homogeneity of reporting functions makes the coefficient of variation zero, trivially satisfying the assumption. More generally, the stringency of the assumption can be evaluated from the data by a nonparametric regression of observed bunching at the endpoints of the scale (0 and $\bar{R}$) on X_i and W_i.

Note that if the additional restriction justifying the second set of bounds holds, and $\mathbb{E}[r'(H_i, V_i)|X_i = x, W_i = w]$ is roughly constant in x, then

$$\frac{\Pi_{x,x'}}{\Pi_x} \approx \frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x + \Pi_{x'})}$$

and we can take the intersection of the two sets of bounds: $[1, 1/(1 - \mathbb{E}[\mathcal{B}_i|x, w])^2]$. This bound will be very narrow if there are few endpoint bunchers when $X_i = x, W_i = w$.

E.5 Simulation evidence on Proposition 6

To gather some further suggestive evidence on the comparability of estimates that use discrete vs. continuous variation in X. I in this section simulate some data-generating-processes (DGPs) for H_i and for the response functions $r(\cdot, V_i)$, computing the quantities $\Pi_{x,x'}$, Π_x and $\Pi_{x'}$ numerically given the population DGP. This section is abbreviated

for brevity; a much more extensive set of simulations is reported in Goff (2025). To summarize the results found there, I focus on a particularly simple class of DGPs.

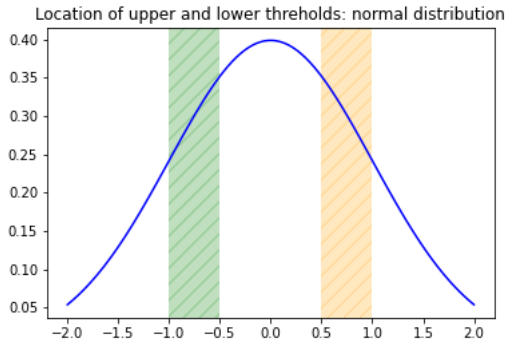
The DGP is set such that EXOG holds with no covariates W_i, and consider a researcher comparing $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ to $\partial_{x_j}\mathbb{E}[R_i|X_i = x]$ and $\partial_{x_j}\mathbb{E}[R_i|X_i = x']$ for some given values x' and x, and regressors X_j. Given the results of the last section, we seek to compare $\Pi_{x,x'}$, Π_x and $\Pi_{x'}$ to understand the relative weights each of these estimands place on causal effects. An "optimistic" picture emerges if $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'}) \approx 1$.

I this section I suppose that heterogeneous linear reporting holds, and investigate deviations from the assumptions of Proposition 6 only in that $\bar{R}$ is finite (so we are outside of the dense response limit), and that Item 2. in the statement of Proposition 6 may not hold. Additional simulations in Goff (2025) report results from DGPs that relax heterogeneous linear reporting, but the range of deviations from $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'}) \approx 1$ found in that case were no greater than in the case with (heterogenous) linear reporting.

I take $H_i|X_i = x$ to have a standard normal distribution. Note that since the overall location and scale of the happiness distribution is not inherently meaningful, this choice of mean and variance is arbitrary. Additional simulations reported in Goff (2025) replicate similar behavior to the results reported below, when the conditional H_i is not normally distributed and exhibits skewness, non-unimodality, or bounded support.

Individual reporting functions can be characterized by $\ell(v)$, the value of happiness at which an individual with $V_i = v$ moves from response category 0 to response category 1, and $\mu(v)$, the value at which this individual would move from category $\bar{R}-1$ to the highest category $\bar{R}$. Next, I suppose that individuals' values of $\ell(V_i)$ are distributed uniformly between -1 and -0.5, and that $\mu(V_i)$ is independent of $\ell(V_i)$ and distributed uniformly from $[0.5, 1]$. The left panel of Figure 8 provides a visualization. These choices aim to reflect a world in which while individuals differ e.g. in the point $\mu(V_i)$ at which they would report $R = \bar{R}$, this threshold for the highest possible category is for all individuals at least above the mean level of happiness in the population.

The right side of Figure 8 computes $\frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x+\Pi_{x'})}$ from Eqs. (38) and (39) by drawing a population of 1000 reporting functions, and assigning each a value of $H_i|X_i = x$ independent of V_i (i.e. imposing $U_i \perp\!\!\!\perp V_i$). Figure 8 generally provides an optimistic picture that $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'}) \approx 1$, even when $\bar{R}$ is small (and thus the dense response limit does not hold). The table on the right side of Figure 8 reports $\frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x+\Pi_{x'})}$ as a function of the number of response categories $\bar{R} \in [2, 5, 11, 100]$, supposing a constant treatment effect Δ which is varied from -0.5 to 5. Note that the results can be interpreted as reporting conditional analogs of the quantity $\frac{\Pi_{x,x'}}{\frac{1}{2}(\Pi_x+\Pi_{x'})}$ among individuals sharing a value of $\Delta_i = h(x', U_i) - h(x, U_i)$, in a setting in which H_i is independent of treatment effects Δ_i, conditional on $X_i = x$.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	1.017758	1.016489	1.018028	1.018884
-0.1	1.000335	1.000441	1.000664	1.000809
0.1	1.000837	1.000283	1.001079	1.001052
0.25	1.003905	1.005432	1.004132	1.003522
0.5	1.020549	1.019535	1.017904	1.014529
1	1.060440	1.062557	1.061607	1.051899
5	0.504738	0.531706	0.544236	0.549753
1/NB	1.867396	1.878186	1.874115	1.873973

Figure 8: $H_i|X_i = x$ is standard normal, and 1000 reporting functions are drawn from $\ell(v) \sim U[-1, 1/2]$, $\mu(v) \sim U[1/2, 1]$. The left panel depicts the supports of $\ell(v)$ (green) and $\mu(v)$ (yellow) with the density of H_i. The right panel reports values of $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

Proposition 6 implies that as $\bar{R} \to \infty$, $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'})$ should lie between 1 and 2, for any values Δ such that $\ell(V_i) < -|\Delta|$ and $\mu(V_i) > |\Delta|$ for all V_i (so that $f_H(h|x)$ is increasing on the interval $[\ell(V_i) - |\Delta|, \ell(V_i) + |\Delta|]$, and analogously for $\ell(v)$). This is true for all of the values reported in Figure 8, aside from $\Delta = 1$ and $\Delta = 5$. In all but the case of $\Delta = 5$, $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'})$ is in fact quite close to unity, well within the refined bounds $[1, 1/NB]$ which holds under the variance restriction in Proposition 6, where $NB = P(0 < R_i < R|X = x)$ is the "non-bunching" probability.

With the exception of $\Delta = 5$, the standard-normal DGP reported in Figure 8 provides an optimistic picture that $\{\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]\}/\partial_{x_j}\mathbb{E}[R_i|X_i = x]$ uncovers something close to a ratio of weighted averages of causal effects $\tilde{\beta}_2/\tilde{\beta}_1$ in the setting described by Equation (40). In this case, results do not differ substantially whether the number of response categories is small (e.g. $\bar{R} = 2$, binary response) or e.g. $\bar{R} = 100$.

The $\Delta = 5$ case nevertheless shows that the ratio in the case of (40) may be quite misleading *in principle*. The $\bar{R} = 2$ value of $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'}) \approx 0.5$ means that the magnitude of β_1 relative to that of β_2 would be under-estimated by a factor of 2, when using $x' = (1, x_2)$ and $x = (0, x_2)$ in a linear model $h(x, u) = \beta_1 x_1 + \beta_2 x_2$. On the other hand, it is implausible that binary treatment variable being analyzed would have an effect on happiness that is 5 times the variance of happiness in the population. Note that $\Delta = 5$ with $H_i|\{X_i = x\} \sim \mathcal{N}(0, 1)$ violates Item 2. in the assumptions of Proposition 6, given e.g. that all $\ell(v)$ within 5 of where the conditional density of H_i begins decreasing.

While the quantity $\Pi_{x,x'}/\frac{1}{2}(\Pi_x + \Pi_{x'})$ averages over the reporting heterogeneity in the population, Figure 9 disaggregates this by V_i. Define $\delta_{\Delta,x,v} := \frac{\sum_r \bar{f}_H(\Delta,\tau_v(r),x,v) - \sum_r f_H(\tau_v(r)|x,v)}{\sum_r f_H(\tau_v(r)|x,v)}$. An individual with $X_i = x$ and $V_i = v$ will receive similar weights when using either dis-

crete or continuous variation at x if $\delta_{\Delta,x,v} \approx 0$. Write Eq. (20) as:

$$\mathbb{E}[R_i|x'] - \mathbb{E}[R_i|x] = \int dF_{V|W}(v|w) \cdot \left(\sum_r f_H(\tau_v(r)|x,v) \right) \cdot \mathbb{E}[\Delta_i|X_i = x, V_i = v] + \int dF_{V|W}(v|w) \cdot \int d\Delta \cdot f_H(\Delta|x,v) \cdot \Delta \cdot \delta_{\Delta,x,v}$$

Figure 9 reports the distributions of $\frac{\sum_r \bar{f}_H(\Delta,\tau_v(r),x,v)}{\sum_r f_H(\tau_v(r)|x,v)} = 1 + \delta_{\Delta,x,v}$, across 1000 reporting functions sampled the same as in Figure 8. The distributions of δ_{Δ,x,V_i} are approximately unimodal in each case, with a variance that tends to increase with the magnitude of Δ.

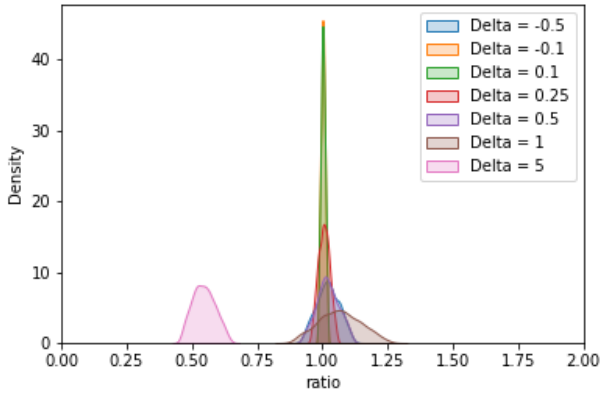


Figure 9: The distribution of $1 + \delta_{\Delta,x,V_i}$ across V_i is depicted across alternative values of Δ_i, with $H_i|X_i = x$ standard normal, $\bar{R} = 100$, and 1000 reporting functions are drawn from $\ell(v) \sim U[-1, 1/2]$, $\mu(v) \sim U[1/2, 1]$.

In line with Proposition 6, the distributions are centered at or slightly above unity, except in the case of $\Delta = 5$.

F Further details on the empirical application

F.1 Sample construction

I use three data sources in my replication and extension of Luttmer (2005). First, I access the public microdata files for the 1987 and 1992 waves of the NLSF from ICPSR, which constitutes a nationally representative sample of individuals nineteen or older (and able to speak English or Spanish). This provides the variables R_i, X_{1i}, and W_i in e.g. Eq. (18). I follow Luttmer (2005) in deflating monetary values using the consumer price index from the Bureau of Labor Statistics CPI-U series.

Although accessing the geo-coded data from the NLSF is not currently supported, I obtained the predicted PUMA-level log-earnings variable X_{2i} and PUMA identifiers (for clustering standard errors) through correspondence with the author and a data sharing agreement with the Social Sciences Research Services at the University of Wisconsin. I thank Erzo F.P. Luttmer for providing this variable to me and the cooperation of the

University of Wisconsin. By merging these data with the publicly available data (by NLSF caseid and wave) and keeping only observations that are matched, I automatically implement the sampling restriction of Luttmer (2005) to respondents who were married or cohabiting in both waves of the NLSF. In the regressions reported, the sample used throughout is that of the OLS regressions with controls, i.e. observations for which none of R_i or any of the components of X_i, W_i are missing.

F.2 Details on the DML estimator, and nonparametric alternatives

The DML estimator for average partial derivatives of Klosin and Vilgalys (2023) uses Lasso to learn the function $\mathbb{E}[R_i|X_i, W_i]$ and, following Chernozhukov et al. (2022), debiases the final estimates using the Riesz representation of the average partial derivative estimand.[48] As in Chernozhukov et al. (2022) (see Appendix B.1 there), two such average partial derivative estimators have a $\sqrt{n}$ jointly normal asymptotic distribution.

I limit the allowed interactions between the variables (X_i, W_i) in the set of basis functions considered by the Lasso. In particular, I allow interactions between own-income, employment, and PUMA-income, and constrain the additional variables W_i to enter linearly. This renders estimator semi-parametric rather than fully nonparametric, but makes the bootstrap standard error calculations computationally feasible. I use bootstrap rather than the analytical standard errors to accomodate clustering by PUMA. Consistency of the bootstrap follows from standard arguments (van der Vaart, 2000), given that the DML estimator is asymptotically linear with a Gaussian limiting distribution.

Effectively Column (3) of Table 2 thus employs a semi-parametric estimator that assumes the partially linear form $\mathbb{E}[R_i|X_i, W_i] = f(X_{1i}, X_{2i}, W_{1i}) + \lambda^T W_{-1,i}$, where W_i is a dummy variable for unemployment and $W_{-1,i}$ denotes the other variables in W_i. Conceptually, the coefficients γ_j from Eq. (18) are now replaced with $\gamma_j(x, w_1) := \partial_{x_j}\mathbb{E}[R_i|X_i = x, W_i = w]$, and first two rows of Column (3) report $\gamma_j(X_i, W_{1,i})$ averaged across the empirical distribution of X_i. The row labeled "Ratio PUMA/own" reports the ratio of these two, i.e. $\mathbb{E}[\partial_{x_2}\mathbb{E}[R_i|X_i, W_i]]/\mathbb{E}[\partial_{x_1}\mathbb{E}[R_i|X_i, W_i]]$.

I use the `rlassoAutoDML` function from the `hdm` package in R. I set the polynomial order to two among (X_i, W_i), and use 5 folds in the cross-fitting stage.

For comparison, Table 3 shows results for the average derivatives and their ratio using a local-linear regression. This procedure is implemented without control variables, and the variance covariance matrix for the average derivatives is obtained by bootstrap (Cattaneo and Jansson, 2018). The standard error for the ratio of average derivatives is then calculated by the delta method. Bandwidths for the local-linear regression are chosen by cross-validation, using the `npregress kernel` function in Stata.[49]

[48] The estimator of Klosin and Vilgalys (2023) involves two modifications of the DML average derivative estimator previously proposed by Chernozhukov et al. (2022). First, it relies on analytical derivatives of the basis functions which reduces the computational burden relative to numerical differentiation. Second, it introduces an iterative procedure for the debiasing step by directly solving the corresponding convex optimization problem.

[49] In computing average derivatives, this command drops observations for which the local kernel-weighted design matrix is close to singular, which results in a loss of some observations. Table 3 reports the size of the full sample passed to `npregress kernel`.

	(1) OLS	(2) Local linear
Own ln income	0.0446***	-0.246***
	(0.0113)	(0.0223)
PUMA ln income	-0.169**	-1.971***
	(0.0614)	(0.0696)
Ratio PUMA/own	-3.792	-1.971
se(ratio)	1.534	0.401
Controls		
Clustered se	X	X
Sample size	7939	7939

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 3: Replication of Luttmer (2005)'s results for the main respondent, without control variables. Without control variables, a fully nonparametric kernel regression (local-linear) estimator is feasible. Standard errors for the average partial derivatives and their ratio obtained by cluster bootstrap, 500 replications.

F.3 Estimates by response category in table form

	(1) R≤1	(2) R≤2	(3) R≤3	(4) R≤4	(5) R≤5	(6) R≤6
Own ln income	-0.000281	0.00205	0.00829	0.0211**	0.0405***	0.0159*
	(-0.14)	(0.73)	(1.96)	(2.95)	(4.67)	(2.05)
PUMA ln income	-0.00798	-0.0151	-0.0242	-0.0448	-0.0891**	-0.0480
	(-1.16)	(-1.62)	(-1.69)	(-1.93)	(-2.75)	(-1.65)
Ratio PUMA/own	28.44	-7.360	-2.920	-2.121	-2.198	-3.009
se(ratio)	201.4	10.82	2.211	1.330	0.896	2.288
Sample size	7939	7939	7939	7939	7939	7939

t statistics in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 4: Coefficients from a linear probability model for each response category. All regressions include the controls from Table 2 and standard errors clustered by PUMA.

F.4 Relationship between the density at each threshold r and $P(R_i = r)$

Suppose that the causal effect were a constant $\partial_{x_1} h(x, U_i) = \beta_1$ for all i. Then, Lemma 2 shows that γ_{1r} would be equal to $\beta_1 \cdot \mathbb{E}[f_H(\tau_{V_i}(r)|X_i, V_i, W_i)]$ for each r.[50] The quantity $\mathbb{E}[f_H(\tau_{V_i}(r)|X_i, V_i, W_i)]$ is unobservable, but note that for any $r \in \{1, 2, \ldots 6\}$ the observable probability $P(R_i = r) = P(R_i \leq r) - P(R_i \leq r-1)$ identifies the quantity

$$\begin{aligned}\mathbb{E}[F_H(\tau_{V_i}(r)|X_i, V_i, W_i) - F_H(\tau_{V_i}(r-1)|X_i, V_i, W_i)] \\ \approx \mathbb{E}\left[\{\tau_{V_i}(r) - \tau_{V_i}(r-1)\} \cdot f_H(\tau_{V_i}(r)|X_i, V_i, W_i)\right]\end{aligned}$$

where the approximation takes the density $f_H(\tau_{V_i}(r)|X_i, V_i, W_i)$ to be roughly constant on the interval $[\tau_{V_i}(r), \tau_{V_i}(r-1)]$. This will be a good approximation if that interval is small with high probability (i.e. in the limit of many categories), in which case $P(R_i = r)$

[50] This uses that since γ_{1r} does not depend on x or w, we must have that $\mathbb{E}[f_H(\tau_{V_i}(r)|x, V_i, w)|W_i = w] = \mathbb{E}[f_H(\tau_{V_i}(r)|x_i, V_i, W_i)]$ for all x and w.

is roughly proportional to $\mathbb{E}[f_H(\tau_{V_i}(r)|X_i, V_i, W_i)]$, if $\mathbb{E}[\tau_{V_i}(r) - \tau_{V_i}(r-1)]$ does not vary much with r.

G Proofs

G.1 Proof of Lemma 1

Fix any v. First we show that if (8) holds for all r then Assumption MONO holds. Indeed, suppose that for some $h' > h$ we had $r(h', v) < r(h, v)$. Substituting $r = r(h', v)$ into (8), we would then have that $r(h, v) > r(h', v) \implies h > \tau_v(r(h', v))$ and hence that $h' > \tau_v(r(h', v))$ since $h' > h$. But $h' > \tau_v(r(h', v))$ violates the definition of τ_v, since then $h' > \sup\{h \in \mathcal{H} : r(h, v) \leq r(h', v)\} \geq h'$.

Left-continuity of r holds by considering any increasing sequence of h converging to $\tau_v(r)$, i.e. I show that $\lim_{h \uparrow \tau_v(r)} r(h, v) = r(\tau_v(r), v)$. First, note that $\lim_{h \uparrow \tau_v(r)} r(h, v) > r(\tau_v(r), v)$ would violate weak monotonicity of r. Suppose instead that $\lim_{h \uparrow \tau_v(r)} r(h, v) = r^*$ where $r^* < r(\tau_v(r), v)$. This limit exists by the increasing property of r. It must then be the case that $\tau_v(r^*) = \tau_v(r)$. To see this, consider the two alternatives. For $\tau_v(r^*) < \tau_v(r)$, there would need to exist an h^* such that $r(h^*, v) > r^*$ but $h^* < \tau_v(r)$. This would violate $\lim_{h \uparrow \tau_v(r)} r(h, v) = r^*$ given that r is increasing. Suppose instead that $\tau_v(r^*) > \tau_v(r)$. Then there would need to exist an h^* such that $r(h^*, v) > r$ but $h^* < \tau_v(r)$. But $h^* < \tau_v(r)$ implies that $r(h^*, v) \leq r$ given that r is increasing. Now, given that $\tau_v(r^*) = \tau_v(r)$, $r^* < r(\tau_v(r), v)$ would violate (8) for $h = \tau_v(r^*)$, because $r(h, v) > r \implies h > \tau_v(r)$.

Now we will show that if Assumption MONO holds then (8) is satisfied for all v, r. First, note that $\tau_v(r)$ is weakly increasing in r, and thus $r(h, v) \leq r \implies \tau_v(r(h, v)) \leq \tau_v(r) \implies h \leq \tau_v(r)$ since by the definition of $\tau_v(r)$: $h \leq \tau_v(r(h, v))$ for any h. Thus we can establish the $\implies$ direction of (8), without even invoking Assumption MONO. In the other direction, assume that for some r and h, $h \leq \tau_v(r)$ but $r(h, v) > r$. By the increasing property of MONO: $h \leq \tau_v(r) \implies r(h, v) \leq r(\tau_v(r), v)$. Thus $r < r(h, v) \leq r(\tau_v(r), v)$ and thus $r(\tau_v(r), v) > r$, so $r(\cdot, v)$ must have a left discontinuity at $\tau_v(r)$.

G.2 Proof of Lemma 2

We begin by showing that under MONO, REG_j, and the first part of EXOG ($\{X_i \perp\!\!\!\perp V_i\} \mid W_i$):

$$\begin{aligned}
&\partial_{x_j} P(R_i \leq r | x, w) \\
&\quad = -\mathbb{E}\left\{ f_H(\tau_{V_i}(r)|x, V_i, w) \cdot \partial_{x_j} Q_{H|XVW}(\alpha|x, V_i, w)\big|_{\alpha = F_{H|XVW}(\tau_{V_i}(r)|x, V_i, w)} \middle| W_i = w \right\}
\end{aligned}$$

This result is of independent interest, because it shows that a regression of the distribution of R on a component of X can be decomposed into a linear combination of quantile regressions of H on X (conditional on V and W). Beyond regularity conditions, this

result only requires reporting heterogeneity V to be conditionally independent of variation in X_j, and no independence assumptions regarding potential outcomes. To interpret this result causally, we then add the second part of EXOG.

To see this, note that by the law of iterated expectation and Lemma 1:

$$\begin{aligned}
P(R_i \leq r | X_i = x, W_i = w) &= \int dF_{UV|XW}(u, v|x, w) \cdot \mathbb{1}(r(h(x, u), v) \leq r) \\
&= \int dF_{UV|XW}(u, v|x, w) \cdot \mathbb{1}(h(x, u) \leq \tau_v(r)) \\
&= \int dF_{V|XW}(v|x, w) \int dF_{U|XVW}(u|x, v, w) \cdot \mathbb{1}(h(x, u) \leq \tau_v(r)) \\
&= \int dF_{V|W}(v|w) \cdot \mathbb{E}\left[\mathbb{1}(h(x, U_i) \leq \tau_v(r)) | \, X_i = x, V_i = v, W_i = w\right] \\
&= \int dF_{V|W}(v|w) \cdot P(H_i \leq \tau_v(r) | X_i = x, V_i = v, W_i = w)
\end{aligned}$$

where I have used $\{X_i \perp\!\!\!\perp V_i\} | W_i$ in the second to last step to replace $F_{V|XW}$ with $F_{V|W}$.

By differentiating the equation $Q_{H|XVW}(F_{H|XVW}(h|x, v)|x, v, w) = h$ with respect to x_j, we have:

$$\partial_{x_j} P(H_i \leq h | X_i = x, V_i = v, W_i = w) = -f_H(h|x, v, w) \cdot \partial_{x_j} Q_{H|XVW}(\alpha|x, v, w)\big|_{\alpha = F_{H|XVW}(h|x,v,w)}$$

By dominated convergence (using Assumption REG) we can move the derivative inside the expectation, and thus:

$$\partial_{x_j} P(R_i \leq r|x) = -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v, w) \cdot \partial_{x_j} \, Q_{H|XVW}(\alpha|x, v, w)\big|_{\alpha = F_{H|XVW}(\tau_v(r)|x,v,w)}.$$

Now we bring in the second part of EXOG, which implies that $\{X_{ji} \perp\!\!\!\perp U_i\} | (X_{-j,i}, V_i, W_i)$, where $X_{-j,i}$ denotes all of the components of X_i aside from the j^{th}. The theorem of Hoderlein and Mammen (2007) implies that given this and REG:

$$\partial_{x_j} \, Q_{H|XVW}(\alpha|x, v, w)\big|_{\alpha = F_{H|XVW}(\tau_v(r)|x,v,w)} = \mathbb{E}\left[\partial_{x_j} h(x, U_i) | H_i = \tau_v(r), x, v, w\right]$$

Therefore:

$$\partial_{x_j} P(R_i \leq r|x) = -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i) | H_i = \tau_v(r), x, v\right]$$

In the case where V is degenerate, a similar proof to the above is used in Chernozhukov et al. (2019) to study derivatives of conditional choice probabilities in multinomial choice models (under somewhat different regularity conditions).

In the proof of Theorem 1 in Hoderlein and Mammen (2007), the conditional expectation function analogous to $\mathbb{E}\left[\partial_{x_j} h(x, U_i) | H_i = h, X_i = x, V_i = v, W_i = w\right]$ appearing in

the expression for $\partial_{x_j} P(R_i \leq r | X_i = x)$ is defined to be the following integral:

$$\int dt \cdot t \cdot \frac{f_{H,\partial_{x_j} h(x,U)|XVW}(h,t|x,v,w)}{f_{H|XVW}(h|x,v,w)} = \int dt \cdot t \cdot \frac{\partial_t \partial_h P(H_i \leq h, \partial_{x_j} h(x,U_i) \leq t | x,v,w)}{\partial_h P(H_i \leq h | x,v,w)}$$
$$\int dt \cdot t \cdot \partial_t \left\{ \frac{\lim_{\epsilon \downarrow 0} P(H_i \in [h, h+\epsilon], \partial_{x_j} h(x,U_i) \leq t | x,v,w)/\epsilon}{\lim_{\epsilon \downarrow 0} P(H_i \in [h,h+\epsilon] | x,v,w)/\epsilon} \right\}$$
$$\int dt \cdot t \cdot \partial_t \left\{ \lim_{\epsilon \downarrow 0} \frac{P(\partial_{x_j} h(x,U_i) \leq t, h(x,U_i) \in [h,h+\epsilon] | x,v,w)}{P(h(x,U_i) \in [h,h+\epsilon] | x,v,w)} \right\} \quad (42)$$

where $f_{H|XVW}$ and $f_{H,\partial_{x_j} h(x,U)|XVW}$ exist and have a ratio that is dominated by an absolutely integrable function $M \cdot c(t)$, by Assumption REG_j. Given that U_i is a random vector in $\mathbb{R}^{d_U}$ with a well-defined probability distribution conditional on $X_i = x, V_i = v, W_i = w$, the limit

$$\lim_{\epsilon \downarrow 0} \frac{P(\partial_{x_j} h(x,U_i) \leq t, h(x,U_i) \in (h,h+\epsilon] | x,v,w)}{P(h(x,U_i) \in (h,h+\epsilon] | x,v,w)}$$

yields a regular conditional probability distribution of $\partial_{x_j} h(x,U_i)$ given $H_i = h(x,U_i) = h$ (and $X_i = x, V_i = v, W_i = w$). See the result of Pfanzagl (1979) for details.

Under an interchange of the limit and the integral in (42), we could also write the quantity $\mathbb{E}[\partial_{x_j} h(x,U_i) | H_i = h, X_i = x, V_i = v, W_i = w]$ as $\lim_{\epsilon \downarrow 0} \mathbb{E}[\partial_{x_j} h(x,U_i) | H_i \in [h,h+\epsilon], X_i = x, V_i = v, W_i = w]$. I employ this limit representation to offer an intuitive description of the Lemma 2 estimand as a weighted average over $\partial_{x_j} h(x,U_i)$, among individuals having H_i "near" $\tau_{V_i}(r)$, in the limit that $\epsilon \downarrow 0$.[51] Sasaki (2015) shows how such outcome-conditioned average derivatives can be written as an explicit integral over the distribution of heterogeneity values $U_i \in \mathbb{R}^{d_U}$ such that $h(x,U_i) = \tau_v(r)$.

[51] One can also establish Lemma 2 intuitively by applying Theorem 2 and letting $x' \to x$ (see Footnote 31).

G.3 Proof of Theorem 1

Let $\beta_r(x,v,w) := \mathbb{E}\left[\partial_{x_j} h(x,U_i)|H_i = \tau_v(r), x, v, w\right]$. Averaging Eq (12) over X_i, W_i yields:

$$\begin{aligned}
&\mathbb{E}[\partial_{x_j} P(R_i \leq r|X_i, W_i)] \\
&= -\int dF_{XW}(x,w) \cdot \int dF_{V|W}(v|w) \cdot f_{H|XVW}(\tau_v(r)|x,v,w) \cdot \beta_r(x,v,w) \\
&= -\int dF_{XW}(x,w) \cdot \int dF_{V|XW}(v|x,w) \cdot f_{H|XVW}(\tau_v(r)|x,v,w) \cdot \beta_r(x,v,w) \\
&= -\int dF_{XVW}(x,v,w) \cdot f_{H|XVW}(\tau_v(r)|x,v,w) \cdot \beta_r(x,v,w) \\
&= -\int dF_{XVW|H}(x,v,w|\tau_v(r)) \cdot f_H(\tau_v(r)) \cdot \beta_r(x,v,w) \\
&= -\int dF_{V|H}(v|\tau_v(r)) \cdot f_H(\tau_v(r)) \int dF_{XW|VH}(x,w|v,\tau_v(r)) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i = \tau_v(r), x, v, w\right] \\
&= -\int dF_{V|H}(v|\tau_v(r)) \cdot f_H(\tau_v(r)) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i = \tau_{V_i}(r), V_i = v\right] \\
&= -\int dF_V(v) \cdot f_{H|V}(\tau_v(r)|v) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i = \tau_{V_i}(r), V_i = v\right]
\end{aligned}$$

using EXOG in the second step. Note that by Bayes' rule:

$$dF_{V|H-\tau_V(r)}(v|0) = f_{H-\tau_V(r)|V}(0|v) \cdot \frac{dF_V(v)}{f_{H-\tau_V(r)}(0)} = f_{H|V}(\tau_v(r)|v) \cdot \frac{dF_V(v)}{f_{H-\tau_V(r)}(0)}$$

and thus $dF_V(v) \cdot f_{H|V}(v|\tau_v(r)) = f_{H-\tau_V(r)}(0) \cdot dF_{V|H-\tau_V(r)}(v|0)$, where $f_{H|V}(\tau_v(r)|v) = f_{H-\tau_V(r)|V}(0|v)$ given that $\tau_V(r)$ is a constant given $V = v$. Note that existence of $f_{H-\tau_V(r)}(0)$ is guaranteed by Assumption REG_j, since by integrating item four of REG_j we know that the density $f_{H|V}$ exists and thus the density $f_{H-\tau_V(r)|V}$ exists as well. Thus:

$$\begin{aligned}
&\mathbb{E}[\partial_{x_j} P(R_i \leq r|X_i, W_i)] \\
&= -\int dF_V(v) \cdot f_{H|V}(\tau_v(r)|v) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i = \tau_{V_i}(r), V_i = v\right] \\
&= -f_{H-\tau_V(r)}(0) \cdot \int dF_{V|H-\tau_V(r)}(v|0) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i - \tau_{V_i}(r) = 0, V_i = v\right] \\
&= -f_{H-\tau_V(r)}(0) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i - \tau_{V_i}(r) = 0\right] = -f_{H-\tau_V(r)}(0) \cdot \mathbb{E}\left[\partial_{x_j} h(X_i,U_i)|H_i = \tau_{V_i}(r)\right]
\end{aligned}$$

G.4 Proof of Corollary 2

The proof will make use of the following lemma:

Lemma. *Let B be a random variable and A a random vector, admitting a joint density and satisfying $A \perp\!\!\!\perp B$. Let g be a scalar-valued differentiable function of A. Then if B has a uniform distribution on an interval of the real line that contains the support of $g(a)$, it follows that $A|\{g(A) = B\} \sim A$.*

Proof. Since g is differentiable, $g(A), B$ admit a joint density by the multivariate change-of-variables theorem. Using independence of A and B, we can write this joint density as $f_{g(A),B}(t,b) = f_{g(A)}(t) \cdot f_B(b)$. B then admits a density conditional on $g(A) = B$, which is given by

$$f_{B|g(A)=B}(b) = \frac{f_{g(A),B}(b,b)}{\int f_{g(A),B}(s,s)ds} = \frac{f_{g(A)}(b) \cdot f_B(b)}{\int f_{g(A)}(s) \cdot f_B(s)ds}$$

Since $B \sim Unif[\ell, u]$ for some ℓ, u we have that $f_B(s) = f_B(s^*)$ if $s \in [\ell, u]$ and zero otherwise, where s^* is an arbitrary point in $[\ell, u]$. We can thus rewrite the above as

$$f_{B|g(A)=B}(b) = \frac{f_B(s^*) \cdot f_{g(A)}(b)}{f_B(s^*) \cdot \int_\ell^u f_{g(A)}(s)ds} \cdot \mathbb{1}(\ell \leq b \leq u) = f_{g(A)}(b) \cdot \mathbb{1}(\ell \leq b \leq u)$$

where $\int_\ell^u f_{g(A)}(s)ds = \int f_{g(A)}(s)ds = 1$ since $supp\{g(A)\} \subseteq [\ell, u]$.

For a (vectored-valued) t, consider the CDF of A conditional on $g(A) = B$, evaluated at t:

$$\begin{aligned}
P(A \leq t|g(A) = B) &= \int f_{B|g(A)=B}(b) \cdot P(A \leq t|g(A) = B, B = b) \cdot db \\
&= \int_\ell^u f_{g(A)}(b) \cdot P(A \leq t|g(A) = b, B = b) \cdot db \\
&= \int_\ell^u f_{g(A)}(b) \cdot P(A \leq t|g(A) = b) \cdot db \\
&= \int f_{g(A)}(b) \cdot P(A \leq t|g(A) = b) \cdot db = P(A \leq t)
\end{aligned}$$

where I've used that $A \perp\!\!\!\perp B$ in the third equality, that $supp\{g(A)\} \subseteq [\ell, u]$ in the fourth, and then finally the law of iterated expectations.

□

To ease notation, let $\tau_{ri} = \tau_{V_i}(r)$ and define τ_i to be a vector of τ_{ri} across $r \in \mathcal{R}$. Given $V_i \perp\!\!\!\perp X_i|W_i$, it follows that $\tau_i \perp\!\!\!\perp X_i|W_i$, and we can rewrite Eq. (12) as a one-dimensional integral over τ_{ri} (by redefining $V_i = \tau_i$ and using the law of total probability):

$$\begin{aligned}
&\partial_{x_j} P(R_i \leq r|x, w) \\
&= -\int dF_{\tau_r|W}(t|w) \cdot f_{h(x,U)}(t|\tau_{ri} = t, x, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|h(x, U_i) = t, \tau_{ri} = t, x, w\right] \\
&= -\int dF_{\tau_r|W}(t|w) \cdot f_{h(x,U)}(t|\tau_{ri} = t, x, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|h(x, U_i) = t, h(x, U_i) = \tau_{ri}, x, w\right]
\end{aligned}$$

Under the assumption that $\tau_{ri}|W_i \sim Unif[\ell_r, u_r]$, we can replace $dF_{\tau_r|W}(t|w)$ with $\frac{dt}{\mu_r - \ell_r} \cdot \mathbb{1}(\ell_r \leq t \leq \mu_r)$. Using $V_i \perp\!\!\!\perp U_i|X_i, W_i$, we have $f_{h(x,U)}(t|\tau_{ri} = t, x, w) = f_{h(x,U)}(t|x, w)$ and by the Lemma above that the distribution of U_i conditional on $h(x, U_i) = \tau_{ri}, X_i = x, W_i = w$ is the same of the distribution of $U_i|X_i = x, W_i = w$. This implies that $\mathbb{E}\left[\partial_{x_j} h(x, U_i)|h(x, U_i) = t, h(x, U_i) = \tau_{ri}, x, w\right] = \mathbb{E}\left[\partial_{x_j} h(x, U_i)|h(x, U_i) = t, x, w\right]$, and

thus

$$\partial_{x_j} P(R_i \leq r|x,w) = -\frac{1}{\mu_r - \ell_r} \cdot \int_{\ell_r}^{u_r} dt \cdot f_{h(x,U)}(t|x,w) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|h(x,U_i)=t,x,w\right]$$

Meanwhile:

$$\begin{aligned}\mathbb{E}\left[\partial_{x_j} h(x,U_i)|x,w\right] &= \int dt \cdot f_{h(x,U)}(t|x,w) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|h(x,U_i)=t,x,w\right] \\ &= \int_{\ell_r}^{u_r} dt \cdot f_{h(x,U)}(t|x,w) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|h(x,U_i)=t,x,w\right]\end{aligned}$$

using that $\text{supp}\{h(x,U_i)\} \subseteq [\mu_r, \ell_r]$ in the second equality. Thus, $\partial_{x_j} P(R_i \leq r|x,w) = -\frac{1}{\mu_r - \ell_r} \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|x,w\right]$.

G.5 Proof of Theorem 2

I begin with a heuristic overview: the detailed proof is below. The logic of the result is as follows: for a given individual having $V_i = v$, R_i will be less than or equal to r when $X_i = x'$, but not when $X_i = x$, if $\Delta_i < 0$ and $h(x,U_i) \in (\tau_v(r), \tau_v(r) + |\Delta_i|]$. This event increases the value of $P(R_i \leq r|x,w') - P(R_i \leq r|x,w)$. On the other hand, R_i will be less than or equal to r when $X_i = x$ but not when $X_i = x'$ when $\Delta_i > 0$ and $h(x,U_i) \in (\tau_v(r) - \Delta_i, \tau_v(r)]$. This event instead decreases the value of $P(R_i \leq r|x',w) - P(R_i \leq r|x,w)$. The RHS of Theorem 2 can be written as

$$\mathbb{E}\left\{ \int_{\tau_{V_i}(r)-\Delta_i}^{\tau_{V_i}(r)} dy \cdot f_H(y|\Delta_i, X_i = x, V_i) \middle| W_i = w \right\},$$

which averages over both positive and negative Δ_i, covering both cases.

Now let us prove the result of Theorem 2. By the law of iterated expectations, Lemma 1, and then EXOG

$$\begin{aligned}&P(R_i \leq r|X_i = x', W_i = w) - P(R_i \leq r|X_i = x, W_i = w) \\ &= \int dF_{UV|XW}(u,v|x',w) \cdot \mathbb{1}(r(h(x',u),v) \leq r) - \int dF_{UV|XW}(u,v|x,w) \cdot \mathbb{1}(r(h(x,u),v) \leq r) \\ &= \int dF_{UV|XW}(u,v|x',w) \cdot \mathbb{1}(h(x',u) \leq \tau_v(r)) - \int dF_{UV|XW}(u,v|x,w) \cdot \mathbb{1}(h(x,u) \leq \tau_v(r)) \\ &= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \leq \tau_v(r)|X_i = x', v, w) - P(h(x,U_i) \leq \tau_v(r)|X_i = x, v, w)\} \\ &= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \leq \tau_v(r)|X_i = x, v, w) - P(h(x,U_i) \leq \tau_v(r)|X_i = x, v, w)\}\end{aligned}$$

using that $X_i \perp\!\!\!\perp U_i | W_i, V_i$ by EXOG in the last step. Thus:

$$
\begin{aligned}
&P(R_i \leq r | X_i = x', W_i = w) - P(R_i \leq r | X_i = x, W_i = w) \\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x', U_i) \leq \tau_v(r) \text{ but not } h(x, U_i) \leq \tau_v(r) | x, v, w) \\
&\qquad\qquad - P(h(x, U_i) \leq \tau_v(r) \text{ but not } h(x', U_i) \leq \tau_v(r) | x, v, w)\} \\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x', U_i) \leq \tau_v(r) < h(x, U_i) | x, v, w) - P(h(x, U_i) \leq \tau_v(r) < h(x', U_i) | x, v, w)\} \\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x, U_i) \in (\tau_v(r), \tau_v(r) - \Delta_i] | x, v, w) - P(h(x, U_i) \in (\tau_v(r) - \Delta_i, \tau_v(r)] | x, v, w)\} \\
&= \int dF_{V|W}(v|w) \cdot \{P(H_i \in (\tau_v(r), \tau_v(r) - \Delta_i] | x, v, w) - P(H_i \in (\tau_v(r) - \Delta_i, \tau_v(r)] | x, v, w)\} \\
&= -\int dF_{V|W}(v|w) \cdot \int dF_{\Delta|XVW}(\Delta | x, v, w) \cdot \{P(H_i \in (\tau_v(r), \tau_v(r) - \Delta] | \Delta, x, v, w) \\
&\qquad\qquad - P(H_i \in (\tau_v(r) - \Delta, \tau_v(r)] | \Delta, x, v, w)\} \\
&= -\int dF_{V|W}(v|w) \cdot \int dF_{\Delta|XVW}(\Delta | x, v, w) \int_{\tau_v(r) - \Delta}^{\tau_v(r)} dy \cdot f_H(h | \Delta, x, v, w) \\
&= -\int dF_{V|W}(v|w) \cdot \int dF_{\Delta|VW}(\Delta | x, v, w) \cdot \bar{f}_H(\tau_v(r) | \Delta, x, v, w) \cdot \Delta \\
&= -\int dF_{V|W}(v|w) \cdot \mathbb{E}[\bar{f}_H(\tau_v(r) | \Delta_i, x, v, w) \cdot \Delta_i | V_i = v, W_i = w] \\
&= -\mathbb{E}[\bar{f}_H(\tau_{V_i}(r) | \Delta_i, x, V_i, w) \cdot \Delta_i | W_i = w]
\end{aligned}
$$

using EXOG and with the definition $\bar{f}_H(y|\Delta, x, v, w) := \frac{1}{\Delta} \int_{y-\Delta}^{y} f_H(h|\Delta, x, v, w) dh$.

G.6 Proof of Proposition 1

To fix the scale normalization, suppose that $g(x^*) = 1$ for some $x^* \in \mathcal{X}$. Then, note that by the fundamental theorem of calculus, we may write

$$\log g(x) = \int_{x^*}^{x} \nabla \log g(x) \circ dv = \sum_{j=1}^{J} \int_{x_j^*}^{x_j} \partial_{x_j} \log g(x_1, \ldots x_{j-1}, t, 0, \ldots, 0) dt$$

where $\circ$ denotes a dot product and dv traces any continuous path in $\mathcal{X}$ from x^* to x, for example the one given after the second equality that integrates over each x_j in turn.

If all components of X are continuous and there are no controls, then note that for any $x \in \mathcal{X}$ we can identify $\partial_{x_j} g(x) / \partial_{x_k} g(x) = \partial_{x_j} \mathbb{E}[R_i|x] / \partial_{x_k} \mathbb{E}[R_i|x]$ for any $j, k \in 1 \ldots J$ by Eq. (17). By assumption that $g(x)$ is homogeneous of degree one, we have that $g(\lambda x) = \lambda g(x)$. "Euler's theorem" of homogeneous functions then implies that $g(x) = \sum_{j=1}^{J} \partial_{x_j} g(x) \cdot x_j$ (this result can be obtained by differentiating $g(\lambda x) = \lambda g(x)$ with respect to λ and evaluating at $\lambda = 1$). Thus $(\partial_{x_k} \log g(x))^{-1} = \frac{g(x)}{\partial_{x_k} g(x)} = 1 + \sum_{j \neq k} \frac{\partial_{x_j} g(x)}{\partial_{x_k} g(x)} \cdot x_j$. We

now arrive at a constructive expression for $g(x)$ in terms of observables

$$g(x) = e^{\int_{x_j^*}^{x_j} \left(1+\sum_{j\neq k} \frac{\partial_{x_j}\mathbb{E}[R_i|(x_1,\ldots x_{j-1},t,0,\ldots,0)]}{\partial_{x_k}\mathbb{E}[R_i|(x_1,\ldots x_{j-1},t,0,\ldots,0)]}\cdot x_j\right)^{-1} dt} \tag{43}$$

G.7 Proof of Proposition 3

For any x and x' that differ in component X_j only:

$$\begin{aligned}
&P(R_i \leq r|X_i = x', A_i) - P(R_i \leq r|X_i = x, A_i) \\
&= \mathbb{E}[\mathbb{1}(H_i \leq \tau_{V_i}(r))|x', A_i] - \mathbb{E}[\mathbb{1}(H_i \leq \tau_{V_i}(r))|x, A_i] \\
&= \int dF_{V|XA}(v|x', A_i) \cdot \mathbb{E}[\mathbb{1}(H_i \leq \tau_v(r))|x', v, A_i] - \int dF_{V|XA}(v|x, A_i) \cdot \mathbb{E}[\mathbb{1}(H_i \leq \tau_v(r))|x, v, A_i] \\
&= \int dF_{V|XA}(v|x, A_i) \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r))|x', v, A_i] - \mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_v(r))|x, v, A_i]\} \\
&= \int dF_{V|XA}(v|x, A_i) \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r)) - \mathbb{1}(h(x, U_i) \leq \tau_v(r))|x, v, A_i]\} \\
&= \int dF_{V|XA}(v|x, A_i) \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r) < h(x, U_i))|x, v, A_i]\} \\
&\qquad - \int dF_{V|XA}(v|x, A_i) \cdot \{\mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_v(r) < h(x', U_i))|x, v, A_i]\}
\end{aligned} \tag{44}$$

where in the second equality I have used that $\{X_{ji} \perp\!\!\!\perp V_i\}|(A_i, W_i)$ so that $F_{V|XA}(v|x', a) = F_{V|XA}(v|x, a)$ for all a, and in the fourth equality that $\{X_{ji} \perp\!\!\!\perp U_i\}|(A_i, W_i, V_i)$ so that

$$\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r))|X_i = x', V_i = v, A_i] = \mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r))|X_i = x, V_i = v, A_i]$$

Given (44), we have that

$$\begin{aligned}
&\mathbb{E}[A_i \cdot \{P(R_i \leq r|x', A_i) - P(R_i \leq r|x, A_i)\}|X_i = x] \\
&\qquad = \int dF_{A|X}(a|x) \cdot a \cdot \int dF_{V|XA}(v|x, a) \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r) < h(x, U_i)|x, v, a]\} \\
&\qquad\qquad - \int dF_{A|X}(a|x) \cdot a \cdot \int dF_{V|XA}(v|x, a) \cdot \{\mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_v(r) < h(x', U_i)|x, v, a]\} \\
&= \mathbb{E}[A_i \cdot \mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i))|X_i = x] - \mathbb{E}[A_i \cdot \mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i))|X_i = x]
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathbb{E}[P(R_i \leq r|x', A_i) - P(R_i \leq r|x, A_i)|X_i = x] \\
&\quad = \mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i))|X_i = x] - \mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i))|X_i = x]
\end{aligned}$$

Note that assuming the numerator and denominator below both exist, we can write

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]} \\
&\quad = \frac{\lim_{x' \downarrow x} \frac{1}{||x'-x||} \cdot \mathbb{E}[A_i \cdot \{P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i)\} | X_i = x]}{\lim_{x' \downarrow x} \frac{1}{||x'-x||} \cdot \mathbb{E}[P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i) | X_i = x]} \\
&\quad = \lim_{x' \downarrow x} \frac{\cancel{\frac{1}{||x'-x||}} \cdot \mathbb{E}[A_i \cdot \{P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i)\} | X_i = x]}{\cancel{\frac{1}{||x'-x||}} \cdot \mathbb{E}[P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i) | X_i = x]} \\
&\quad = \lim_{x' \downarrow x} \frac{\mathbb{E}[A_i \cdot \{P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i)\} | X_i = x]}{\mathbb{E}[P(R_i \leq r | x', A_i) - P(R_i \leq r | x, A_i) | X_i = x]}
\end{aligned}$$

where x' is a sequence of vectors that differ from x only in the j^{th} component, and I've assumed dominated convergence so that we can interchange the limits and expectations.

Thus, by the above:

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]} \\
&= \lim_{x' \downarrow x} \quad \frac{\mathbb{E}[A_i \cdot \mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) | X_i = x]}{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) | X_i = x] - \mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i)) | X_i = x]} \\
&- \lim_{x' \downarrow x} \quad \frac{\mathbb{E}[A_i \cdot \mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i)) | X_i = x]}{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) | X_i = x] - \mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i)) | X_i = x]}
\end{aligned} \tag{45}$$

Suppose for example that $\partial_{x_j} h(x, U_i) \geq 0$ with probability one. Then:

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]} \\
&= \lim_{x' \downarrow x} \quad \frac{\mathbb{E}[A_i \cdot \mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) | X_i = x]}{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) | X_i = x]} \\
&= \lim_{x' \downarrow x} \quad \frac{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x) \cdot \mathbb{E}[A_i | h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i), X_i = x]}{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x)} \\
&= \lim_{x' \downarrow x} \quad \mathbb{E}[A_i | h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i), X_i = x] = \mathbb{E}[A_i | h(x, U_i) = \tau_{V_i}(r), X_i = x] \\
&= \mathbb{E}[A_i | H_i = \tau_{V_i}(r), X_i = x]
\end{aligned}$$

provided that the RHS of the last line is well-defined. Similarly, if $\partial_{x_j} h(x, U_i) \leq 0$ with probability one, then the LHS above evaluates to $\lim_{x' \downarrow x} \quad \mathbb{E}[A_i | h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i), X_i = x] = \mathbb{E}[A_i | H_i = \tau_{V_i}(r), X_i = x]$ and we thus obtain the same expression.

More generally, if the sign of treatment effects vary by unit:

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]} \\
&= \lim_{x' \downarrow x} \frac{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x) \cdot \mathbb{E}[A_i | h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i), X_i = x]}{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x) - P(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i) | X_i = x)} \\
&\quad - \lim_{x' \downarrow x} \frac{P(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i) | X_i = x) \cdot \mathbb{E}[A_i | h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i), X_i = x]}{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x) - P(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i) | X_i = x)}
\end{aligned} \tag{46}$$

and the estimand $\frac{\mathbb{E}[A_i \cdot \partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}{\mathbb{E}[\partial_{x_j} P(R_i \leq r | X_i = x, A_i) | X_i = x]}$ yields a non-convex combination of $\mathbb{E}[A_i | \mathbb{1}(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i), X_i = x]$ and $\mathbb{E}[A_i | \mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i), X_i = x]$.

A sufficient condition for $\partial_{x_j} h(x, U_i)$ to have the same sign for all i "uniformly" in the above sense is that for all x' within some neighborhood of x, $h(x', U_i)$ is either strictly increasing or strictly decreasing in component j of x', for all U_i. Specifically, let $x'(\delta)$ be the vector x but with δ added to the j^{th} component. Then, for some $\bar{\delta} > 0$, we have we have that $P(h(x, U_i) \geq h(x'(\delta), U_i)) = 1$ or $P(h(x, U_i) \geq h(x'(\delta), U_i)) = 0$ for any $\delta \leq \bar{\delta}$ (i.e. x' and x are sufficiently close). Then given that $\mathbb{E}[A_i | H_i = \tau_{V_i}(r), X_i = x]$ is well-defined we have either that $\lim_{x' \downarrow x} \frac{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x)}{P(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i) | X_i = x)} = 0$ if $\partial_{x_j} h(x, U_i) \geq 0$ with probability one, or that $\lim_{x' \downarrow x} \frac{P(h(x, U_i) \leq \tau_{V_i}(r) < h(x', U_i) | X_i = x)}{P(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i) | X_i = x)} = 0$ if $\partial_{x_j} h(x, U_i) \leq 0$ with probability one. In either case one term of (46) evaluates to zero and the other to $\mathbb{E}[A_i | H_i = \tau_{V_i}(r), X_i = x]$.

To see that (32) holds under the stronger condition that $\{X_{ji} \perp\!\!\!\perp (A_i, U_i, V_i)\} | W_i$, we have in this case by similar steps as above:

$$\begin{aligned}
&\mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r) | X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r) | X_i = x] \\
&= \mathbb{E}[A_i \cdot \mathbb{1}(H_i \leq \tau_{V_i}(r)) | x'] - \mathbb{E}[A_i \cdot \mathbb{1}(H_i \leq \tau_{V_i}(r)) | x] \\
&= \int dF_{VA|X}(v, a | x') \cdot a \cdot \mathbb{E}[\mathbb{1}(H_i \leq \tau_v(r)) | x'] - \int dF_{VA|X}(v, a | x) \cdot a \cdot \mathbb{E}[\mathbb{1}(H_i \leq \tau_v(r)) | x] \\
&= \int dF_{VA|X}(v, a | x) \cdot a \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r)) | x'] - \mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_v(r)) | x]\} \\
&= \int dF_{VA|X}(v, a | x) \cdot a \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r)) - \mathbb{1}(h(x, U_i) \leq \tau_v(r)) | x]\} \\
&= \int dF_{VA|X}(v, a | x) \cdot a \cdot \{\mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_v(r) < h(x, U_i)) | x]\} \\
&\qquad - \int dF_{VA|X}(v, a | x) \cdot a \cdot \{\mathbb{E}[\mathbb{1}(h(x, U_i) \leq \tau_v(r) < h(x', U_i)) | x]\} \\
&= \mathbb{E}[A_i \cdot \{\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) - \mathbb{1}(h(x, U_i) \leq \tau_v(r) < h(x', U_i))\} | X_i = x]
\end{aligned} \tag{47}$$

using that $\{X_{ji} \perp\!\!\!\perp (V_i, A_i)\} | W_i$ so that $F_{VA|X}(v, a | x') = F_{VA|X}(v, a | x)$ for all a in the

third equality. Similarly:

$$\begin{aligned}&P(R_i \leq r)|X_i = x') - P(R_i \leq r)|X_i = x) \\ &= \mathbb{E}[\mathbb{1}(h(x', U_i) \leq \tau_{V_i}(r) < h(x, U_i)) - \mathbb{1}(h(x, U_i) \leq \tau_v(r) < h(x', U_i))|X_i = x] \quad (48)\end{aligned}$$

And thus

$$\frac{\partial_{x_j}\mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r)|X_i = x]}{\partial_{x_j} P(R_i \leq r|X_i = x)} = \lim_{x' \downarrow x} \frac{\mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r)|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \leq r)|X_i = x]}{P(R_i \leq r)|X_i = x') - P(R_i \leq r)|X_i = x)}$$

under suitable regularity conditions to take the derivative outside of the expectation. Given (47) and (48), the above yields the same estimand as (45), again simplifying to $\mathbb{E}[A_i|h(x, U_i) = \tau_{V_i}(r), X_i = x]$ given the common sign of derivative $\partial_{x_j} h(x, U_i)$ across all individuals i.

G.8 Proof of Proposition 4

By the law of iterated expectations: $\mathbb{E}[R_i|X_i = x, W_i = w] = \int dF_{V|W}(v|w) \cdot \int dh \cdot r(h, v) \cdot f_H(h|x, v, w)$. Now use REG to move the derivative inside the integral:

$$\partial_{x_j}\mathbb{E}[R_i|X_i = x, W_i = w] = \int dF_{V|W}(v|w) \cdot \int dh \cdot r(h, v) \cdot \partial_{x_j} f_H(h|x, v, w)$$

Theorem 1 of Kasy (2022) (for a one-dimensional outcome) implies that $\partial_{x_j} f_H(h|x, v, w) = -\frac{\partial}{\partial h}\left\{f_H(h|x, v, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = h, x, v, w\right]\right\}$. Thus

$$\partial_{x_j}\mathbb{E}[R_i|x, w] = -\int dF_{V|W}(v|w) \int dh \cdot r(h, v) \cdot \frac{\partial}{\partial h}\left\{f_H(h|x, v, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = h, x, v, w\right]\right\}$$

Now use integration by parts, applying the assumed boundary condition eliminates the first term, establishing the result:

$$\partial_{x_j}\mathbb{E}[R_i|x, w] = 0 + \int dF_{V|W}(v|w) \int dh \cdot r'(h, v) \cdot f_h(h|x, v, w) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = h, x, v, w\right]$$

G.9 Proof of Proposition 5

With the substitution $h = \tau_v(r)$, $dr = r'(h, v) \cdot dh$:

$$\begin{aligned}&\sum_r \int_{\tau_v(r)-\Delta}^{\tau_v(r)} dy \cdot f_H(y|\Delta, x, v, w) \xrightarrow{R} \bar{R} \cdot \int dr \int_{\tau_v(r)-\Delta}^{\tau_v(r)} dy \cdot f_H(y|\Delta, x, v, w) \\ &= \bar{R} \cdot \int dh \cdot r'(h, v) \int_{h-\Delta}^{h} dy \cdot f_H(y|\Delta, x, v, w) = \bar{R} \cdot \int dy \int_{y}^{y+\Delta} dh \cdot r'(h, v) \cdot f_H(y|\Delta, x, v, w) \\ &= \bar{R} \cdot \int dy \cdot \Delta \cdot \bar{r}'(y, \Delta, v) \cdot f_H(y|\Delta, x, v, w) = \Delta \cdot \bar{R} \cdot \mathbb{E}[\bar{r}'(H_i, \Delta, v)|\Delta_i = \Delta, X_i = x, V_i = v, W_i = w]\end{aligned}$$

where $\bar{r}'(y,\Delta,v) := \frac{1}{\Delta}\int_y^{y+\Delta} r'(h,v)dh$. Thus:

$$\begin{aligned}
&\mathbb{E}[R_i|X_i = x', W_i = w] - \mathbb{E}[R_i|X_i = x, W_i = w] \\
&\quad = \bar{R} \cdot \int dF_{V|W}(v|w) \cdot \int dF_{\Delta|XVW}(\Delta|x,v,w) \cdot \Delta \cdot \mathbb{E}[\bar{r}'(H_i,\Delta,v)|\Delta_i = \Delta, X_i = x, V_i = v, W_i = w] \\
&\quad = \bar{R} \cdot \int dF_{V|W}(v|w) \cdot \int dF_{\Delta|XVW}(\Delta|x,v,w) \cdot \mathbb{E}[\Delta \cdot \bar{r}'(H_i,\Delta,v)|\Delta_i = \Delta, X_i = x, V_i = v, W_i = w] \\
&\quad = \bar{R} \cdot \int dF_{V|W}(v|w) \cdot \mathbb{E}\left[\mathbb{E}[\Delta_i \cdot \bar{r}'(H_i,\Delta_i,V_i)|\Delta_i = \Delta, X_i = x, V_i = v]\middle| X_i = x, V_i = v\right] \\
&\quad = \bar{R} \cdot \int dF_{V|W}(v|w) \cdot \mathbb{E}[\Delta_i \cdot \bar{r}'(H_i,\Delta_i,V_i)|X_i = x, V_i = v, W_i = w] \\
&\quad = \bar{R} \cdot \mathbb{E}[\Delta_i \cdot \bar{r}'(H_i,\Delta_i,V_i)|X_i = x, W_i = w]
\end{aligned}$$

Note that if we assume that Δ_i and $\bar{r}'(H_i,\Delta_i,V_i)$ are uncorrelated conditional on $X_i = x, W_i = w$, this reduces to

$$\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x] = \bar{R} \cdot \mathbb{E}[\Delta_i|X_i = x] \cdot \mathbb{E}[\bar{r}'(H_i,\Delta_i,V_i)|X_i = x]$$

G.10 Proof of Proposition 6

Starting with Proposition 5, observe that $\bar{r}'(y,\Delta,v) := \frac{1}{\Delta}\int_y^{y+\Delta} r'(h,v)dh$ is equal to

$$r'(v) \cdot \begin{cases} \frac{y-(\ell(v)-\Delta)}{|\Delta|} \cdot \mathbb{1}(y \in [\ell(v)-\Delta, \ell(v)]) + \mathbb{1}(y \in [\ell(v), \mu(v)-\Delta]) & \\ \qquad + \frac{\mu(v)-y}{\Delta} \cdot \mathbb{1}(y \in [\mu(v)-\Delta, \mu(v)]) & \text{if } \Delta > 0 \\ \frac{y-\ell(v)}{\Delta} \cdot \mathbb{1}(y \in [\ell(v), \ell(v)+|\Delta|]) + \mathbb{1}(y \in [\ell(v)+|\Delta|, \mu(v)]) & \\ \qquad + \frac{\mu(v)+|\Delta|-y}{|\Delta|} \cdot \mathbb{1}(y \in [\mu(v), \mu(v)+|\Delta|]) & \text{if } \Delta < 0 \end{cases}$$

where $r'(v) = \frac{|\mathcal{R}|}{\ell(v)-\mu(v)}$. To ease notation, let us for the moment make the conditioning implicit and let $f(y)$ denote $f_H(y|\Delta,x,v,w)$ and $F(y)$ the corresponding conditional CDF. Let us keep v also implicit in both ℓ and μ. If we let ϕ denote the quantity $\frac{1}{r'(v)}\int dy \cdot \bar{r}'(y,\Delta,v)$ for a fixed Δ, then:

$$\phi = \begin{cases} [F(\ell)-F(\ell-\Delta)]\mathbb{E}\left[\frac{H_i-(\ell-\Delta)}{\Delta}\middle| H_i \in [\ell-\Delta, \ell]\right] + F(\mu-\Delta) & \\ \qquad -F(\ell) + [F(\mu)-F(\mu-\Delta)]\mathbb{E}\left[\frac{\mu-H_i}{\Delta}\middle| H_i \in [\mu-\Delta, \mu]\right] & \text{if } \Delta > 0 \\ [F(\ell+|\Delta|)-F(\ell)]\mathbb{E}\left[\frac{H_i-\ell}{|\Delta|}\middle| H_i \in [\ell, \ell+|\Delta|]\right] + F(\mu) & \\ \qquad -F(\ell+|\Delta|) + [F(\mu+|\Delta|)-F(\mu)]\mathbb{E}\left[\frac{\mu+\Delta-H_i}{|\Delta|}\middle| H_i \in [\mu, \mu+|\Delta|]\right] & \text{if } \Delta < 0 \end{cases} \tag{49}$$

To get a lower bound on ϕ, we use the assumption that $f(y)$ is increasing on the interval $[\ell - |\Delta|, \ell + |\Delta|]$, as well as decreasing on the interval $[\mu - |\Delta|, \mu + |\Delta|]$:

$$\begin{aligned}\phi &\geq \begin{cases} \frac{1}{2}[F(\ell) - F(\ell - \Delta)] + F(\mu - \Delta) - F(\ell) + \frac{1}{2}[F(\mu) - F(\mu - \Delta)] & \text{if } \Delta > 0 \\ \frac{1}{2}[F(\ell + |\Delta|) - F(\ell)] + F(\mu) - F(\ell + |\Delta|) + \frac{1}{2}[F(\mu + |\Delta|) - F(\mu)] & \text{if } \Delta < 0 \end{cases} \\ &= \begin{cases} \frac{1}{2}[F(\mu - \Delta) - F(\ell - \Delta)] + \frac{1}{2}[F(\mu) - F(\ell)] & \text{if } \Delta > 0 \\ \frac{1}{2}[F(\mu + |\Delta|) - F(\ell + |\Delta|)] + \frac{1}{2}[F(\mu) - F(\ell)] & \text{if } \Delta < 0 \end{cases} \\ &= \frac{1}{2}[F(\mu - \Delta) - F(\ell - \Delta)] + \frac{1}{2}[F(\mu) - F(\ell)] \\ &= \frac{1}{2}[F(\mu(v)|\Delta, x', v) - F(\ell(v)|\Delta, x', v)] + \frac{1}{2}[F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x, v, w)],\end{aligned}$$

reintroducing conditioning values with the notation $F(\cdot|\Delta, x, v, w) := F_{H|\Delta XVW}(\cdot|\Delta, x, v, w)$. A lower bound on the weight $\Pi_{x,x'}$ on causal effects in $\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w]$ can thus given by averaging over V_i (c.f. Proposition 5):

$$\begin{aligned}\Pi_{x,x'} \geq \int dF_{V|W}(v|w) \cdot \int dF(\Delta|x, v, w) \cdot \Big\{ &\frac{1}{2}[F(\mu(v)|\Delta, x', v, w) - F(\ell(v)|\Delta, x', v, w)] \\ &+ \frac{1}{2}[F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x, v, w)] \Big\}\end{aligned}$$

Note that this exactly the same as the average between the weights Π_x and $\Pi_{x'}$ corresponding to using continuous variation at $X_i = x$ and $X_i = x'$, respectively. For example (c.f. Eq. 36):

$$\Pi_x = \int dF_{V|W}(v|w) \cdot \int dF(\Delta|x, v, w) \cdot [F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x, v, w)]$$

This leads to the lower bound of $\Pi_{x,x'}/(\frac{1}{2}\Pi_x + \frac{1}{2}\Pi_{x'}) \geq 1$ in Proposition 5.

Now, to obtain an upper bound, notice that an upper bound on ϕ occurs if we imagine putting all of the mass in each of the interval conditional expectations in (49) to the right in the intervals that depend on ℓ, and at the left end for the intervals that depend on μ. Then:

$$\begin{aligned}\phi &\leq \begin{cases} \cancel{F(\ell)} - F(\ell - \Delta) + \cancel{F(\mu - \Delta)} - \cancel{F(\ell)} + F(\mu) - \cancel{F(\mu - \Delta)} & \text{if } \Delta > 0 \\ \cancel{F(\ell + |\Delta|)} - F(\ell) + \cancel{F(\mu)} - \cancel{F(\ell + |\Delta|)} + F(\mu + |\Delta|) - \cancel{F(\mu)} & \text{if } \Delta < 0 \end{cases} \\ &= \begin{cases} F(\mu) - F(\ell - \Delta) & \text{if } \Delta > 0 \\ = F(\mu + |\Delta|) - F(\ell) & \text{if } \Delta < 0 \end{cases} = \begin{cases} F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x', v, w) & \text{if } \Delta > 0 \\ F(\mu(v)|\Delta, x', v, w) - F(\ell(v)|\Delta, x, v, w) & \text{if } \Delta < 0 \end{cases}\end{aligned}$$

where I've used that $F(y|\Delta, x', v, w) = F(y - \Delta|\Delta, x, v, w)$ in the last step. An upper bound for ϕ that applies to both cases can be obtained by adding them together:

$$\phi \leq F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x, v, w) + F(\mu(v)|\Delta, x', v, w) - F(\ell(v)|\Delta, x', v, w) \quad (50)$$

where I've used that $F(\mu) \geq F(\ell - \Delta)$ and $F(\mu + |\Delta|) \geq F(\ell)$ are implied by the assumption that $f(y)$ is increasing on the interval $[\ell - |\Delta|, \ell + |\Delta|]$, while decreasing on the interval $[\mu - |\Delta|, \mu + |\Delta|]$, which implies that $\mu - |\Delta| \geq \ell + |\Delta|$.

Thus, an upper bound on the weight $\Pi_{x,x'}$ on causal effects in $\mathbb{E}[R_i|x', w] - \mathbb{E}[R_i|x, w]$ is:

$$\Pi_{x,x'} \geq \int dF_{V|W}(v|w) \cdot \int dF_{\Delta|XVW}(\Delta|x, v, w) \cdot \{F(\mu(v)|\Delta, x', v, w) - F(\ell(v)|\Delta, x', v, w) \\ + F(\mu(v)|\Delta, x, v, w) - F(\ell(v)|\Delta, x, v, w)\}$$

leading to the upper bound of $\Pi_{x,x'}/(\frac{1}{2}\Pi_x + \frac{1}{2}\Pi_{x'}) \leq 2$ in Proposition 5.

Now consider the final condition in Proposition 5. That $\Pi_{x,x'}/\Pi_x \geq 1/2$ follows from the above since $F(\mu(v)|\Delta, x', v, w) - F(\ell(v)|\Delta, x', v, w) \geq 0$ for all Δ, x, v, w. For the upper bound we have

$$\begin{aligned}
\frac{\Pi_x}{\Pi_{x,x'}} &\geq \frac{\mathbb{E}\left\{ \frac{NB(x,V_i,w)}{\mu(V_i)-\ell(V_i)} \middle| X_i = x, W_i = w \right\}}{\mathbb{E}\left[\frac{1}{\mu(V_i)-\ell(V_i)} \middle| \right]} \\
&= \frac{\mathbb{E}\left[\frac{1}{\mu(V_i)-\ell(V_i)} \middle| X_i = x, W_i = w \right] \cdot NB(x,w) - Cov\left[\frac{1}{\mu(V_i)-\ell(V_i)}, NB(x,V_i,w) \middle| X_i = x, W_i = w \right]}{\mathbb{E}\left\{ \frac{1}{\mu(V_i)-\ell(V_i)} \middle| X_i = x, W_i = w \right\}} \\
&\geq NB(x,w) - \sqrt{\frac{Var\left[\frac{1}{\mu(V_i)-\ell(V_i)} \middle| X_i = x, W_i = w \right]}{\mathbb{E}\left\{ \frac{1}{\mu(V_i)-\ell(V_i)} \middle| X_i = x, W_i = w \right\}^2} \cdot Var\left[NB(x,V_i,w) \middle| X_i = x, W_i = w \right]} \\
&\geq NB(x,w) - Var\left[NB(x,V_i,w) \middle| X_i = x, W_i = w \right] \\
&\geq NB(x,w) - NB(x,w) \cdot (1 - NB(x,w)) = NB(x,w)^2
\end{aligned}$$

where $NB(x, v, w) := P(0 < R_i < \bar{R}|x, v, w) = P(\ell(V_i) \leq H_i \leq \mu(V_i)|x, v, w)$ and $NB(x, w) = \mathbb{E}[NB(x, V_i, w)|x, w]$ is the observable probability of not bunching given $(X_i, W_i) = (x, w)$. The third inequality uses the assumption that $\frac{Var\left[\frac{1}{\mu(V_i)-\ell(V_i)} \middle| x,w \right]}{\mathbb{E}\left\{ \frac{1}{\mu(V_i)-\ell(V_i)} \middle| x,w \right\}^2} \leq Var\left[NB(x,V_i,w) \middle| x,w \right]$ and the last one that $Var\left[NB(x,V_i,w) \middle| x,w \right] \leq NB(x,w) \cdot (1 - NB(x,w))$ since $NB(x, v, w) \in [0, 1]$ for all x, v, w.

From this notation, we obtain the form written in Proposition 6 by noting that $NB(X_i, V_i, W_i) = 1 - \mathcal{B}_i$. Note that $Var\left[NB(x,V_i,w) \middle| x,w \right] = Var\left[\mathcal{B}_i \middle| x,w \right]$.

References

ABADIE, A. (2003). "Semiparametric instrumental variable estimation of treatment response models". *Journal of Econometrics* 113 (2), pp. 231–263.

BARREIRA, P., BASILICO, M. and BOLOTNYY, V. (2021). "Graduate Student Mental Health: Lessons from American Economics Departments". *Journal of Economic Literature* Forthcoming.

BARRINGTON-LEIGH, C. (2024). "The econometrics of happiness: Are we underestimating the returns to education and income?" *Journal of Public Economics* 230, p. 105052.

BENJAMIN, D. J., HEFFETZ, O., KIMBALL, M. S. and REES-JONES, A. (2014). "Can Marginal Rates of Substitution Be Inferred from Happiness Data? Evidence from Residency Choices". *American Economic Review* 104 (11), pp. 3498–3528.

BLOMQUIST, S., KUMAR, A., LIANG, C.-Y. and NEWEY, W. (2021). "On Bunching and Identification of the Taxable Income Elasticity". *Journal of Political Economy* 129 (8).

BLUNDELL, R., KRISTENSEN, D. and MATZKIN, R. (Dec. 2017). *Individual counterfactuals with multidimensional unobserved heterogeneity*. CeMMAP working papers 60/17. Institute for Fiscal Studies.

BOND, T. N. and LANG, K. (2019). "The Sad Truth about Happiness Scales". *Journal of Political Economy* 127 (4), pp. 1629–1640.

CARD, D., MAS, A., MORETTI, E. and SAEZ, E. (2012). "Inequality at Work: The Effect of Peer Salaries on Job Satisfaction". *American Economic Review* 102 (6), pp. 2981–3003.

CATTANEO, M. D. and JANSSON, M. (2018). "Kernel-Based Semiparametric Estimators: Small Bandwidth Asymptotics and Bootstrap Consistency". *Econometrica* 86 (3), pp. 955–995.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I., HODERLEIN, S., HOLZMANN, H. and NEWEY, W. (2015). "Nonparametric identification in panels using quantiles". *Journal of Econometrics* 188 (2). Heterogeneity in Panel Data and in Nonparametric Analysis in honor of Professor Cheng Hsiao, pp. 378–392.

CHERNOZHUKOV, V., FERNÁNDEZ-VAL, I. and NEWEY, W. K. (2019). "Nonseparable multinomial choice models in cross-section and panel data". *Journal of Econometrics* 211 (1). Annals Issue in Honor of Jerry A. Hausman, pp. 104–116.

CHERNOZHUKOV, V., NEWEY, W. K. and SINGH, R. (2022). "Automatic Debiased Machine Learning of Causal and Structural Effects". *Econometrica* 90 (3), pp. 967–1027.

CUNHA, F., HECKMAN, J. J. and NAVARRO, S. (2007). "The Identification and Economic Content of Ordered Choice Models with Stochastic Thresholds". *International Economic Review* 48 (4).

DEATON, A. (2018). "What do self-reports of wellbeing say about life-cycle theory and policy?" *Journal of Public Economics* 162. In Honor of Sir Tony Atkinson (1944-2017), pp. 18–25.

D'HAULTFŒUILLE, X. and FÉVRIER, P. (2015). "Identification of Nonseparable Triangular Models With Discrete Instruments". *Econometrica* 83 (3), pp. 1199–1210.

DI TELLA, R., MACCULLOCH, R. J. and OSWALD, A. J. (2001). "Preferences over Inflation and Unemployment: Evidence from Surveys of Happiness". *American Economic Review* 91 (1), 335–341.

DWYER, R. J. and DUNN, E. W. (2022). "Wealth redistribution promotes happiness". *Proceedings of the National Academy of Sciences* 119 (46), e2211123119.

FAN, Y. and PARK, S. S. (2010). "Sharp bounds on the distribution of treatmetn effects and their statistical inference". *Econometric Theory* 26 (3), pp. 931–951.

FENG, J. and LEE, S. (2025). "Individual welfare analysis: Random quasilinear utility, independence, and confidence bounds". *Journal of Econometrics* 247, p. 105927.

FLEMING, M. (1952). "A Cardinal Concept of Welfare". *The Quarterly Journal of Economics* 66 (3), pp. 366–384.

GALLUP (2021). *Gallup Worldwide Research Methodology and Codebook*. Gallup, Inc.

GOFF, L. (2022). *Treatment Effects in Bunching Designs: The Impact of Mandatory Overtime Pay on Hours*. arXiv: 2205.10310 [econ.EM].

— (2025). *Identifying causal effects with subjective ordinal outcomes*. arXiv: 2212.14622v4 [econ.EM].

GOFF, L., KÉDAGNI, D. and WU, H. (2024). *Testing Identifying Assumptions in Parametric Separable Models: A Conditional Moment Inequality Approach*. arXiv: 2410.12098 [econ.EM].

GREENE, W. (2005). *Econometric Analysis, 7th Edition*. Pearson.

HAMERMESH, D. S. (2004). "Subjective Outcomes in Economics". *Southern Economic Journal* 71 (1), pp. 1–11.

HARSANYI, J. C. (1955). "Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility". *Journal of Political Economy* 63 (4), pp. 309–321.

HELLIWELL, J. F. and BARRINGTON-LEIGH, C. P. (2010). "Viewpoint: Measuring and understanding subjective well-being". eng. *The Canadian journal of economics* 43 (3), pp. 729–753.

HODERLEIN, S. and MAMMEN, E. (2008). "Identification and estimation of local average derivatives in non-separable models without monotonicity". *Econometrics Journal* 00 (501), pp. 1–25.

HODERLEIN, S., HOLZMANN, H., KASY, M. and MEISTER, A. (June 2016). "Corrigendum: Instrumental Variables with Unrestricted Heterogeneity and Continuous Treatment". *The Review of Economic Studies* 84 (2), pp. 964–968.

HODERLEIN, S. and MAMMEN, E. (2007). "Identification of Marginal Effects in Nonseparable Models without Monotonicity". *Econometrica* 75 (5), pp. 1513–1518.

HODERLEIN, S. and SASAKI, Y. (Aug. 2013). *Outcome Conditioned Treatment Effects*. Boston College Working Papers in Economics 840. Boston College Department of Economics.

HODERLEIN, S., SIFLINGER, B. and WINTER, J. (2015). *Identification of structural models in the presence of measurement error due to rounding in survey responses*.

HOSSEINI, R. (2010). *Quantiles Equivariance*. arXiv: 1004.0533.

HU, Y. (2008). "Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution". *Journal of Econometrics* 144 (1), pp. 27–61.

IMBENS, G. W. and ANGRIST, J. D. (1994). "Identification and Estimation of Local Average Treatment Effects". *Econometrica* 62 (2), pp. 467–475.

IMBENS, G. W. and NEWEY, W. K. (2009). "Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity". *Econometrica* 77 (5), pp. 1481–1512.

KAISER, C. and VENDRIK, M. C. M. (2022). "How much can we learn from happiness data?" *Working Paper*.

KASY, M. (2022). "Who wins, who loses? Identification of conditional causal effects, and the welfare impact of changing wages". *Journal of Econometrics* 226 (1). Annals Issue in Honor of Gary Chamberlain, pp. 155–170.

KLEVEN, H. J. (2016). "Bunching". *Annual Review of Economics* 8 (1), pp. 435–464.

KLOSIN, S. and VILGALYS, M. (2023). *Estimating Continuous Treatment Effects in Panel Data using Machine Learning with a Climate Application*. arXiv: `2207.08789 [econ.EM]`.

LINDQVIST, E., ÖSTLING, R. and CESARINI, D. (Feb. 2020). "Long-Run Effects of Lottery Wealth on Psychological Well-Being". *The Review of Economic Studies* 87 (6), pp. 2703–2726.

LUTTMER, E. F. P. (Aug. 2005). "Neighbors as Negatives: Relative Earnings and Well-Being*". *The Quarterly Journal of Economics* 120 (3), pp. 963–1002.

MANSKI, C. F. and TAMER, E. (2002). "Inference on Regressions with Interval Data on a Regressor or Outcome". *Econometrica* 70 (2), pp. 519–546.

MATZKIN, R. L. (1992). "Nonparametric and Distribution-Free Estimation of the Binary Threshold Crossing and The Binary Choice Models". *Econometrica* 60 (2), pp. 239–270.

— (1994). "Chapter 42 Restrictions of economic theory in nonparametric methods". Vol. 4. Handbook of Econometrics. Elsevier, pp. 2523–2558.

— (2019). "Constructive identification in some nonseparable discrete choice models". *Journal of Econometrics* 211 (1). Annals Issue in Honor of Jerry A. Hausman, pp. 83–103.

MILGROM, P. and SHANNON, C. (1994). "Monotone Comparative Statics". *Econometrica* 62 (1), pp. 157–180.

NOCKE, V. and SCHUTZ, N. (2017). "Quasi-linear integrability". *Journal of Economic Theory* 169, pp. 603–628.

OPARINA, E. and SRISUMA, S. (2022). "Analyzing Subjective Well-Being Data with Misclassification". *Journal of Business & Economic Statistics* 40 (2), pp. 730–743.

PEREZ-TRUGLIA, R. (2020). "The Effects of Income Transparency on Well-Being: Evidence from a Natural Experiment". *American Economic Review* 110 (4), pp. 1019–54.

PFANZAGL, P. (1979). "Conditional Distributions as Derivatives". *The Annals of Probability* 7 (6), pp. 1046 –1050.

ROOIJ, M. van, COIBION, O., GEORGARAKOS, D., CANDIA, B. and GORODNICHENKO, Y. (2024). *Keeping Up with the Jansens: Causal Peer Effects on Household Spending, Beliefs and Happiness*. Working Paper 32107. National Bureau of Economic Research.

SAEZ, E. (2010). "Do Taxpayers Bunch at Kink Points?" *American Economic Journal: Economic Policy* 2 (3), pp. 180–212.

SASAKI, Y. (2015). "What do Quantile Regression Identify for General Structural Functions?" *Econometric Theory* 31 (5), 1102–1116.

SCHENNACH, S. M. and HU, Y. (2013). "Nonparametric Identification and Semiparametric Estimation of Classical Measurement Error Models Without Side Information". *Journal of the American Statistical Association* 108 (501), pp. 177–186.

SCHRÖDER, C. and YITZHAKI, S. (2017). "Revisiting the evidence for cardinal treatment of ordinal variables". *European Economic Review* 92, pp. 337–358.

TORGOVITSKY, A. (2015). "Identification of Nonseparable Models Using Instruments With Small Support". *Econometrica* 83 (3), pp. 1185–1197.

VAN DER VAART, A. (2000). *Asymptotic Statistics*. Asymptotic Statistics. Cambridge University Press.

VAN PRAAG, B. M. (1991). "Ordinal and cardinal utility: An integration of the two dimensions of the welfare concept". *Journal of Econometrics* 50 (1), pp. 69–89.