Identifying causal effects with subjective outcomes

Leonard Goff[*]

This version: February 19, 2024

Abstract

Survey questions often elicit responses on ordered scales for which the definitions of the categories are subjective, possibly varying by individual. This paper clarifies what is learned when these subjective responses are used as an outcome in regression-based causal inference. When a continuous treatment variable is statistically independent of both i) potential outcomes; and ii) heterogeneity in reporting styles, a nonparametric regression of response category number on that variable uncovers a positively-weighted linear combination of causal effects among individuals who are on the margin between adjacent response categories. Though the weights do not integrate to one, the *ratio* of local regression derivatives with respect to two such explanatory variables identifies the relative magnitudes of convex averages of their effects. When results are extended to discrete treatment variables, different weighting schemes apply to different regressors, making comparisons of magnitude less informative. Under further assumptions, I obtain bounds on the possible bias when comparing the effects of a discrete treatment variable to those of another discrete or continuous treatment variable.

1 Introduction

Survey questions often ask respondents to choose from a set of ordered categories that lack clear definitions, and are thus interpreted subjectively by those respondents. These categorical responses are then commonly used as outcome variables in social science research.[1] Examples include self-reported health status in health economics, job satisfaction in labor economics, or life-satisfaction questions as a measure of overall wellbeing.

A key question when analyzing these responses as an outcome is how "reporting functions"—the way that individuals map an underlying latent variable into one of the available response categories—impact conclusions drawn from the data.[2] 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 of individuals, absent strong restrictions on its unobservable distribution. More fundamentally, if the response categories lack objective definitions, reporting functions might also vary between individuals, possibly confounding any attempt to study relationships between explanatory variables and the latent variable.

This paper shows that the observed categorical responses can still be informative about causal effects on this latent outcome, despite the dual threats of reporting functions being

[*] Department of Economics, University of Calgary. For useful conversations, I thank Christopher Barrington-Leigh, Carol Caetano, Andrew Clark, Ben Crost, John Helliwell, Peter Hull, Caspar Kaiser, Jean-William Laliberté, Simon Lee, Guy Mayraz, Max Norton, Bernard Salanié, Adam Rosen, Kevin Song and Sam Viavant.

[1] See e.g. Hamermesh (2004) for an overview.

a) unknown to the researcher; and b) heterogenous across individuals. This paper takes the perspective of Bond and Lang (2019) that the latent variable driving individuals' responses is the researcher's ultimate outcome of interest, and establish new results relating the observed joint distribution of responses and covariates to the causal effects of those covariates on the latent variable. I do so by extending the selection-on-observables assumption—familiar from causal inference—that explanatory variables are (conditionally) independent of potential outcomes, adding to it the assumption that explanatory variables are also (conditionally) independent of heterogeneity in reporting functions. I then consider the practice that is common in empirical work of regressing categorical response numbers on explanatory variables, and show how the regression estimand from this exercise can be interpreted in terms of the causal effects of those regressors on the latent variable of interest.

Specifically, I consider a model of the general form:

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

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 $\mathcal{R}$ (e.g. $\mathcal{R} = \{0, 1\}$ for a binary yes/no question). For simplicity, I consider a scalar latent variable before later generalizing to $K > 1$.

The function $h_i(x)$ above denotes the potential outcomes of H for individual i, indicating the value of H that would occur if a vector of observed explanatory variables X took counterfactual value x. The function $r_i(h)$ represents individual i's 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, and the main statistical assumption of the model can be stated simply as $X_i \perp\!\!\!\perp (U_i, V_i)$. This is later relaxed to *conditional* independence given control variables. The researcher's objective is to learn how $h_i(x)$ varies with x, while observing only R_i and X_i.

One of the main results of this paper is that if x_1 and x_2 reflect two continuously distributed components of the vector x, then

$$\frac{\frac{\partial}{\partial_{x_1}}\mathbb{E}[R_i|X_i = x]}{\frac{\partial}{\partial_{x_2}}\mathbb{E}[R_i|X_i = x]} = \frac{\tilde{\beta}_1(x)}{\tilde{\beta}_2(x)} \tag{1}$$

where $\tilde{\beta}_j(x)$ reflects a weighted average of the causal effect of a small change in the j^{th} component of X on H, when $X = x$. In particular, when $\mathcal{R}$ reflects a set of integers $\mathcal{R} = \{0, 1, \ldots \bar{R}\}$ for some $\bar{R}$, the quantity $\tilde{\beta}_j(x)$ averages $\partial_{x_j} h(x, U_i)$ over individuals i who are on the margin between two response categories $r-1$ and r, for any $r \in \{1, \ldots \bar{R}\}$.

[2] The use of the term "reporting function" for subjective data appears to have first appeared in the economics literature in Oswald (2008), but the general concept certainly predates its discussion in economics (e.g. Banks and Coleman 1981).

More generally, when one compares the mean of R between any two values x and x' of the vector X, we have that:

$$\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x] = \mathbb{E}\left[\bar{f}(\Delta_i, V_i, x) \cdot \Delta_i\right] \tag{2}$$

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, and $\bar{f}(\Delta, v, x) \geq 0$ for all Δ, v, x.[3] Eq. (2) 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. Meanwhile, Eq. (1) implies that when the conditional expectation of R given X is linear in X (justifying *linear* regression), the ratio of regression coefficients corresponding to two continuously distributed explanatory variables can capture the relative magnitudes of marginal causal effects corresponding to each.

Despite a growing trend in papers leveraging natural experiments with subjective outcome data,[4] empiricists have lacked formal results to interpret precisely what is estimated by regressions in which "subjective" ordinal responses R are used as the outcome. This paper helps to fill this gap by showing that when the selection-on-observables research design is extended to include reporting-function heterogeneity, the shape of how the regression function of integer category numbers varies with X recovers positive linear aggregations of causal effects of X on H, with intuitive weights.[5] The results illuminate how mean regression can remain a useful—while in some ways limited—tool for causal inference about H, without assuming cardinality or interpersonal comparability of H.

For ease of exposition, I take as a running example survey questions that ask respondents about their overall satisfaction with life, and refer to H as "happiness".[6] This draws connections to the notion of cardinal utility as a measure of welfare (Fleming, 1952; Harsanyi, 1955), and motivates the treatment of H as an outcome of normative interest. However, results are equally applicable to other outcomes elicited on ordered scales, e.g. self-reported health status, mental health indicators, job satisfaction, ratings of products and services, or other settings in which ordered response models might be employed with random individual-specific thresholds.[7]

Intuition for the main results: To appreciate the the role of the independence assumption $X_i \perp\!\!\!\perp (U_i, V_i)$ in the above results, consider a study investigating the connection between wealth and H_i conceived of as general satisfaction with life. Suppose that the study makes use of life evaluations from the popular "Cantril Ladder" question, which asks (Gallup,

[3] The function $\bar{f}$ is defined in Sec. 4, and no longer depends upon Δ as $x' \to x$ and the difference becomes a derivative.

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

[5] When the researcher is interested in establishing correlations rather than causation, the same results capture changes to the conditional quantile function of the underlying latent variable, rather than causal effects.

[6] I do this for simplicity only, ignoring e.g. important distinctions between hedonic, affective and evaluative notions of wellbeing (Deaton, 2018; Helliwell and Barrington-Leigh, 2010).

[7] A broad class of such examples are survey questions that 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").

2021): *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?*. In this example, the response space is $\mathcal{R} = \{0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10\}$.

We can appreciate the importance of independence between X_i and V_i by supposing that the mean response R_i among wealthy individuals $X_i = 1$ is higher than among less-wealthy individuals $X_i = 0$. While this correlation could be due to wealthier individuals tending to have higher H_i, it could in principle instead simply reflect that wealthier individuals tend to operationalize the question differently, having lower threshold values of experienced happiness H_i at which they would evaluate their life as being in a particular category, e.g. an "eight" or a "nine".

Contrast the above study with one that compares the life satisfaction of lottery winners to those who played but did not win.[8] Before the lottery, heterogeneity V_i in reporting functions is balanced between treatment and control groups by virtue of the random selection of winners. This guarantees balance after the win as well, if individuals have fixed reporting functions that do not themselves change due to winning the lottery. Equations (1) and (2) show that randomization can be helpful both in confronting the classic problem of selection bias (wealthier individuals tend to differ from the less-wealthy in many unobserved dimensions relevant to their happiness—captured by U_i) *and* the possibility of confounding heterogeneity V_i in how they map that happiness into a subjectively-defined response category in $\mathcal{R}$, as allowed by the question.

Relationship to literature: The results in this paper are related both to the literature on subjective well-being (SWB)—which is often measured using survey questions involving integer scales such as the Cantril Ladder—as well as to ordered response models which are used widely throughout economics.

Empirical studies using SWB data often regress the integer category number R_i on explanatory variables X_i, either implicitly or explicitly treating observed responses as a direct observation of the ultimate outcome of interest (amounting to $r_i(h) = h$, or "cardinality"). This justifies familiar regression-based approaches to studying the average effects of X_i on R_i, given selection on observables. Other authors instead estimate parametric ordered response models (such as the ordered probit or logit) to recover causal effects, on the grounds that R_i represents an ordinal variable only. However, such approaches generally trade the restrictiveness of imposing cardinality for the potential for misspecification bias, by introducing strong parametric assumptions. The ordered probit model for example assumes that $h(x, u) = x'\beta + u$ where u follows a normal distribution, a structure that cannot be verified empirically.

[8] As an example, Lindqvist et al. (2020) document that lottery winners in Sweden report higher life satisfaction than lottery losers, persistent for many years after their payout.

A recent methodological literature that has considered the robustness of conclusions drawn from SWB data whether using either the ordered response or regression approach, if one accepts that comparisons of R_i are only *ordinally* meaningful (Schröder and Yitzhaki 2017; Bond and Lang 2019; Kaiser and Vendrik 2022; Chen et al. 2022; Chesher et al. 2022). These papers mostly abstract away from of heterogeneity V_i in reporting functions, representing a special case of my setup in which $r_i(\cdot) = r(\cdot)$ is common across individuals.[9] A common reporting function r amounts to assuming it possible to make interpersonal comparisons of happiness, since if $R_i > R_j$ for two individuals i and j it must then be the case that $H_i > H_j$, provided that r is weakly increasing. The possibility of interpersonal comparisons of utility is widely disputed (Binmore, 2009).

However, comparisons of H between specific pairs of individuals is rarely the goal of analysis. Instead, researchers are typically interested in documenting features of how the conditional distribution of H varies with X, or to establish the causal effects of components of X on H.

A major challenge to this enterprise is articulated by Bond and Lang (2019) (henceforth, BL), who suggest that using observations of R_i to learn about H_i can be misleading. BL point out that the implied ranking of mean happiness H between two groups can differ depending on which function $r(\cdot)$ generated the data. To see the issue, let x' and x stand for two groups defined by X. In general, $\mathbb{E}[R_i|X_i = x'] \geq \mathbb{E}[R_i|X_i = x]$ is insufficient to conclude that $\mathbb{E}[H_i|X_i = x'] \geq \mathbb{E}[H_i|X_i = x]$. Intuitively, this also depends on how "concave" or "convex" the function $r(h)$ is in h. BL show that even if the distribution of R_i given $X_i = x'$ stochastically dominates that of R_i given $X_i = x$ (a much stronger condition than having a higher mean), it is still possible that $\mathbb{E}[H_i|X_i = x'] < \mathbb{E}[H_i|X_i = x]$.[10]

This result may appear to be odds with Eqs. (1) and (2) above, which suggest that comparing $\mathbb{E}[R_i|X_i = x]$ across values of x can be informative about causal effects. However BL focus their attention on the sign of $\mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x]$, which given random assignment of X (ignoring control variables for now) corresponds to the overall average treatment effect $\mathbb{E}[\Delta_i]$, in the notation of Eq. (2). Although this paper finds that regressions of R on X identify positively-weighted combinations of the causal effects of X on H, the weights cannot be chosen by the researcher, and relative to the uniform weighting of $\mathbb{E}[\Delta_i]$ may over or under-represent particular types of individuals. In principle, studying the conditional mean of R_i could therefore lead to misleading inferences if causal effects are sufficiently heterogeneous: for example if they are large and positive for some while large and negative for others.[11] As a means of addressing this, I show that average characteristics of the individuals that are suitably "marginal" between response categories to be reflected in the regression function can be identified

[9] A common reporting function r is also assumed in the literature on measuring inequality from ordinal data (Allison and Foster, 2004; Cowell and Flachaire, 2017); though Kaplan and Zhao (2022) allow group-level differences in reporting.

[10] By saying that $R|X = x'$ (first order) stochastically dominates $R|X = x$, I mean that $P(R_i \leq r|X_i = x') \leq P(R_i \leq r|X_i = x')$ for all values r.

[11] This is arguably less of a problem when conducting causal inference as compared with comparing actual distributions of happiness, as BL do. Causal effects may be fairly homogeneous within observable strata, while the distribution functions of realized happiness for two groups are likely to cross or exhibit more widely varying gaps.

from the data (assuming that these characteristics are themselves unaffected by X_i). This allows the researcher to reason about the subpopulations that contribute.

Existing results in the SWB literature that allow for heterogeneity in reporting functions (e.g. Layard et al. 2008) incorporate additive heterogeneity in the mapping between R and H into the error term of fully linear models relating X, H and R.[12] Instead, I let latent happiness H and responses R exist on entirely different scales, in common with ordered response models. This also distinguishes the approach of this paper from models of rounding (Hoderlein et al., 2015), classical measurement error (Schennach and Hu, 2013), or discrete missclasification (Hu, 2008; Oparina and Srisuma, 2022).[13]

My results show what simple mean regressions having R on the left-hand side identify in a model in which reporting functions are left fully unrestricted except that heterogeneity in these functions is unrelated to the regressors, and without any side information. My setup nests familiar econometric models for ordered response, such as the ordered probit or logit, but drops any parametric or functional-form assumptions. It also generalizes nonparametric ordered response models that assume scalar or additively separable heterogeneity (e.g. Matzkin 1992; Matzkin 1994).

Outline: The remainder of this paper proceeds as follows. In Section 2, I propose a nonparametric model of ordered response with non-separable heterogeneity, which assumes only that i) each individual's responses are weakly increasing in their value of the latent outcome of interest (Appendix C extends to a multivariate latent variable); and ii) that a vector of treatment variables X is conditionally independent of all unobserved heterogeneity in the model. While the model serves as a potential outcomes notation when the goal is causal inference, it can also serve as a representation of the unobserved conditional distribution of the latent variable, without reference to causality.

Section 3 establishes my main identification result when there is continuous variation in a component of X, showing that the slope of changes in the conditional distribution of responses with respect to that variable identifies a positively-weighted linear combination of heterogeneous causal responses, among individuals who are on the margin between adjacent response categories. When the sign of those effects is common across individuals, the regression derivative of integer response numbers with respect to X thus reveals the direction of causal effects. More strikingly, it remains meaningful to compare the *magnitudes* of two partial derivatives (at the same value of X) of the conditional expectation function of responses with general heterogeneity in causal effects, as we saw through Eq. (1). I show that under additional conditions, the ratio of regression derivatives in fact

[12] Further results that allow for reporting heterogeneity rely on auxiliary data sources or particular models of that heterogeneity. The approach of "anchoring vignettes" (King et al., 2004) adjusts for heterogeneity by asking respondents to rate the hypothetical well-being of others (see e.g. Kapteyn et al. 2013; Molina 2017; Montgomery 2022 for applications). Kaiser (2022) uses memories of one's own past life satisfaction, and Liu and Netzer (2023) use survey response times. Barrington-Leigh (2018) extends parametric ordered response models to allow for individuals to differ with respect to whether they make use of the entire response scale, or restrict themselves only to "focal" values.

[13] Other measurement error models (Hu and Schennach, 2008; Nadai and Lewbel, 2016; Breunig and Martin, 2020) might also be applied to subjective outcomes, but identification typically relies on instrumental variables or special regressors.

captures a local average marginal rate of substitution between the two X.

In Section 4, I turn to identification with a discrete treatment variable. In this case, a comparison of mean responses at two values of the covariates X again captures a positively weighted combination of causal effects, as we saw in Eq. (2). However, the weights are no longer "local" to particular threshold values of happiness, and unfortunately the *total* weight placed on causal effects generally differs from that recovered by a derivative using a continuous X. This suggests that comparing the magnitudes of discrete regression coefficients to one another or comparing the coefficients of a discrete and a continuous regressor is not guaranteed to be quantitatively meaningful. I assess this implication through simulations with a variety of assumed distributions of latent happiness. While the distortion due to differential weighting can be severe in principle, I only find evidence that it is in practice when treatment effects are made implausibly large.

To shed light on these simulation results, I consider a "dense response limit" in which we view the number of response categories along a sequence that tends towards infinity. This delivers analytical results that hold approximately when there are many categories of response. I use the dense response limit to derive bounds on the ratio of the total weight that the conditional expectation applies to causal effects when comparing continuous to discrete variation in X. In particular, when individual reporting functions are approximately linear, discrete contrasts will tend to overstate causal effects relative to continuous ones, by a factor that is upper bounded by two. I also demonstrate a second set of bounds that are typically much narrower, and hold when individuals furthermore do not differ too much in how "sensitive" their reporting functions are. These theoretical bounds are quite conservative when compared with the simulation results, but hold without parametric knowledge of the underlying distribution of happiness.

Some further results are given in the appendices. In Appendix A I discuss testable implications of my main model, in particular given the assumption that reporting functions are unaffected by covariates. Appendix B illustrates the results of the paper by estimating causal effects in a synthetic dataset. In Appendix C.1, I extend results for the continuous X case to a nonparametric instrumental variables setup. Appendix C.2 extends the model to embed a second notion of "subjectivity" in subjective responses: not only do individuals differ in their definitions of the response categories in terms of a latent variable, but also in how they conceptualize the latent variable that the survey question is asking about. In Appendix D, I discuss identification when the researcher is interested in understanding the correlation between the underlying latent variable and covariates, rather than establishing causal effects.

2 Model

To begin let us take for granted that there exists a meaningful latent value H_i for each individual, that the researcher is ultimately interested in as an outcome. With the life satisfaction example in mind, I will often refer to H_i as $i's$ underlying "happiness", which the researcher aims to learn about given those individuals' responses R_i.[14]

The researcher observes a sample of (R_i, X_i) across individuals i generated as:

$$R_i = r_i(H_i) = r(H_i, V_i) \tag{3}$$

$$H_i = h_i(X_i) = h(X_i, U_i) \tag{4}$$

where $r_i(h)$ is in individual-specific function mapping happiness h to the space of possible responses $\mathcal{R}$. 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 $r_i(h)$. Figure 1 depicts some examples of reporting functions when $\mathcal{R} = \{0, 1, 2\}$.

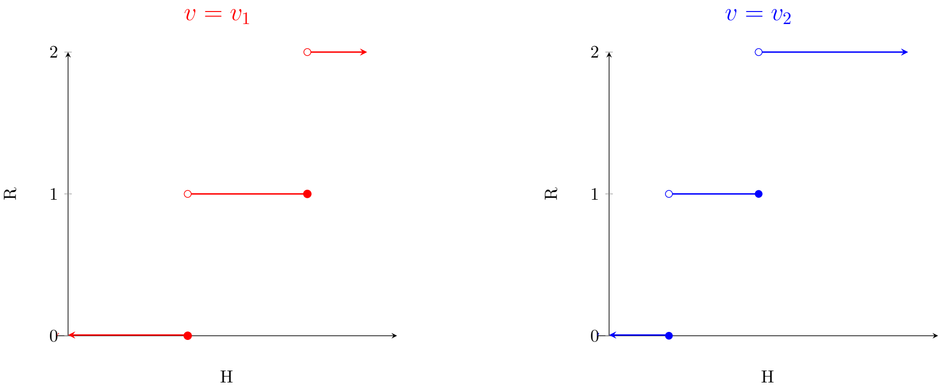


Figure 1: Examples of two different reporting functions, in a case with three categories: $\mathcal{R} = \{0, 1, 2\}$.

Similarly, for each individual there is a function $h_i(\cdot)$ mapping values of a vector of explanatory variables X into a value of H via (4), where heterogeneity in the function $h_i(\cdot)$ is represented by parameter $U_i \in \mathcal{U} \subseteq \mathbb{R}^{d_u}$. The intended interpretation of the function $h_i(x)$ is that it denotes potential outcomes for individual i as a function of counterfactual values of x, in some set of possible treatments $\mathcal{X} \subseteq \mathbb{R}^{d_x}$. The function h is our object of interest: how it varies with x holding u fixed yields the causal effect of that change on H. Since the dimension d_u is again left unrestricted, the above model places no restriction on how heterogeneous these causal effects can be across individuals.

Remark: An alternative interpretation of $h(x, u)$ that requires no causal assumptions is that it 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.[15] This representation is helpful when causal effects are not the target, and the researcher is instead interested in the more

[14] The model extends naturally to a setting in which the definition of "H" is itself subjective, in the sense that different individuals use different latent variables when constructing their responses. The key requirement is that these subjectively defined latent variables in turn reflect increasing transformations of an objective variable of interest. See Appendix C.2.

[15] In particular, let $\theta_i := F_{H|XV}(H_i|X_i, V_i)$ be i's "rank" in the conditional happiness distribution of individuals sharing their value of X and V, where $F_{H|XV}$ denotes a cumulative distribution function of H. Now let $U_i = (\theta_i, V_i)^T$, and define $h(x, u) := Q_{H|XV}(\theta|x, v)$ for any $u = (\theta, v)^T$, where $Q_{H|XV}$ denotes the conditional quantile function of H given X and V. Eq. (4) now follows from these definitions. See Appendix D for details.

modest goal of uncovering statistical features of the joint distribution of H_i and X_i.

Note that model (3)-(4) embeds an exclusion restriction: X does not directly enter in the equation for R, and only affects reports through H. This is important for drawing inferences about the relationship between H and X from the observable joint distribution of R and X. The model can be generalized slightly to allow reporting behavior to depend directly on observables, as described in Appendix A. The appendix also describes how restrictions on effect heterogeneity yield overidentification restrictions that can then be used to test the assumption that reporting functions are invariant with respect to X.

The following two subsections introduce the additional identifying assumptions of the model: first, that reporting functions are weakly increasing in h; and second, that the researcher as exogenous variation in some components of X_i.

2.1 Weakly increasing reporting

The main assumption that I make about the reporting functions $r_i(\cdot)$ themselves is that they are *increasing* in H_i:

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

The left-continuity assumption of HONEST is essentially a normalization, since any weakly increasing function of bounded variation is continuous except at isolated points within its support.[16] The first part of Assumption HONEST rules out cases in which individuals would report a lower value of R if H were increased.

The following lemma shows that Assumption HONEST is equivalent to there being a set of "thresholds" $\tau_v(r)$ that separate the ordered categories in $\mathcal{R}$. This characterization is useful in developing formal results.

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

$$r(h, v) \leq r \iff h \leq \tau_v(r) \tag{5}$$

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.

Proof. See Appendix G. □

As an illustration of Lemma 1, suppose that $\mathcal{R} = \{0, 1, \dots \bar{R}\}$ for some integer $\bar{R}$. Then

[16] 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.

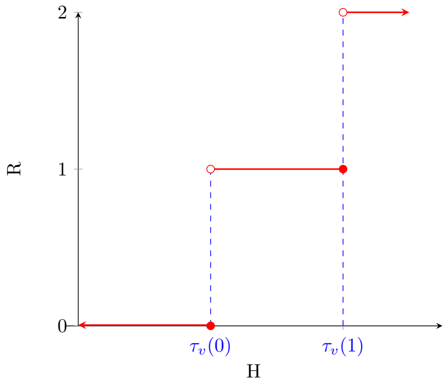


Figure 2: Depiction of the thresholds corresponding to reporting function $r(\cdot, v)$ for $v = v_1$ from Figure 1.

Lemma 1 implies that any given reporting function $r(h, v)$ can be written as:

$$r(h,v) = \begin{cases} 0 & \text{if } h \le \tau_v(0) \\ 1 & \text{if } \tau_v(0) < h \le \tau_v(1) \\ 2 & \text{if } \tau_v(1) < h \le \tau_v(2) \\ \vdots & \\ \bar{R} & \text{if } h > \tau_v(\bar{R}-1) \end{cases} \tag{6}$$

Remark: Assumption HONEST does not require that respondents are motivated only by "honesty" when choosing R_i. Instead, they may have direct preferences for certain response categories. 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. As an example, let us further assume that the utility function takes the form

$$u(r, h, v) = \phi_v(r) - |h_v^*(r) - h|$$

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 $h_v^*(r)$ that is perceived by them to correspond to response r. Provided that $h_v^*(r)$ is strictly increasing in r (i.e. higher responses are subjectively associated with higher values of happiness), then u satisfies the property of *increasing differences* (cf. Milgrom and Shannon 1994) in (r, h), which in turn implies HONEST.[17]

2.2 Conditional independence

The final piece of the model is a conditional independence assumption for variation in X. Denote the elements of the vector X as $X_i = (X_{1i}, X_{2i} \dots X_{J,i}, W_i)^T$, where the first J components of X will be variables for which causal effects are of interest, while the remaining components $W_i = (X_{J+1,i} \dots X_{d_x,i})'$, will serve as control variables. For a given

[17] Note that heterogeneity v in this form for utility need not be additively separable from quantities that depend on x (i.e. h). Such separability is shown by Allen and Rehbeck (2019) to admit important identification results for latent utility.

value x of X, I will use w to denote these final $d_x - J$ components of x.[18]

With this notation, I now suppose that conditional on W, each X_j is as good as randomly assigned in the following sense:

Assumption EXOG (conditionally exogenous components of X). *Both*

- $\{X_{ji} \perp\!\!\!\perp V_i\} \mid W_i$
- $\{X_{ji} \perp\!\!\!\perp U_i\} \mid (W_i, V_i)$

for each $j = 1 \dots J$.

Note that Assumption EXOG follows if for each j:

$$\{X_{ji} \perp\!\!\!\perp (U_i, V_i)\} \mid W_i \tag{7}$$

Eq. (7) provides a natural foundation for EXOG and is easier to interpret, but is technically stronger than the results require. For causal inference, an assumption like $\{X_j \perp\!\!\!\perp U\}|W$ is generally already necessary for identification: one needs some kind of experiment or natural experiment providing exogenous variation in X_j. Eq. (7) then simply requires this natural experiment to make X_j (conditionally) independent of V as well. Note that under EXOG, U and V may be arbitrarily correlated with one another (e.g. if happier individuals have more optimistic reporting functions).[19] In Section C.1, I relax EXOG to consider identification using instrumental variables.

The assumption that response behavior is independent of a treatment variable may be restrictive in many contexts, especially in the absence of a credible research design. For example, Barrington-Leigh (2018) notes that the tendency to bunch at endpoints or the mid-point of scales for life-satisfaction questions is higher among individuals with less formal education, which suggests that a regression of life satisfaction on years of schooling might conflate reporting heterogeneity with variation in actual life satisfaction.[20] While a natural experiment could yield variation in schooling that is orthogonal to this heterogeneity, the assumption that education does not directly change individuals' reporting functions (i.e. their definition of say an "eight" out of ten in life satisfaction) may be strong. I show that when causal effects are assumed to be homogeneous, the model has testable implications that can be used to assess this assumption of reporting function invariance. More broadly, whether reporting functions are themselves plausibly affected by a given treatment variable must be considered on a case-by-case basis.

[18] Although I use the terminology of "potential outcomes" to describe the structural function $h_i(x)$, the components w of x that simply serve as controls need not be manipulable or enter the function h directly.

[19] This is a feature that distinguishes my approach from the treatment of measurement error by Abrevaya and Hausman. (1999), who assume (in my notation) that $R \perp\!\!\!\perp X|H$, which amounts to $V \perp\!\!\!\perp U|H$. They also restrict the model functionally, with a linear index structure for h and scalar errors with monotonicity.

[20] See also Conti and Pudney (2011) and Montgomery (2022) for evidence of non-independence between V and gender.

2.3 Relationship to existing ordered response models

The model described above 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 often σ is normalized to 1. This fits into the general model if V_i is taken to be degenerate (all units share a value v), $\tau(0) = 0$, U_i is a scalar and $h(x, u) = x^T\beta + u$ for some $\beta \in \mathbb{R}^{d_x}$. The assumption that U is normally distributed independent of X_i implies EXOG. In the probit model, the effect on H of a switch from x to x' is common across units, given by $(x' - x)^T\beta$. The ordered probit model maintains the structural function $h(x, u) = x^T\beta + u$ and distributional assumption on U_i, but allows for a reporting function that maps to a larger set of categories $\mathcal{R} = \{0, 1, \dots \bar{R}\}$. The reporting function continues to be homogeneous across units, with thresholds $\tau(0), \tau(1), \dots, \tau(\bar{R} - 1)$.

Despite the popularity of probit and ordered 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 provided that 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 a matter of normalization.[21] However, a separable model like $h(x, u) = g(x) + u$ for potential outcomes also imposes homogeneity of treatment effects, which is often an unpalatable assumption when conducting causal inference. The Matzkin (1992) result also requires all of the X to be continuously distributed. My results allow for treatment effect heterogeneity, and nests a leading case of her identification result when regressors are continuous (see Proposition 3).

3 What is identified from continuous variation in X

Given the model outlined in the last section, let us now turn to what can be identified by looking at responses given variation in X. In this section, we suppose that at least one component of X is continuously distributed.

Let ∂_{x_j} denote a partial derivative with respect to x_j, and assume the following:

Assumption REG (regularity conditions). *The following hold:*

[21] 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 (see Appendix A for a generalization). 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.

- *component X_{ji} of X_i is continuously distributed*
- *H_i is continuously distributed conditional on X_i and V_i*
- *$\frac{\partial}{\partial x_j} Q_{H|XV}(\alpha|h, v) \leq M < \infty$ for all $\alpha \in [0, 1]$, $h \in \mathcal{H}$, where $Q_{H|X}$ is the conditional quantile function of H given X and V*
- *$f_{H,\partial_{x_j}h|XV}(h, h'|x, v)$ exists and is upper bounded by some $c(h')$ where $\int c(h')|h'|dh' < \infty$, for all $v \in \mathcal{V}$.*

I denote by $f_H(h|x, v)$ the conditional density of H_i at h, conditional on $X_i = x$ and $V_i = v$. With this notation in mind, we have the following result:

Theorem 1. *Assume HONEST and EXOG, and that REG holds for some $j \in \{1, \dots, J\}$. Then:*

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

Proof. See Appendix G. □

The inner expectation in Theorem 1 (indicated by square brackets [ ]) is over heterogeneity in causal effects U_i, while the outer expectation (indicated by curly brackets { }) is over heterogeneity V_i in reporting functions. Expanding this second expectation out, we have

$$\partial_{x_j} P(R_i \leq r|X_i = x) = -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v) \cdot \mathbb{E}\left[\partial_{x_j} h(x, U_i)|H_i = \tau_v(r), x, v\right] \tag{8}$$

where recall that for a value x, we let w denote it's final components $(x_{J+1} \dots x_{d_x})$, fixed as control variables.

Theorem 1 shows that the derivative of $P(R_i \leq r|X_i = x)$ with respect to changes in x_j provides a positively-weighted linear combination of the structural change in H due to X_j. When h is interpreted as potential outcomes, $\partial_{x_j} h(x, U_i)$ yields marginal causal effects. Under the weaker interpretation of Footnote 15, $\partial_{x_j} h(x, U_i)$ captures an average of how the conditional quantile of H varies with X_j.[22]. The proof of Theorem 1 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).[23]

By (8), the "weight" on an individual with happiness close to $\tau_v(r)$ is positive and proportional to $dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v)$. Figure 3 provides some intuition for this particular weighting. 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

[22] In particular, Theorem 1 shows that

$$\frac{\partial}{\partial x_j} P(R_i \leq r|X_i = x) = -\mathbb{E}\left\{ f_H(\tau_{V_i}(r)|x, V_i) \cdot \frac{\partial}{\partial x_j} \, Q_{H|XV}(\alpha|x, V_i)\big|_{\alpha = F_{H|XV}(\tau_{V_i}(r)|x, V_i)]} \Big| \, X_i = x\right\}$$

[23] A version of Theorem 1 that analyzes the estimand of quantile regression $\partial_{x_j} Q_{R|X}$ instead of the CDF does not admit the first of these steps, which uses of the law of iterated expectations to condition on latent reporting heterogeneity V_i.

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 can have negative effects: $\partial_{x_j} h(x, U_i) < 0$, denoted by arrows to the left in Figure 3. Others can have positive effects $\partial_{x_j} h(x, U_i) > 0$, indicated by rightward arrows in Figure 3. 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: $f_H(\tau_v(r)|x, v) > 0$. 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.

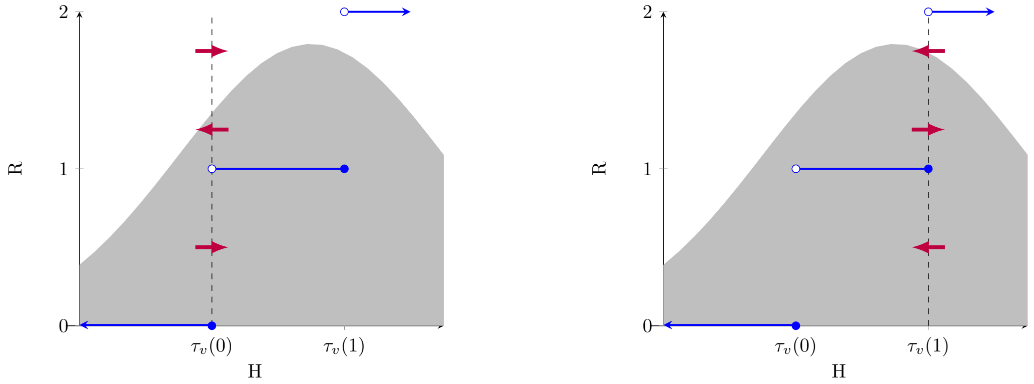


Figure 3: Intuition for Theorem 1: the derivative of $P(R_i \leq r|X_i = x)$ with respect to x_j captures the "flow" of individuals over threshold $\tau_v(r)$ due to a small change in x_j. Left: $\partial_{x_j} P(R_i \leq 1|X_i = x)$ captures flows over $\tau_v(0)$. Right: $\partial_{x_j} P(R_i \leq 1|X_i = x)$ captures flows over $\tau_v(1)$. The gray shaded curve in the background depicts the density of H_i.

Note that the marginal respondents averaged over in the RHS of Theorem 1 cannot be individually identified, since neither H_i nor $\tau_{V_i}(r)$ are observed for a given i. However, if the sign of causal effects is common across individuals, the following proposition shows that average characteristics of these marginal respondents can be identified.

Proposition 1. *Let $A_i \in \{0, 1\}$ be a binary covariate that is unaffected by X_i. Then if the sign of $\partial_{x_j} h(x, U_i)$ is the same for all individuals:*

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

Proof. See Appendix G. □

One could estimate, for example, the proportion of respondents at a particular reponse margin r that are women, and compare this to the population as a whole.[24]

[24] This result parallels identification of average complier characteristics in instrumental variables models (Abadie, 2003), with the restriction of a common effect sign playing a role analagous to the LATE monotonicity assumption.

3.1 Implications

Theorem 1 generalizes the well-known 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. In the probit model, v is degenerate and the single threshold $\tau_v(0) = 0$, while $h(x,u) = x^T\beta + u$ and $H_i|X_i = x \sim \mathcal{N}(x^T\beta, \sigma^2)$. Thus, $f_H(\tau(0)|x) = f_H(0|x) = 1/\sigma \cdot \phi(-x'\beta/\sigma) = \phi(x'\beta)$.

It is well-known that β in the probit model is only identified up to an overall scale normalization, often achieved by fixing the variance of the error distribution $\sigma^2 = 1$. Similarly, we lack from Theorem 1 the ability to pin down the overall scale of derivatives of the structural function $\partial_{x_j} h(x, U_i)$. Put another way, the weights $dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x,v)$ do not necessarily integrate to one. However, considering the ratio of *two* derivatives removes any scale-dependence of the estimand:

$$\frac{\frac{\partial}{\partial x_1} P(R_i = 1|X_i = x)}{\frac{\partial}{\partial x_2} P(R_i = 1|X_i = x)} = \frac{\mathbb{E}\left\{w_r(x,V_i) \cdot \mathbb{E}\left[\partial_{x_1} h(x,U_i)|H_i = \tau_v(0), x, v\right]|X_i = x\right\}}{\mathbb{E}\left\{w_r(x,V_i) \cdot \mathbb{E}\left[\partial_{x_2} h(x,U_i)|H_i = \tau_v(0), x, v\right]|X_i = x\right\}} \tag{9}$$

where

$$w_r(x,v) := f_H(\tau_v(r)|x,v)/\mathbb{E}[f_H(\tau_{V_i}(r)|x,V_i)|X_i = x]$$

gives a weighting function that is positive and integrates to one, i.e. $\mathbb{E}[w_r(x,V_i)|X_i = x] = 1$. To contrast this with the positive but non-normalized measure that appears in (8), I refer to weights such as those appearing in (9) as “convex”. Note that the convex weight applied to each group characterized by $H_i = \tau_v(r), X_i = x, V_i = v$ is exactly the same in both the numerator and denominator of (9).

Beyond the case of binary survey questions, it is not typical to estimate regressions of response CDF at some fixed category r, as contemplated by Theorem 1. However, the result allows us to study the more common practice of modeling the conditional mean of R_i given X_i. To see this, suppose that $\mathcal{R}$ consists of integers $\{0, 1, \ldots, \bar{R}\}$ for some $\bar{R}$. Note that the following identity holds for all i:

$$R_i = \sum_{r=1}^{\bar{R}} \mathbb{1}(r \le R_i) = \sum_{r=0}^{\bar{R}-1} \mathbb{1}(r < R_i) \tag{10}$$

From this it then follows that for any x: $\mathbb{E}[R_i|X_i = x] = \sum_{r=0}^{\bar{R}-1} P(r < R_i|X_i = x) = \bar{R} - \sum_{r=0}^{\bar{R}-1} P(R_i \le r|X_i = x)$. Then, applying Theorem 1:

$$\partial_{x_j}\mathbb{E}[R_i|X_i = x] = \int dF_{V|W}(v|w) \cdot \sum_{r=0}^{\bar{R}-1} f_H(\tau_v(r)|x,v) \cdot \mathbb{E}\left[\partial_{x_j} h(x,U_i)|H_i = \tau_v(r), x, v\right] \tag{11}$$

For brevity, I use the shorthand $\sum_r$ for the definite sum $\sum_{r=0}^{\bar{R}-1}$.

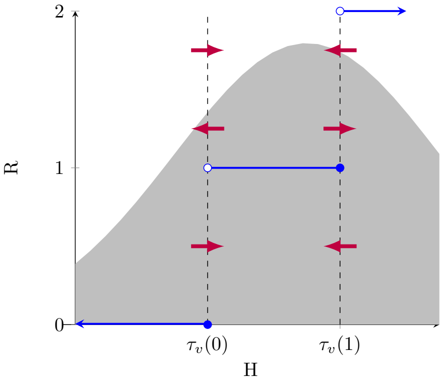


Figure 4: With $\mathcal{R}$ integers, the derivative of $\mathbb{E}[R_i \le r|X_i = x]$ with respect to x_j captures the "flow" of individuals over either threshold $\tau_v(r)$ due to a small change in x_j. Compare to left and right panels of Fig 3.

Collecting (11) across all continuous regressors, we can summarize as:

Corollary 1. *Under the assumptions of Theorem 1, if $\mathcal{R} = \{0, 1, \dots, \bar{R}\}$ then*

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

where for any function $g(x)$ we let $\nabla_x g(x) = (\partial_{x_1} g(x), \partial_{x_2} g(x), \dots)^T$ be a vector of partial derivatives with respect to whichever of the first J components of X are continuously distributed and satisfy REG.

Remark: if instead of the integers, the researcher associates alternative numerical values r_j with the ordered responses $\mathcal{R}$, where $r_0 < r_1 < \dots$, then instead of (10) 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 $f_H(\tau_v(r_j)|x, v)$ upweighted by the positive factor $(r_{j+1} - r_j)$. While different labeling schemes could be used in estimation to achieve different weightings over local causal effects, the most information one could learn is by simply repeating Theorem 1, one r at a time. When considering mean regression, using integer category labels is natural in that it weighs each threshold equally.

Another way to express Corollary 1 is as follows. Let $\tau_v := \{h : \tau_v(r) \text{ for some } r\}$ denote the set of all thresholds for individuals with reporting function v. Then

$$\nabla_x \mathbb{E}[R_i|x] = \mathbb{E}\left[w(x, V_i) \cdot \nabla_x h(x, U_i)|H_i \in \tau_{V_i}, X_i = x\right] \tag{12}$$

where $w(x, v) := \sum_r f_H(\tau_v(r)|x, v)$. This expression shows that $\nabla_x \mathbb{E}[R_i|x]$ averages over all units having $X_i = x$, located at any of their individual-specific happiness thresholds, with (positive but not convex) weights $w(X_i, V_i)$.

In the subjective well-being literature, it is common to regress R_i on multiple explanatory variables and compare coefficients. For example, Luttmer (2005) compares the regression coefficient for i's own income to that of the income of i's neighbors, finding that the latter is negative and roughly twice as large in magnitude. If the conditional expectation of R_i given X_i is in fact linear in two continuous x_1 and x_2 (see Sec. 6), then the OLS coefficient γ_1 on x_1 captures $\partial_{x_1}\mathbb{E}[R_i|X_i = x]$, and similarly $\gamma_2 = \partial_{x_2}\mathbb{E}[R_i|X_i = x]$. Accordingly, the ratio γ_1/γ_2 is equal to $\partial_{x_1}\mathbb{E}[R_i|X_i = x]/\partial_{x_2}\mathbb{E}[R_i|X_i = x]$, which by Corollary 1 can be written

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

where $\tilde{\beta}_j(x) := \mathbb{E}\left[\tilde{w}(x, V_i) \cdot \partial_{x_j} h(x, U_i)|H_i \in \tau_{V_i}, X_i = x\right]$ and

$$\tilde{w}(x, v) := \frac{w(x, V_i)}{\mathbb{E}\left[w(x, V_i)|H_i \in \tau_{V_i}, X_i = x\right]}$$

For any given x, $\tilde{\beta}_j(x)$ is a convex combination of treatment effects with respect to X_j across individuals in the population. Note that the weights appearing in the numerator and denominator are again the same for a fixed value of x, as in (9).

A special case in which (13) simplifies further when the structural function $h(x, u)$ is weakly separable between x and u, i.e.

$$h(x, u) = \mathtt{h}(g(x), u) \tag{14}$$

where some function $g : \mathcal{X} \to \mathbb{R}$ aggregates over the components in X into some scalar $g(x)$, which is then combined through $\mathtt{h}$ with heterogeneity u in a way that may or may not be additively separable. For example, a linear structural function $h(x, u) = x^T\beta + u$ sets $g(x) = X^T\beta$ and $\mathtt{h}(g, u) = g + u$, combining a linear causal response with an additive scalar error term. Weakly separable models for ordered response in which u is a scalar have been studied by Matzkin (1994). When (14) holds, Corollary 1 yields

$$\frac{\partial_{x_1}\mathbb{E}[R_i|x]}{\partial_{x_2}\mathbb{E}[R_i|x]} = \frac{\int \sum_r f_H(\tau_v(r)|x, v) \cdot \partial_{x_1} g(x) \cdot \mathbb{E}\left[\partial_{x_1}\mathtt{h}(x, U_i)|H_i = \tau_v(r), x, v\right]}{\int \sum_r f_H(\tau_v(r)|x, v) \cdot \partial_{x_2} g(x) \cdot \mathbb{E}\left[\partial_{x_2}\mathtt{h}(x, U_i)|H_i = \tau_v(r), x, v\right]} = \frac{\partial_{x_1} g(x)}{\partial_{x_2} g(x)} \tag{15}$$

where the highlighted factors cancel out in the numerator and denominator, since the derivatives of $g(x)$ do not depend on v or r.

In the still simpler case of a partially linear h function, (15) leads to the following:

Corollary 2. *If the assumptions of Theorem 1 hold for $j = 1$ and $j = 2$, and $h(x, u)$ takes the form $h(x, u) = x_1\beta_1 + x_2\beta_2 + g(x_3, \ldots x_{d_x}) + u$ (e.g. $h(x, u) = x^T\beta + u$) with $\beta_2 > 0$, then the observable conditional expectation function $\mathbb{E}[R_i|x]$ is also weakly separable, i.e. $\mathbb{E}[R_i|x] = \phi(\gamma_1 x_1 + \gamma_2 x_2, x_3 \ldots x_{d_x})$ for some function ϕ, and $\gamma_1/\gamma_2 = \beta_1/\beta_2$.*

Proof. Let $m(x) := \mathbb{E}[R_i|X_i = x]$. By (15), we have that $\partial_{x_1} m(x)/\partial_{x_2} m(x) = \beta_1/\beta_2$, independent of x. This implies that m takes the form of ϕ above. □

A convenient feature of a weakly separable model like (14) 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, (13) 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 Theorem 1 with $j = 1, 2$, we have*

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

then

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

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

Proof. See Appendix G. □

Proposition 2 requires reporting heterogeneity V_i to be conditionally orthogonal to structural function heterogeneity U_i, similar to Assumption IDR (but here also conditional on $H_i \in \tau_{V_i}$.). Further, one must be able to 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.

As a final note, we can see how Theorem 1 recovers an identification result of Matzkin (1992) for the function $g(x)$ when $h(x, u) = g(x)+u$ (i.e. causal effects are homogeneous). 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 (1992), is to suppose $g(x)$ is homogeneous of degree one. In this case, I show below that g is identified up to scale, under somewhat different assumptions than those of Matzkin (1992).

Proposition 3. *Suppose HONEST and EXOG hold, there are no controls W, and each of the $X_1 \dots X_J$ are continuously distributed satisfying REG. Suppose further that $h(x, u) = \boldsymbol{h}(g(x), u)$, 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.*

Proof. See Appendix G. Eq. (37) gives an explicit expression for $g(x)$. □

4 What is identified from discrete variation in X

Section 3 considered what was identified by examining the distribution of R over infinitesimal differences in X. Now let us instead consider any two fixed values x and x' (differing only in their first J components), and let $\Delta_i := h(x', U_i) - h(x, U_i)$ be the "treatment effect" of moving from $X_i = x$ to $X_i = x'$ for unit i. Further, let $f_H(y|\Delta, x, v)$ denote the density of H_i conditional on $\Delta_i = \Delta$, $X_i = x$ and $V_i = v$. The following expression shows what can be identified from the conditional distribution of R_i across this discrete change:

Theorem 2. *Under HONEST and EXOG:*

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

where $\bar{f}(\Delta, y, x, v) := \frac{1}{\Delta}\int_{y-\Delta}^{y} f_H(h|\Delta, x, v)dh$ *is the average density between* $y - \Delta$ *and* y*, among units with reporting function* v*, treatment effect* Δ*, and* $X_i = x$*.*[25]

Proof. See Appendix G. □

Similar to Theorem 1, Theorem 2 shows that the change in $P(R_i \leq r|x)$ over discrete changes in x can be written as a positive linear combination of the causal effect of that variation in X on H.

A similar expression to the above shows up in the "bunching design", which leverages bunching at kinks in decision-makers' choice sets for identification of behavioral elasticities. An assumption sometimes used in that literature is that $f_H(h|\Delta, x, v)$ is approximately constant for all h between $\tau_v(r) - \Delta$ and $\tau_v(r)$ (see e.g. Saez 2010; Kleven 2016; Goff 2022 for a discussion).[26] Under this assumption, Theorem 2 would simplify to:

$$\begin{aligned} &P(R_i \leq r|X = x') - P(R_i \leq r|X = x) \\ &\qquad = -\int dF_{V|W}(v|w) \cdot \int d\Delta \cdot \Delta \cdot f_H(\Delta, \tau_v(r)|X_i = x, V_i = v) \\ &\qquad = -\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v) \cdot \mathbb{E}[\Delta_i | H_i = \tau_v(r), X_i = x, V_i = v] \end{aligned} \tag{16}$$

Eq. (16) exactly recovers the weighting over individuals achieved by Theorem 1 using continuous variation in x. In particular, the quantity $\mathbb{E}[\Delta_i|H_i = \tau_v(r), X_i = x, V_i = v]$ appears above with the same weight $-dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v)$ as $\mathbb{E}[\partial_{x_j} h(x, U_i)|H_i = \tau_v(r), X_i = x, V_i = v]$ does in Eq. (8). Unfortunately, the constant density assumption used to derive the above is hard to justify except as an approximation if Δ is very small.[27]

[25] By "between $y - \Delta$ and y" I mean in the interval $[\min\{y - \Delta, y\}, \max\{y - \Delta, y\}]$, without taking a stand on the sign of Δ. Note that $\bar{f}(\Delta, y, x, v)$ is positive even if $\Delta < 0$, in which case it is equal to the average density between y and $y + |\Delta_i|$.

[26] The constant density restriction could also help motivate an identification condition used by Kaiser and Vendrik (2022) for the signs of coefficients in wellbeing regressions. See Appendix D for details.

[27] If we consider the limit $x' \to x$ with the two differing only in component j, then this approximation becomes exact and Eq. (16) applied to $(P(R_i \leq r|x') - P(R_i \leq r|x))/(x'_j - x_j)$ reduces to Theorem 1. See Lemma SMALL in Goff (2022).

Note that Theorem 2 exhausts all implications of the observable data (R_i, X_i) regarding variation in the structural functions $h(x, u)$ with respect to x, when the control variables w are held fixed. Once $P(R_i \leq r|X_i = x)$ is known for all r for some fixed reference value x of the explanatory variables, along with the marginal distribution of X_i, the only remaining information available from the joint distribution of (R_i, X_i) takes the form of differences $P(R_i \leq r|X_i = x') - P(R_i \leq r|X_i = x)$ for various values of x' and r.

As our main focus is the conditional mean of responses with an integer response scale, let us as in Eq. (11) aggregate Theorem 2 across the response categories r to obtain:

$$\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x] = \mathbb{E}\left[\sum_{r=0}^{\bar{R}-1} \bar{f}(\Delta_i, \tau_v(r), x, V_i) \cdot \Delta_i \middle| X_i = x\right] \tag{17}$$

To obtain the notation of Eq. (2), define $\bar{f}(\Delta, v, x) := \sum_{r=0}^{\bar{R}-1} \bar{f}(\Delta, \tau_v(r), x, v)$.

Recall from Theorem 1 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, 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 $\sum_r \bar{f}(\Delta_i, \tau_v(r), x, v)$. Since this quantity depends not only on the density of H at response thresholds $\tau_v(r)$ but also the density at points within a treatment effect Δ of such thresholds, the two weighting schemes do not lead to estimands that can obviously be directly compared.

Note: whether or not $\mathbb{E}[R_i|x'] - \mathbb{E}[R_i|x]$ is positive or negative does not reflect the sign of the average treatment effect: $\mathbb{E}[\Delta_i]$. Rather, it depends on how positive and negative treatment effects are aggregated over by the weights $\sum_r \bar{f}(\Delta, \tau_v(r), x, v)$. If the CDF functions (or equivalently, quantile 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 $\Delta_i > 0$.[28]. While the weights that emerge from a discrete contrast x, x' are less "local" than the ones that emerge when leveraging continuous variation in x (Section 3), they are still not uniform over the support of H: the ATE or its sign is not identified.

5 Comparing discrete and continuous regressors

Given the results of the last two sections, suppose we are now interested in comparing the magnitude of a local regression derivative to the mean difference across two discrete groups, i.e.

$$\frac{\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]}{\partial_{x_j} \mathbb{E}[R_i|X_i = x'']} \tag{18}$$

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

for some x,x', and x''. By Corollary 1 and Eq. (17), we know that this ratio is equal to

$$\frac{\mathbb{E}\left[\sum_r \bar{f}(\Delta_i, \tau_{V_i}(r), x, V_i) \cdot \Delta_i | X_i = x\right]}{\mathbb{E}\left\{\sum_r f_H(\tau_{V_i}(r)|x'', V_i) \cdot \mathbb{E}\left[\partial_{x_j} h(x'', U_i) | H_i = \tau_{V_i}(r), x'', V_i\right] \middle| X_i = x''\right\}}$$

To interpret this ratio quantitatively in terms of averages of Δ_i and $\partial_{x_2} h(x'', U_i)$, the relevant question is how similar the sum $\sum_r f_H(\tau_{V_i}(r)|x'', V_i)$ over densities at the thresholds is to the corresponding sum over averaged densities: $\sum_r \bar{f}(\Delta_i, \tau_{V_i}(r), x, V_i)$. The total "weight" placed on causal effects in the numerator, after averaging over reporting function heterogeneity V_i, is $\mathbb{E}\left[\sum_r \bar{f}(\Delta_i, \tau_{V_i}(r), x, V_i) | X_i = x\right]$. In the denominator, the total weight is $\mathbb{E}\left[\sum_r f_H(\tau_{V_i}(r)|x'', V_i) \middle| X_i = x''\right]$. If these quantities are close to one another in magnitude, then Eq. (18) uncovers something close to the ratio of two convex averages of causal effects. If they differ by an unknown amount, then interpreting (18) in terms of the relative magnitudes of causal effects is not possible.

Given the definition of $\bar{f}$, notice that $\sum_r f_H(\tau_v(r)|x'', v)$ and $\sum_r \bar{f}(\tau_v(r), \Delta, x, v)$ are similar for a given (Δ, v) if

$$\sum_r \frac{1}{\Delta} \int_{\tau_v(r)-\Delta}^{\tau_v(r)} f_H(y|\Delta, x, v) dy \approx \sum_r f_H(\tau_v(r)|x'', v) \tag{19}$$

Observe that the two sides of (19) can *only* differ because the summation occurs over H_i evaluated at the thresholds $\tau_v(r)$. If instead the sums over r were replaced by integrals over all possible values of H_i, we would have

$$\int \left\{\frac{1}{\Delta} \int_{h-\Delta}^{h} f_H(y|\Delta, x, v) dy\right\} dh = \int f_H(h|x'', v) \cdot dh$$

which holds trivially because both sides evaluate to unity, regardless of the values of Δ, v, x, and x''. This is immediate for the RHS, which integrates a density. To see it for the LHS, reverse the order of the integrals to obtain $\int dy \cdot f_H(y|\Delta, x, v) \cdot \left\{\frac{1}{\Delta} \int_y^{y+\Delta} dh\right\} = 1$.

However, we know from Section E that discrete sums over the thresholds do not correspond to equal-weighted integrals over h, even in the limit of a continuum of response categories. Rather, the integrals also involve the quantity $r'(h, v)$, which measures how responsive response function v is at h: the local "density" of thresholds $\tau_v(r)$ around h. Nevertheless, the intuition provided by the above logic suggests that looking at the dense response limit may be informative in evaluating the approximation (19). The next section does exactly this, while Section 5.2 evaluates the approximation across a variety of simulated data generating processes.

5.1 Analytical results in the dense response limit

Suppose that the response categories are "dense" in the sense described in Section E, and defined formally in Appendix G.7. From Proposition 9, we know that the derivative of $\mathbb{E}[R_i|X_i = x]$ with respect to a continuous component of x depends on the steepness of response curves $r'(h,v)$ across h. Analogously, when using discrete variation in X_i, what matters for a unit with treatment effect Δ is instead $\bar{r}'(y,\Delta,v) := \frac{1}{\Delta}\int_y^{y+\Delta} r'(h,v)dh$, the average slope of the response function $r(h,v)$ for h between y and $y+\Delta$:

Proposition 4. *Under HONEST, EXOG, and REG, then in the dense response limit*

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

Proof. See Appendix G. □

Since $\bar{r}'$ is weakly positive, we see from Proposition 4 that $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ identifies a positive aggregation of treatment effects $\Delta_i = h(x', U_i) - h(x, U_i)$. Furthermore, the weights on Δ_i aggregate up to[29]

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

By comparison, note that the total weight on causal effects in a derivative $\partial_{x_j}\mathbb{E}[R_i|X_i = x]$ are, by Proposition 9:

$$w_x := \int dF_{V|W}(v|w) \int dh \cdot f_H(h|\Delta, x, v) \cdot r'(h,v) = \bar{R} \cdot \mathbb{E}[r'(H_i, V_i)|X_i = x]$$

A comparison of w_x and $w_{x,x'}$ allows us to interpret the relative magnitudes of discrete and continuous differences in $\mathbb{E}[R_i|X_i = x]$, as in Eq. (18). 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:

$$\frac{\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]}{\partial_{x_2}\mathbb{E}[R_i|X_i = x]} \xrightarrow{R} \frac{\tilde{\beta}_1}{\tilde{\beta}_2} \cdot \frac{w_{x,x'}}{w_x} \tag{20}$$

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

Let us say that *heterogenous linear reporting* holds if with $\mathcal{R} = \{0, 1, \dots \bar{R}\}$ for some integer $\bar{R}$, we have that:

$$\tau_v(r) = \ell(v) + r \cdot \frac{\mu(v) - \ell(v)}{\bar{R}} \tag{21}$$

where $\ell(v) = \tau_v(r)$ is the threshold between the lowest and second-lowest category for an individual with $V_i = v$, and $\mu(v)$ is the threshold between the second-highest and highest category. Note that in the limit of many categories $\bar{R}$, (21) can be well approximated by the piecewise-linear reporting function $\lim_{\bar{R}\to\infty} \frac{r(h,v)}{\bar{R}} = \mathbb{1}(\ell(v) \le h \le \mu(v)) \cdot \frac{h - \ell(v)}{\mu(v) - \ell(v)}$.

Heterogeneous linear reporting may be a reasonable assumption if individuals aim to maximize the informativeness of their responses by equally spreading out the response

[29] Note that if Δ_i and $\bar{r}'(H_i, \Delta_i, V_i)$ are uncorrelated conditional on $X_i = x$, then we can further write the RHS of Proposition 4 as $\mathbb{E}[\Delta_i|X_i = x] \cdot \bar{R} \cdot \mathbb{E}[\bar{r}'(H_i, \Delta_i, V_i)|X_i = x]$.

categories (van Praag, 1991), given their subjective definitions $\ell(v)$ and $\mu(v)$ of the minimum and maximum category thresholds.[30] Kaiser and Vendrik (2022) summarize empirical evidence in support of linearity, for example from asking individuals directly about their response thresholds, or using objectively verifiable outcomes such as an individual's height.

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

Proposition 5. *Suppose that the following hold in addition to HONEST,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*,* $f_H(h|\Delta,x,v)$ *is increasing on the interval* $[\ell(v)-|\Delta|,\ell(v)+|\Delta|]$*, and decreasing on the interval* $[\mu(v)-|\Delta|,\mu(v)+|\Delta|]$

Then

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

where $w_{x,x'} = \bar{R}\cdot\mathbb{E}[\bar{r}'(H_i,\Delta_i,V_i)|X_i=x]$*,* $w_x = \bar{R}\cdot\mathbb{E}[r'(H_i,V_i)|X_i=x] \xrightarrow{R}$*, and* $w_{x'} = \bar{R}\cdot\mathbb{E}[r'(H_i,V_i)|X_i=x]$*.*

Furthermore, suppose that

$$Var\left[\left.\frac{1}{\mu(V_i)-\ell(V_i)}\right| X_i = x\right] \leq Var\left[P(0<R_i<\bar{R}|x,V_i)\middle| x\right]\cdot\mathbb{E}\left\{\left.\frac{1}{\mu(V_i)-\ell(V_i)}\right| X_i=x\right\}^2,$$

i.e. the lengths of reporting intervals are not too variable relative to variability in bunching at the endpoints 0 *and* $\bar{R}$*, then*

$$\frac{1}{2} \leq \frac{w_{x,x'}}{w_x} \leq \frac{1}{P(0<R_i<\bar{R}|X_i=x)^2},$$

Proof. See Appendix G. □

Proposition 5 provides two sets of bounds on the ratio of the total weight on causal effects in $\mathbb{E}[R_i|X_i=x'] - \mathbb{E}[R_i|X_i=x]$, to the total weight on causal effects in $\nabla_x\mathbb{E}[R_i|X_i=x]$. The first bound, $\frac{w_{x,x'}}{\frac{1}{2}(w_x+w_{x'})} \in [1,2]$ implies that, in the setup of Eq. (20):

$$\frac{\mathbb{E}[R_i|X_i=x'] - \mathbb{E}[R_i|X_i=x]}{\frac{1}{2}\partial_{x_2}\mathbb{E}[R_i|X_i=x'] + \frac{1}{2}\partial_{x_2}\mathbb{E}[R_i|X_i=x]} \xrightarrow{R} \theta\cdot\frac{\beta_1}{\beta_2} \tag{22}$$

where θ is some number between 1 and 2. This bound requires no assumptions on how variable the happiness scale lengths $\mu(V_i)-\ell(V_i)$ can be across individuals with

[30] 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.

different V_i. By contrast, the second set of bounds requires us to place an upper bound on $Var\left[\frac{1}{\mu(V_i)-\ell(V_i)}\middle| X_i = x\right]$. While its upper bound is a universal one, the upper bound can be estimated from the data by a nonparametric regression of observed bunching at the endpoints of the scale (0 and $\bar{R}$) on X_i.

Proposition 5 does not appear to generalize easily away from the heterogeneous linear reporting benchmark, which is particularly tractable. However, we can obtain a result that holds more generally (but is less informative), by observing that we may understand Equation (17) in terms of a *convolution* between the happiness random variable H_i and a fictional second random variable T_i that represents the thresholds of a given response function. Let T_i take values $\tau_v(r)$ for $r = 0 \dots \bar{R} - 1$, with equal probability $p(r) = P(T_i = \tau_v(r)) = 1/\bar{R}$ for each, and let T_i be independent of all other random variables. The conditional CDF of $T_i - H_i$ would be:[31]

$$F_{T-H|\Delta,v,x}(t) = \sum_{r=0}^{\bar{R}-1} p(r) \cdot P(\tau_v(r) - H_i \leq t|\Delta, v, x) = \frac{1}{\bar{R}} \sum_{r=0}^{\bar{R}-1} (1 - F_{H|\Delta,v,x}(\tau_v(r) - t))$$

Some algebra shows that the weight $\sum_r \bar{f}(\Delta, \tau_v(r), x, v$ appearing in Eq. (17) is given by the average density of $T - H$ between $t = 0$ and $t = \Delta$:

$$\sum_r \bar{f}(\Delta, \tau_v(r), x, v) = \bar{R} \cdot \frac{F_{T-H|\Delta,v,x}(\Delta) - F_{T-H|\Delta,v,x}(0)}{\Delta}$$

In the dense response limit, the variable T_i becomes continuously distributed with density $r'(h, v)$. To understand the resulting distribution of $T - H$, we can make use of the fact that convolution with a log concave density preserves quasiconcavity (Uhrin, 1984):

Proposition 6. *Suppose that $f_H(y|\Delta, v, x) = f_H(y|v, x)$, i.e. treatment effects are independent of H_i (conditional on X_i and V_i), where $f_H(y|\Delta, v, x)$ is log-concave, and that $r_n(h, v) \overset{R}{\to} r(h, v)$ where $r'(h, v)$ is quasiconcave in h. Then $\frac{d}{dt} F_{T-H|\Delta,v,x}(t)$ is a quasiconcave function of t.*

Quasiconcavity of $r'(h, v)$ is a reasonable assumption if reporting functions have a roughly "sigmoidal" shape, in which they are less sensitive r' is lower in the extremes and have a slope that peaks for some some intermediate value. Log-concavity of $f_H(y|\Delta, v, x)$ would hold if happiness were, for example, normally distributed.

As a quasiconcave function, the derivative of $F_{T-H|\Delta,v,x}(t)$ will under Proposition 6 be increasing up to some point $t = t^*$ and then decreasing for all values $t > t^*$. Accordingly, if $t^* < 0$, and treatment effects are positive for all individuals $\Delta \geq 0$, then individuals with larger values of Δ will receive smaller weights. Similarly, if treatment effects are negative and $t^* > 0$, then individuals with larger magnitudes of treatment effects will receive smaller weights. However, when t^* belongs to the interior of the support of treatment effects, the weights will be non-monotonic in Δ.

[31] If instead of $\mathcal{R}$ being represented by integers $0 \dots \bar{R}$, a value r_j is associated with each element in $\mathcal{R}$ as in the discussion following Eq. (11), then we replace $p(r)$ with $p(j) = (r_{j+1} - r_j)$, which must be kept inside the sum.

5.2 Simulation evidence

To gather some further suggestive evidence on the comparability of estimates that use discrete vs. continuous variation in X. I in this section simulate several data-generating-processes (DGPs) for H_i and for the response functions $r(\cdot, V_i)$. Throughout, I take the response space $\mathcal{R}$ to be a set of integers $0, 1, 2 \dots \bar{R}$, where the value of $\bar{R}$ will be varied across DGPs.

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 $w_{x,x'}$, w_x and $w_{x'}$ to understand the relative weights each of these estimands place on causal effects.

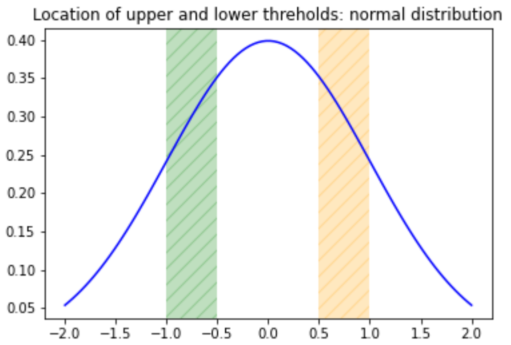
For now, I suppose heterogeneous linear reporting, so that Proposition 5 holds in the dense-response limit $\bar{R} \to \infty$. Individual reporting functions can be charaterized by $\ell(v)$, the value of happiness at which an individual with $V_i = v$ moves from response category 0 to response category 1, and $\mu(v)$, the value at which this individual would move from category $\bar{R} - 1$ to the highest category $\bar{R}$. Response functions are sampled independently of everything else, which implies $U_i \perp V_i$.

In a first set of simulations, I take H_i to have a standard normal distribution, conditional on $X_i = x$. Note that since the overall location and scale of the happiness distribution is not inherently meaningful, this choice of mean and variance is arbitrary. 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 drawn uniformly from $[0.5, 1]$. The left panel of Figure 5 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 $R = 10$, this threshold for the highest possible category is for all individuals at least above the mean level of happiness in the population.

The table on the right side of Figure 5 reports $\frac{w_{x,x'}}{\frac{1}{2}(w_x + w_{x'})}$ 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. Alternatively, the results can be interpreted as reporting conditional analogs of the quantity $\frac{w_{x,x'}}{\frac{1}{2}(w_x + w_{x'})}$ 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 $X_i = x$.

Proposition 5 implies that as $\bar{R} \to \infty$, $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ 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 5, aside from $\Delta = 1$ and $\Delta = 5$. In all but the case of $\Delta = 5$, $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ is in fact quite close to unity, well within the refined bounds $[1, 1/NB]$ which holds under the variance restriction in Proposition 5, 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 5 provides



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 5: $H_i|X_i = x$ 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 $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

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, i.e. β_1/β_2 in the case described by Equation (20). In this case, results do not differ substantially whether the number of response categories is small (e.g. $\bar{R} = 2$, the case of binary response) or e.g. $\bar{R} = 100$. Appendix Table 1 shows that results also do not differ much whether there are few or many different reporting functions present in the population.

The $\Delta = 5$ case nevertheless shows that the ratio in (20) may be quite misleading *in principle*, even with this distribution of H_i. The $\bar{R} = 2$ value of $w_{x,x'}/\frac{1}{2}(w_x + w_{x'}) \approx 0.5$ 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.

While the quantity $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ averages over the reporting heterogeneity in the population, Figure 6 disaggregates this by V_i. Define $\delta_{\Delta,x,v} := \frac{\sum_r \bar{f}(\Delta,\tau_v(r),x,v) - \sum_r f_H(\tau_v(r)|x,v)}{\sum_r f_H(\tau_v(r)|x,v)}$. An individual with $X_i = x$ and $V_i = v$ will receive similar weights when using either discrete or continuous variation at x if $\delta_{\Delta,x,v} \approx 0$. Write Eq. (17) 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 6 reports the distributions of $\frac{\sum_r \bar{f}(\Delta,\tau_v(r),x,v)}{\sum_r f_H(\tau_v(r)|x,v)} = 1 + \delta_{\Delta,x,v}$, across 1000 reporting functions sampled the same as in Figure 5. 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.

Figures 7,8 and 9 repeat the exercise of Figure 5 with alternative distributions assumed for $H_i|X_i = x$. Figure 7 first relaxes unimodality of the normal distribution by letting H_i be distributed as a mixture of two normals, leading to a "double-peaked" shape. Upper

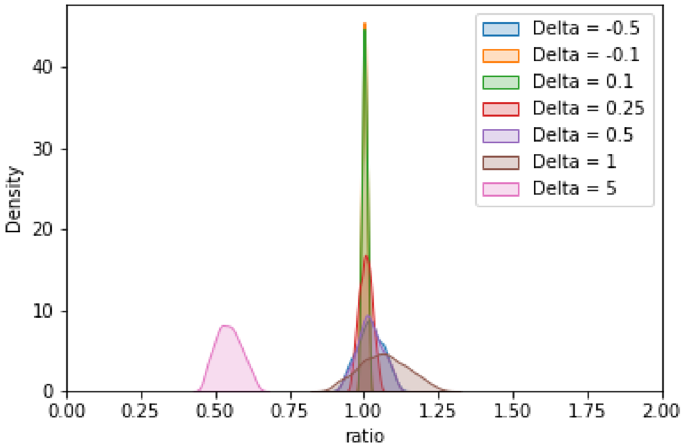
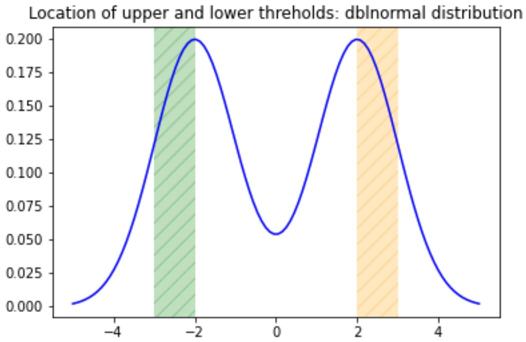


Figure 6: The distribution of $1 + \delta_{\Delta,x,V_i}$ across V_i is depicted across alternative values of Δ_i, with $H_i|X_i = x$ standard normal, $\bar{R} = 100$, and 1000 reporting functions are drawn from $\ell(v) \sim U[-1, 1/2]$, $\mu(v) \sim U[1/2, 1]$.

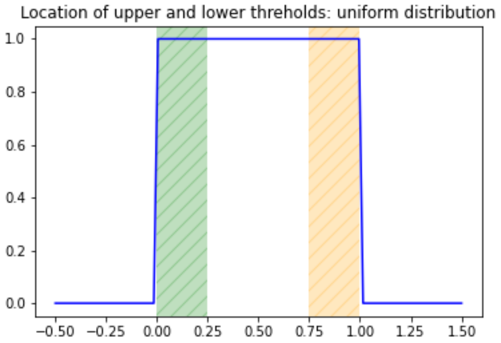
and lower thresholds μ and ℓ are sampled from the decreasing and increasing (respectively) portions of this distribution's density. The table shows that $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ is again close to unity across a wide range of treatment effect sizes, with β_1/β_2 now being overestimated in the case of an extremely large treatment effect $\Delta = 5$. Figure 16 reports the distributions of δ_{Δ,x,V_i}, as in Figure 6.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.945228	0.999726	1.002822	1.002350
-0.1	0.996596	1.000027	1.000005	1.000072
0.1	0.998344	1.000034	1.000106	1.000211
0.25	0.989510	0.999942	1.000275	1.001007
0.5	0.958623	1.000036	1.003705	1.004092
1	0.901081	0.999569	1.009504	1.013429
5	3.335567	1.361256	1.225963	1.154723
1/NB	1.470781	1.481656	1.480439	1.483049

Figure 7: $H_i|X_i = x$ is an equal mixture of $\mathcal{N}(-2, 1)$ and $\mathcal{N}(2, 1)$, and 1000 reporting functions are drawn from $\ell(v) \sim U[-3, -2]$, $\mu(v) \sim U[2, 3]$. 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 $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

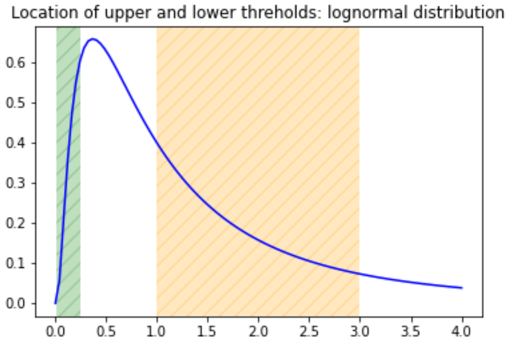
Figure 8 instead uses a uniform distribution for H_i. This allows us to sample the thresholds μ and ℓ from regions that abut the extremes of the population happiness distribution. Results here are encouraging, except in the cases where Δ moves a significant portion of the population outside of $[0, 1]$ (e.g. $|\Delta| \geq 0.5$. In such cases, there is significant non-overlap between the distributions of $H_i|X_i = x'$ and $H_i|X_i = x$). Notably, $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ is non-monotonic in the magnitude of Δ, first increasing above unity and then falling much below it, with opposing effects canceling out when $\Delta = 1$. Figure 17 reports the distributions of δ_{Δ,x,V_i}, as in Figure 6.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	1.348617	1.334445	1.332001	1.334231
-0.1	1.0	1.0	1.005277	1.010828
0.1	1.0	1.0	1.004095	1.002599
0.25	1.0	1.031953	1.031220	1.035741
0.5	1.320108	1.135912	1.101178	1.081512
1	0.999805	0.998796	0.995304	0.995515
5	0.199587	0.199893	0.199174	0.200451
1/NB	1.350946	1.361273	1.355604	1.357831

Figure 8: $H_i|X_i = x$ uniform $[0, 1]$, and 1000 reporting functions are drawn from $\ell(v) \sim U[0, 1/4]$, $\mu(v) \sim U[3/4, 1]$. 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 $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

Finally, Figure 9 introduces skewness by letting happiness have a standard log-normal distribution. Corresponding to the long right-tail in the happiness distribution, I take $\mu(V_i)$ to have support over a large range of values relative to $\ell(V_i)$. The results are less optimistic, as compared with the normally distributed case. For $|\Delta| > 0.1$, $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ differs from unity by more than 10%. However, the worst-case $\Delta = 5$ is not much worse than in the normally-distributed DGP, with $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ at least about 0.45 for all $\bar{R}$. Figure 18 reports the distributions of δ_{Δ,x,V_i}, as in Figure 6.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.684805	0.840033	0.891200	0.928110
-0.1	0.909362	0.931933	0.954190	0.979041
0.1	1.098319	1.075785	1.052181	1.026204
0.25	1.256482	1.182417	1.118010	1.066896
0.5	1.505123	1.278124	1.185238	1.120666
1	1.643653	1.258917	1.196697	1.137716
5	0.494329	0.453087	0.439618	0.452964
1/NB	1.467527	1.447327	1.460737	1.445131

Figure 9: $H_i|X_i = x$ is standard log-normal, and 1000 reporting functions are drawn from $\ell(v) \sim U[1/100, 1/4]$, $\mu(v) \sim U[1, 3]$. 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 $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

Appendix F reports further results and variations on the DGPs discussed above. Appendix Tables 2, 3 and 4 show that as with the normal DGP, results also do not differ much whether there are few or many different reporting functions present in the population. Taking the lognormal distribution of H as representing the worst-case among the distributions considered, Appendix F also considers some variations on the reporting-function DGP used above. Appendix Figure 19 allows the support of ℓ and μ to "overlap" so that the minimum threshold ℓ for some individuals is higher than that *maximum* threshold μ is for others. Appendix Figure 20 eliminates all heterogeneity in reporting functions.

Appendix Figures 21 and 22 dispense with (heterogeneously) linear reporting, instead sampling the thresholds for a given individual from a specified distribution and sorting them in ascending order to define that individual's reporting function. In all cases, results fall within the range of those presented above.

6 Implications for regression analysis

From Theorems 1 and 2, it is clear that learning from the conditional distribution of responses R_i given variation in X_i, one can uncover positive linear combinations of causal effects, but with weights that are not under the researcher's control. Rather, they depend on individuals' unobserved and heterogeneous reporting functions, and the distribution of underlying happiness H_i near the thresholds at which those individuals move between successive response categories.

One immediate implication is that if causal effects have the same sign for all individuals, this sign can be identified empirically by mean regression of responses R_i on variation in X_i, whether that variation is continuous or discrete.[32] The same-sign assumption in fact leads to over-identification restrictions, because $\partial_{x_j} P(R_i \leq r | X_i = x)$ or $P(R_i \leq r | X_i = x') - P(R_i \leq r | X_i = x)$ must have the same sign for all r.

However, researchers often want to be more ambitious and compare the magnitudes of the effects of multiple explanatory variables on H_i. The results of the preceding sections show that if $\mathbb{E}[R_i | X_i = x] = m(x_1, x_2, \ldots x_J, w)$ is modeled as a fully flexible function of the regressors and estimated nonparametrically, features of the function m can be interpreted causally: derivatives of m uncover positive weighted combinations of partial effects (Section 3) and discrete differences uncover positively-weighted combinations of treatment effects (Section 4). In general, these weights vary not only with regressor x_j but by value of the entire vector x, making interpretation somewhat tedious.

Although nonparametric approaches allow one to estimate the entire function $m(x)$ consistently, it is difficult to report and interpret an infinite-dimensional object, and the curse of dimensionality looms large with several X. One path forward for a continuous X_1 is to estimate and report the average of $\partial_y \mathbb{E}[R_i | X_{1i} = y, X_{-1,i}]$ over the distributions of $y = X_{1i}$ and of the other regressors $X_{-1,i}$. For e.g. a binary regressor X_2, one could instead report the average difference $\mathbb{E}[R_i | X_{-2,i}, X_{2i} = 1] - \mathbb{E}[R_i | X_{-2,i}, X_{2i} = 0]$ over the distribution of the other regressors $X_{-2,i}$. Such averages can be estimated at the $\sqrt{n}$ rate (Ichimura and Todd, 2007), and their ratios can still be interpreted in terms of ratios of convex averages of causal effects as in Section 5—the averaging is now over x as well. In Appendix B, I follow this approach using the estimator of Li and Racine (2004) (which is implemented in the Stata command `npregress kernel`) to synthetic data with two explanatory variables. This estimator applies kernel regression techniques to setups in

[32] When the goal is not causal inference but understanding the joint distribution of H_i and X_i, we have from Corollary 3 that if the sign of $\partial_{x_j} Q_{H|X}(\alpha|x)$ is the same for all α, this sign will be reflected in $\partial_{x_j} \mathbb{E}[R_i|x]$. Analogously, with discrete variation in X_i, if the conditional distribution of H_i given $X_i = x'$ stochastically dominates that of $X_i = x$, then this will be reflected in the sign of the observable conditional mean difference $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$.

which there may both be continuous and discrete regressors.

Notwithstanding the above, in practice researchers often instead estimate parsimonious specifications of the function m, most frequently applying OLS to linear models of the form:

$$R_i = \gamma_1 X_{1i} + \gamma_2 X_{2i} + \cdots + \gamma_J X_{Ji} + \lambda^T W_i + \epsilon_i \tag{23}$$

where the vector of control variables W includes a constant. The remainder of this section studies the interpretation of the estimands γ_j in Eq. (23) in light of the results of the proceeding sections.

6.1 Case 1: linear model is correctly specified

The most straightforward case arises when Eq. (23) is correctly specified in the sense that the conditional expectation function is in fact linear in the x, i.e.

$$\mathbb{E}[R_i|X_i = x] = \gamma_1 x_1 + \cdots + \gamma_J x_J + \lambda^T w \tag{24}$$

or equivalently that $\mathbb{E}[\epsilon_i|X_i = x] = 0$ in (23). It should be emphasized that a linear model for causal effects: $h(x, u) = x^T\beta + u$, does *not* imply that a linear relationship holds between R_i and X_i, given non-linearity in the response functions. However, whether or not $\mathbb{E}[R_i|X_i = x]$ exhibits a linear functional form can be examined empirically, given that R_i and X_i are both observable.

In the context of Eq. (24), consider comparing the regression coefficient γ_1 with γ_2, if X_1 and X_2 are both continuously distributed. Since each γ_j is then equal to $\partial_{x_j}\mathbb{E}[R_i|X_i = x]$, the ratio γ_1/γ_2 recovers a ratio of two convex averages of causal effects by Eq. (13). If in addition to Eq. (24), the structural function is linear with $h(x, u) = x^T\beta + u$, then $\gamma_1/\gamma_2 = \beta_1/\beta_2$.

Now consider comparing the coefficients for a continuously distributed X_2 (e.g. income) and a binary X_1 (e.g. an indicator for being married). For any values $x_2, x_3 \dots x_{d_x}$, note that:

$$\frac{\gamma_1}{\gamma_2} = \frac{\mathbb{E}[R_i|X_{1i} = 1, X_{2i} = x_2, \dots X_{d_x i} = x_{d_x}] - \mathbb{E}[R_i|X_{1i} = 0, X_{2i} = x_2, \dots X_{d_x i} = x_{d_x}]}{\frac{1}{2}\partial_{x_2}\mathbb{E}[R_i|X_{1i} = 1, X_{2i} = x_2, \dots X_{d_x i} = x_{d_x}] + \frac{1}{2}\partial_{x_2}\mathbb{E}[R_i|X_{1i} = 0, X_{2i} = x_2, \dots X_{d_x i} = x_{d_x}]}$$

If a linear model again holds *both* for $\mathbb{E}[R_i|X_i = x]$ and for the structural function $h(x, u) = x^T\beta + u$, then under the assumptions of Proposition 5, Eq. (22) with $x' = (1, x_2, \dots)$ and $x = (0, x_2, \dots)$ implies that $\frac{\gamma_1}{\gamma_2}$ estimates $\frac{\beta_1}{\beta_2}$ up to a scaling factor θ that lies between 1 and 2. More generally, if linearity holds only for $\mathbb{E}[R_i|X_i = x]$ but not necessarily for $h(x, u)$, and the assumptions of Proposition 5 are still satisfied, then γ_1/γ_2 identifies a ratio of two weighted averages of causal effects, again up to a factor $\theta \in [1, 2]$, where the weights aggregate to one both the numerator and the denominator. In the numerator, the averaging is over $\Delta_i = h(x', U_i) - h(x, U_i)$ among units with $X_i = x$ while in the denominator it is over both $\mathbb{E}[\partial_{x_2}h(x, U_i)|H_i \in \tau_{V_i}, X_i = x]$ and $\mathbb{E}[\partial_{x_2}h(x', U_i)|H_i \in \tau_{V_i}, X_i = x']$, cf. Eq. (12).

Finally, suppose that we wish to compare regression coefficients for two discrete variables X_1 and X_2. For simplicity, suppose that they are both binary. Then, :

$$\frac{\gamma_1}{\gamma_2} = \frac{\mathbb{E}[R_i|X_{1i}=1, X_{2i}=x_2, \ldots X_{d_x i}=x_{d_x}] - \mathbb{E}[R_i|X_{1i}=0, X_{2i}=x_2, \ldots X_{d_x i}=x_{d_x}]}{\mathbb{E}[R_i|X_{1i}=x_1, X_{2i}=1, \ldots X_{d_x i}=x_{d_x}] - \mathbb{E}[R_i|X_{1i}=x_1, X_{2i}=0, \ldots X_{d_x i}=x_{d_x}]}$$

for any $x = (x_1, x_2, \ldots, x_{dx})$. To analyze this case we can apply Proposition 5 twice while using a continuously distributed third variable X_3 as a common comparison. This implies that under the assumptions of Proposition 5, in the dense response limit γ_1/γ_2 identifies a ratio of two weighted averages of causal effects (with respect to x_1 in the numerator, and x_2 in the denominator) up to a factor that lies between 1/2 and 2.[33]

6.2 Case 2: misspecified regression function

When Eq. (23) is misspecified, the estimands of γ_j in (23) remain well-defined as population linear projection coefficients, but these do not always bear a straightforward relationship to the features of the conditional expectation function $m(x) = \mathbb{E}[R_i|X_i = x]$ of interest. Nevertheless, some existing results on linear regression are useful to gain some intuition.

One case in which the estimand of Eq. (23) remains causally interpretable without assuming linearity of the expectation (24) occurs when we have a single continuously distributed X_i and no control variables beyond a constant. In this case, Eq. (23) amounts to simple linear regression: $R_i = \gamma_0 + \gamma_1 X_i + \epsilon_i$. Yitzhaki (1996) shows that the regression coefficient $\gamma_1 = \frac{Cov(R,X)}{Var(X)}$ can then be written as a weighted average over the local derivative of $\mathbb{E}[R_i|X_i = x]$ even if it is non-linear:

$$\gamma_1 = \int w(x) \cdot \frac{d}{dx}\mathbb{E}[R_i|X_i = x] \cdot dx$$

where $w(x) := \frac{1}{Var(X)} \int_{-\infty}^{x} f_X(t)(t - \mathbb{E}[X_i])dt$ is a positive function that integrates to unity, with f_X denoting the density of X_i. By Theorem 1, γ_1 thus still captures a positively weighted combination of causal effects $\partial_x h(x, U_i)$, where the averaging is now also over x. If all units in the population have the same sign of $\partial_x h(x, U_i)$, then this sign can be recovered as that of γ_1. Angrist and Pischke (2008) extend the above expression to a case with covariates: if $\mathbb{E}[X_1|W_i]$ is linear in W_i, then γ_1 can be written as $\mathbb{E}[\gamma_1(W_i)]$, where the quantities that define $\gamma_1(w)$ are analogous to the above but condition on $W_i = w$, with weights again integrating to unity.[34] An analogous expression can also be derived for a setting with a binary X_1 (Angrist, 1998; Angrist and Pischke, 2008) or an ordered X_1 (Angrist and Krueger, 1999). Thus with a single treatment variable X_1 of any type, a

[33] To see this, let $\gamma_3 = \partial_{x_3}\mathbb{E}[R_i|x]$, and write $\gamma_1/\gamma_2 = \gamma_1/\gamma_3 \cdot \gamma_3/\gamma_2$. Let $\tilde{\beta}_1$, $\tilde{\beta}_2$, and $\tilde{\beta}_3$ denote the convex combinations of causal effects associated with γ_1, γ_2 and γ_3 (cf Propositions 1 and 2 after normalizing the weights). By Proposition 5 $\gamma_1/\gamma_3 \overset{R}{\to} \theta_1 \cdot \tilde{\beta}_1/\tilde{\beta}_3$ and $\gamma_2/\gamma_3 \overset{R}{\to} \theta_2 \cdot \tilde{\beta}_2/\tilde{\beta}_3$, where $\theta_1, \theta_2 \in [1, 2]$. Thus, $\gamma_1/\gamma_2 \overset{R}{\to} \theta_1/\theta_2 \cdot \tilde{\beta}_1/\tilde{\beta}_2$. Note that the regression X_3 need not actually be observed.

[34] The linearity assumption is not restrictive if W_i consists of indicators for an exhaustive set of covariate cells, a so-called "fully-saturated" regression.

linear regression equation (23) with fully saturated controls simply re-averages the causal effects of X_1 on H_i derived in this paper over a second set of positive weights.

Unfortunately, the results mentioned above do not carry over to the general setting with multiple treatment variables $X_1 \dots X_J$ and controls estimated by Eq. (23). This type of regression is common the happiness literature, with OLS coefficients γ_j and γ_k compared quantitatively to assess the relative contributions of two or more factors (see e.g. Luttmer 2005). Goldsmith-Pinkham et al. (2022) show that regressions like (23) with controls W can be subject to "contamination bias", in which the coefficient γ_1 on X_1 includes not only effects from X_1, but also effects from the other treatments $X_2 \dots X_J$. In other words, γ_1 does not cleanly separate variation in X_1 from variation in the other treatments.

Since Goldsmith-Pinkham et al. (2022) consider a standard setup in which the outcome variable of interest is directly observed, we can facilitate the connection by phrasing our examination of regression (23) in terms of the causal effects of X_i on R_i (rather than on H_i).[35] This interpretation is justified under assumption EXOG, because we can define potential outcomes $R(x)$ with respect to $x \in \mathcal{X}$ as $R_i(x) = R(H(x, U_i), V_i)$ with $\{R_i(x) \perp\!\!\!\perp X_{ji}\}|W_i$ for $j = 1 \dots J$. Unfortunately, contamination bias is possible even under fairly optimistic linearity assumptions, for example that $\mathbb{E}[R_i(x) - R_i(x_0)|W_i = w] = x'\beta(w) + \lambda' w$ for some fixed reference treatment $x_0 \in \mathcal{X}$, and vectors $\beta(w)$ and λ, i.e. conditional-on-W average treatment effects are linear in all treatment variables. If the per-unit conditional effects $\beta(w)$ vary with w, then e.g. the estimand γ_1 may not capture a clean average of the $\beta_1(w)$ but instead include a second term that depends on the $\beta_2 \dots \beta_J$. The threat of contamination bias is not in any way specific to the use of subjective outcome variables, but may be particularly pernicious in this context given the motivation to compare magnitudes across regressors. Goldsmith-Pinkham et al. (2022) provide detail on possible solutions.

7 Conclusion

This paper has investigated identification 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. Nevertheless, researchers may be willing to suppose that individual responses reflect the value of a well-defined latent variable H_i for each individual i.

This paper has shown that without observing H_i and without assuming possibility of interpersonal comparisons of H, the conditional distribution of responses given exogenous covariates X can still be informative about the effects of X on H. While this allows for testing the sign of causal effects when this sign is common across individuals, different

[35] Note that given any consistent estimator for the average treatment effect of some covariate contrast x, x' (differing only in the first J components) on R, one can interpret this using the methods of the present paper by translating it back into a statement about conditional means, since $\mathbb{E}[R_i(x') - R_i(x)] = \mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$.

conditional mean comparisons impose different weighting systems over the causal effects of individuals in the population. I have provided both simulation evidence and theoretical results that suggest the impact of this problem is somewhat limited in practice. Nevertheless, the results suggest that caution is warranted in comparing the magnitude of regression coefficients across explanatory variables, even when they are as good as randomly assigned.

References

ABADIE, A. (2003). "Semiparametric instrumental variable estimation of treatment response models". *Journal of Econometrics* 113 (2), pp. 231–263. URL: https://www.sciencedirect.com/science/article/pii/S0304407602002014.

ABREVAYA, J. and HAUSMAN., J. (1999). "Semiparametric Estimation with Mismeasured Dependent Variables: An Application to Duration Models for Unemployment Spells". *Annales d'Economie et de Statistique* (55-56).

ALLEN, R. and REHBECK, J. (2019). "Identification With Additively Separable Heterogeneity". *Econometrica* 87 (3), pp. 1021–1054. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA15867.

ALLISON, R. and FOSTER, J. E. (2004). "Measuring health inequality using qualitative data". *Journal of Health Economics* 23 (3), pp. 505–524. URL: https://www.sciencedirect.com/science/article/pii/S0167629604000232.

ANGRIST, J. and KRUEGER, A. (1999). "Empirical strategies in labor economics". *Handbook of Labor Economics*. Ed. by O. ASHENFELTER and D. CARD. 1st ed. Vol. 3, Part A. Elsevier. Chap. 23, pp. 1277–1366. URL: https://EconPapers.repec.org/RePEc:eee:labchp:3-23.

ANGRIST, J. D. (1998). "Estimating the Labor Market Impact of Voluntary Military Service Using Social Security Data on Military Applicants". *Econometrica* 66 (2), pp. 249–288. URL: http://www.jstor.org/stable/2998558 (visited on 12/05/2022).

ANGRIST, J. D. and PISCHKE, J.-S. (2008). *Mostly Harmless Econometrics*. Princeton, NJ: Princeton University Press.

BANKS, W. P. and COLEMAN, M. J. (1981). "Two subjective scales of number". *Perception and Psychophysics* 29, pp. 95–105.

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. P. (2018). *The econometrics of happiness: Are we underestimating the returns to education and income?* URL: https://arxiv.org/abs/1807.11835.

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. URL: https://www.aeaweb.org/articles?id=10.1257/aer.104.11.3498.

BINMORE, K. (Jan. 2009). "Interpersonal Comparison of Utility". *Kincaid, H. and Ross, D., (eds.) The Oxford Handbook of Philosophy of Economics. Oxford Handbooks in Philosophy . Oxford University Press, New York, US, pp. 540-559. ISBN 9780195189254.*

BOND, T. N. and LANG, K. (2019). "The Sad Truth about Happiness Scales". *Journal of Political Economy* 127 (4), pp. 1629–1640. URL: https://doi.org/10.1086/701679.

BREUNIG, C. and MARTIN, S. (2020). *Nonclassical Measurement Error in the Outcome Variable.* URL: https://arxiv.org/abs/2009.12665.

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. URL: https://www.aeaweb.org/articles?id=10.1257/aer.102.6.2981.

CHEN, L.-Y., OPARINA, E., POWDTHAVEE, N. and SRISUMA, S. (2022). "Robust Ranking of Happiness Outcomes: A Median Regression Perspective". *Journal of Economic Behavior and Organization* 200, pp. 672–686. URL: https://www.sciencedirect.com/science/article/pii/S0167268122002062.

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. URL: https://www.sciencedirect.com/science/article/pii/S0304407618302379.

CHERNOZHUKOV, V. and HANSEN, C. (2005). "An IV Model of Quantile Treatment Effects". *Econometrica* 73 (1), pp. 245–261. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00570.x.

CHESHER, A., ROSEN, A. and SIDDIQUE, Z. (2022). *Estimating Endogenous Effects on Ordinal Outcomes.*

CONTI, G. and PUDNEY, S. (2011). “Survey Design and the Analysis of Satisfaction”. *The Review of Economics and Statistics* 93 (3), pp. 1087–1093. URL: `http://www.jstor.org/stable/23016097` (visited on 11/20/2023).

COWELL, F. A. and FLACHAIRE, E. (2017). “Inequality with Ordinal Data”. *Economica* 84 (334), pp. 290–321. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1111/ecca.12232`.

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. URL: `https://www.sciencedirect.com/science/article/pii/S0047272718300409`.

DWYER, R. J. and DUNN, E. W. (2022). “Wealth redistribution promotes happiness”. *Proceedings of the National Academy of Sciences* 119 (46), e2211123119. URL: `https://www.pnas.org/doi/abs/10.1073/pnas.2211123119`.

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. URL: `http://www.jstor.org/stable/40664510` (visited on 12/22/2022).

FLEMING, M. (1952). “A Cardinal Concept of Welfare”. *The Quarterly Journal of Economics* 66 (3), pp. 366–384. URL: `http://www.jstor.org/stable/1885309` (visited on 12/13/2023).

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

GOFF, L. (2022). *Treatment Effects in Bunching Designs: The Impact of the Federal Overtime Rule on Hours.* URL: `https://arxiv.org/abs/2205.10310`.

GOLDSMITH-PINKHAM, P., HULL, P. and KOLESÁR, M. (2022). *Contamination Bias in Linear Regressions.* Working Paper 30108. National Bureau of Economic Research. URL: `http://www.nber.org/papers/w30108`.

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

HAMERMESH, D. S. (2004). “Subjective Outcomes in Economics”. *Southern Economic Journal* 71 (1), pp. 1–11. URL: `https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2325-8012.2004.tb00619.x`.

HARSANYI, J. C. (1955). “Cardinal Welfare, Individualistic Ethics, and Interpersonal Comparisons of Utility”. *Journal of Political Economy* 63 (4), pp. 309–321. URL: http://www.jstor.org/stable/1827128 (visited on 12/13/2023).

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. and MAMMEN, E. (2007). “Identification of Marginal Effects in Nonseparable Models without Monotonicity”. *Econometrica* 75 (5), pp. 1513–1518. URL: http://www.jstor.org/stable/4502038 (visited on 09/06/2022).

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.* URL: https://arxiv.org/abs/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. URL: https://www.sciencedirect.com/science/article/pii/S0304407607002436.

HU, Y. and SCHENNACH, S. M. (2008). “Instrumental Variable Treatment of Nonclassical Measurement Error Models”. *Econometrica* 76 (1), pp. 195–216. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.0012-9682.2008.00823.x.

ICHIMURA, H. and TODD, P. (2007). “Implementing Nonparametric and Semiparametric Estimators”. *Handbook of Econometrics.* Ed. by J. HECKMAN and E. LEAMER. 1st ed. Vol. 6B. Elsevier. Chap. 74.

IMBENS, G. W. and NEWEY, W. K. (2009). “Identification and Estimation of Triangular Simultaneous Equations Models Without Additivity”. *Econometrica* 77 (5), pp. 1481–1512. URL: https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA7108.

KAISER, C. (2022). “Using memories to assess the intrapersonal comparability of well-being reports”. *Journal of Economic Behavior and Organization* 193, pp. 410–442.

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

KAPLAN, D. M. and ZHAO, W. (Nov. 2022). "Comparing latent inequality with ordinal data". *The Econometrics Journal*. utac030. URL: https://doi.org/10.1093/ectj/utac030.

KAPTEYN, A., SMITH, J. P. and VAN SOEST, A. (2013). "Are Americans Really Less Happy with Their Incomes?" *Review of Income and Wealth* 59 (1), pp. 44–65. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1475-4991.2012.00532.x.

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. URL: https://www.sciencedirect.com/science/article/pii/S0304407621000452.

KING, G., MURRAY, C. J. L., SALOMON, J. A. and TANDON, A. (2004). "Enhancing the Validity and Cross-Cultural Comparability of Measurement in Survey Research". *American Political Science Review* 196, pp. 65–78.

KLEVEN, H. J. (2016). "Bunching". *Annual Review of Economics* 8 (1), pp. 435–464. URL: https://doi.org/10.1146/annurev-economics-080315-015234.

LAYARD, R., MAYRAZ, G. and NICKELL, S. (2008). "The marginal utility of income". *Journal of Public Economics* 92 (8). Special Issue: Happiness and Public Economics, pp. 1846–1857. URL: https://www.sciencedirect.com/science/article/pii/S0047272708000248.

LI, Q. and RACINE, J. (2004). "CROSS-VALIDATED LOCAL LINEAR NONPARAMETRIC REGRESSION". *Statistica Sinica* 14 (2), pp. 485–512. URL: http://www.jstor.org/stable/24307205 (visited on 12/20/2022).

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. URL: https://doi.org/10.1093/restud/rdaa006.

LIU, S. and NETZER, N. (2023). "Happy Times: Measuring Happiness Using Response Times". *American Economic Review* 113 (12), pp. 3289–3322. URL: https://www.aeaweb.org/articles?id=10.1257/aer.20211051.

LUTTMER, E. F. P. (Aug. 2005). “Neighbors as Negatives: Relative Earnings and Well-Being*”. *The Quarterly Journal of Economics* 120 (3), pp. 963–1002. URL: `https://doi.org/10.1093/qje/120.3.963`.

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. URL: `http://www.jstor.org/stable/2951596` (visited on 12/07/2022).

— (1994). “Chapter 42 Restrictions of economic theory in nonparametric methods”. Vol. 4. Handbook of Econometrics. Elsevier, pp. 2523–2558. URL: `https://www.sciencedirect.com/science/article/pii/S157344120580011X`.

MILGROM, P. and SHANNON, C. (1994). “Monotone Comparative Statics”. *Econometrica* 62 (1), pp. 157–180. URL: `http://www.jstor.org/stable/2951479` (visited on 09/16/2022).

MOLINA, T. (2017). “Adjusting for heterogeneous response thresholds in cross-country comparisons of self-reported health”. *The Journal of the Economics of Ageing* 10, pp. 1–20.

MONTGOMERY, M. (2022). “Reversing the gender gap in happiness”. *Journal of Economic Behavior and Organization* 196, pp. 65–78.

NADAI, M. D. and LEWBEL, A. (2016). “Nonparametric errors in variables models with measurement errors on both sides of the equation”. *Journal of Econometrics* 191, pp. 19–32.

OPARINA, E. and SRISUMA, S. (2022). “Analyzing Subjective Well-Being Data with Misclassification”. *Journal of Business & Economic Statistics* 40 (2), pp. 730–743. URL: `https://doi.org/10.1080/07350015.2020.1865169`.

OSWALD, A. J. (2008). “On the curvature of the reporting function from objective reality to subjective feelings”. *Economics Letters* 100 (3), pp. 369–372. URL: `https://www.sciencedirect.com/science/article/pii/S0165176508000645`.

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. URL: `https://www.aeaweb.org/articles?id=10.1257/aer.20160256`.

SAEZ, E. (2010). “Do Taxpayers Bunch at Kink Points?” *American Economic Journal: Economic Policy* 2 (3), pp. 180–212. URL: `https://www.aeaweb.org/articles?id=10.1257/pol.2.3.180`.

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. URL: http://www.jstor.org/stable/23427520 (visited on 02/02/2023).

SCHRÖDER, C. and YITZHAKI, S. (2017). "Revisiting the evidence for cardinal treatment of ordinal variables". *European Economic Review* 92, pp. 337–358. URL: https://www.sciencedirect.com/science/article/pii/S0014292117300016.

UHRIN, B. (1984). "Some Remarks About the Convolution of Unimodal Functions". *The Annals of Probability* 12 (2), pp. 640 –645. URL: https://doi.org/10.1214/aop/1176993312.

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. URL: https://www.sciencedirect.com/science/article/pii/030440769190090Z.

YITZHAKI, S. (1996). "On Using Linear Regressions in Welfare Economics". *Journal of Business and Economic Statistics* 14 (4), pp. 478–486. URL: http://www.jstor.org/stable/1392256 (visited on 10/07/2022).

Appendices

A Relaxing and testing reporting function invariance

This section relaxes the implicit assumption from the main text that reporting behavior V_i is fixed for each individual and therefore unaffected by variation in X_i. In particular, I show that Assumption EXOG is compatible with reporting functions depending directly on observables, in a limited way. I then discusses how even the weakest version of this assumption still leads to testable implications when homogeneity assumptions are placed on causal effects.

To formalize the idea of reporting function invariance, introduce counterfactual notation V_i^x to represent the reporting function that would occur for individual i if $X_i = x$. In this notation, the *actual* reporting function for this individual is $V_i^{X_i}$.[36] The following assumption says that components $1 \dots J$ of X are excludable from the reporting function, so that only W_i can enter directly:

[36] This counterfactual notation is equivalent to instead treating V_i as fixed for an individual and letting X_i enter directly into the reporting function: $R_i = r(H_i, X_i, V_i)$.

Assumption EXCLUSION (full reporting function invariance). *For all i, $V_i^x = V_i^{x'}$ for any x and x' that differ only in components $1 \dots J$.*

Given EXCLUSION, we may let $V_i = V_i^{W_i}$ and proceed with Assumption EXOG as stated above. However, EXLUSION is stronger than necessary for my main results, and can be relaxed along similar lines to the "rank similarity" assumption of Chernozhukov and Hansen (2005):

Assumption INVARIANT (invariant reporting functions in distribution). *Conditional on $W_i = w$, $V^x \sim V^{x'}$ for any x and x' that differ only in components $1 \dots J$ and for which the remaining components equal w. Also, in addition to the second item of EXOG we have: $\{X_{ji} \perp\!\!\!\perp V_i^x\} \mid W_i = w$ for all w and x consistent with w.*

Given INVARIANT, we can proceed the definition $V_i = V_i^{X_i}$, and Assumption EXOG now follows.

A.1 Testing reporting-function invariance in separable models

Given either EXCLUSION or INVARIANT, it is plausible to make Assumption EXOG under randomization or selection-on-observables type variation in X_i. A violation of the "exclusion restriction" that X_i does not enter into an individual's reporting function $r_i(\cdot)$ would threaten the first condition of Assumption EXOG that $\{V_i \perp\!\!\!\perp X_{ji}\}|W_i$. This condition has testable implications, when additional structure is assumed on the causal response function $h(x, u)$.

In particular, consider the weak separability condition Eq. (14) considered in Section 3, that $h(x, u) = \mathtt{h}(g(x), u)$ for some function $\mathtt{h}$. Then:

$$\frac{\partial_{x_1} P(R_i \le r|x)}{\partial_{x_2} P(R_i \le r|x)} = \frac{\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v) \cdot \mathbb{E}\left[\partial_{x_1} h(x, U_i)|H_i = \tau_v(r), x, v\right]}{\int dF_{V|W}(v|w) \cdot f_H(\tau_v(r)|x, v) \cdot \mathbb{E}\left[\partial_{x_2} h(x, U_i)|H_i = \tau_v(r), x, v\right]} = \frac{\partial_{x_1} g(x)}{\partial_{x_2} g(x)} \tag{25}$$

where the first equation generalizes (9) for arbitrary values r and the second equality uses the weak-separability condition and generalizes (13) to hold for the CDF of responses rather than the mean. Importantly, the expression $\frac{\partial_{x_1} g(x)}{\partial_{x_2} g(x)}$ does not depend on r, leading to a set of overidentification restrictions when there are multiple thresholds (the number of response categories is 3 or greater).

This restriction can be leveraged to construct a test for $\{V_i \perp\!\!\!\perp X_{ji}\}|W_i$, with $h(x, u) = \mathtt{h}(g(x), u)$, HONEST, and $\{X_{ji} \perp\!\!\!\perp U_i\} \mid (W_i, V_i)$ (the second component of EXOG) as maintained assumptions. Some algebra shows, using Assumption HONEST (see proof of

Theorem 1), that:

$$\begin{aligned}\frac{\partial}{\partial_{x_j}}P(R_i \leq r|x) &= \frac{\partial}{\partial_{x_j}}\int P(H_i \leq \tau_v(r)|X_i = x, V_i = v)\cdot dF(v|x) \\ &= \int \left\{\frac{\partial}{\partial_{x_j}}P(H_i \leq \tau_v(r)|X_i = x, V_i = v)\right\}\cdot dF(v|x) \\ &\qquad + \int P(H_i \leq \tau_v(r)|X_i = x, V_i = v)\cdot \frac{\partial}{\partial_{x_j}}\{dF(v|x)\}\end{aligned} \tag{26}$$

The first term above evaluates to the quantity in Theorem 1 while the second term may be nonzero if V_i is correlated with X_{ji}.

Instead of Eq. (25) which assumed EXOG, we now have using (26)

$$\frac{\partial_{x_1}P(R_i \leq r|x)}{\partial_{x_2}P(R_i \leq r|x)} = \frac{\partial_{x_1}g(x) + \frac{\int P(H_i \leq \tau_v(r)|x,v)\cdot\{\partial_{x_1}F_{V|X}(v|x)\}}{\int f_H(\tau_v(r)|x,v)\cdot dF_{V|X}(v|x)}}{\partial_{x_2}g(x) + \frac{\int P(H_i \leq \tau_v(r)|x,v)\cdot\{\partial_{x_2}F_{V|X}(v|x)\}}{\int f_H(\tau_v(r)|x,v)\cdot dF_{V|X}(v|x)}} \tag{27}$$

where the second term in both the numerator and the denominator depend on r through the quantity $\tau_v(r)$, highlighted. Under the maintained assumptions, the only way that $\frac{\partial_{x_1}P(R_i \leq r|x)}{\partial_{x_2}P(R_i \leq r|x)}$ can vary by r is through a failure of $\{V_i \perp\!\!\!\perp X_{ji}\}|W_i$. If we further assume linearity of the structural function $g(x) = x^T\beta$, then we obtain additional overidentification restrictions that we can use with (27). In particular, $\partial_{x_1}P(R_i \leq r|x)/\partial_{x_2}P(R_i \leq r|x) = \beta_1/\beta_2$ should not depend on x, if $\{V_i \perp\!\!\!\perp X_{ji}\}|W_i$ holds.

Additional indirect tests for reporting function invariance can be found in the literature. For example, Luttmer (2005) compares life satisfaction to other outcome measures often associated with well-being, such as depression and open disagreements within the household. Seeing effects in the same direction, he concludes that his main results are not likely to driven by individuals changing their "definition" of happiness with X_i.

Eq. (26) can also be used to study the nature of the bias that occurs when the implication $V \perp\!\!\!\perp X|W$ of EXOG fails. Using integration by parts, the second term of (26) can be rewritten as

$$\begin{aligned}&\int P(H_i \leq \tau_v(r)|X_i = x, V_i = v)\cdot\frac{\partial}{\partial_{x_j}}\{dF(v|x)\} \qquad (28)\\ &\qquad = (-1)^{d_v}\int\left\{\frac{\partial}{\partial_{x_j}}F(v|x)\right\}\cdot\left\{\partial_{v_1,v_2,\ldots v_{d_v}}P(H_i \leq \tau_v(r)|X_i = x, V_i = v)\right\}\cdot dv_1\ldots dv_{d_v}\end{aligned}$$

provided that $\frac{\partial}{\partial_{\tilde{v}_1}}\frac{\partial}{\partial_{\tilde{v}_2}}\ldots\frac{\partial}{\partial_{\tilde{v}_M}}\left\{\frac{\partial}{\partial_{x_j}}F(v|x)\right\}$ vanishes on the boundary of $\mathcal{V}$, for any subset $\tilde{v}_1\ldots\tilde{v}_M$ of the components of V.

Expression (28) will be positive if, for example, V is a scalar independent of U (conditional on X), higher values of v are represent more "optimistic" reporting functions (that is, lower thresholds $\tau_v(r)$), and X_j is positively correlated with V (so that $F(v|x)$ decreases as x_j is increased).[37] As a simple example, suppose heterogeneity in repoorting functions is scalar and takes the form as an additive shift in all thresholds between individuals: $\tau_v(r) = \tau(r) - v$. Individuals with high V are more "optimistic reporters", since they require lower values of H to report a given response r. If furthermore $U \perp\!\!\!\perp V|X$, then (28) reduces to:

$$\begin{aligned}\partial_{x_j} P(R_i \leq r|x) &= \text{causal term} \\ &\quad - \int \left\{ \frac{\partial}{\partial_{x_j}} F(v|x) \right\} \cdot \{\partial_v P(H_i \leq \tau(r) - v|x)\} \cdot dv \\ &= \text{causal term} - \int f_H(\tau(r) - v|x) \cdot \left\{ -\frac{\partial}{\partial_{x_j}} F(v|x) \right\} \cdot dv\end{aligned}$$

The second term reflects a positively-weighted integral over the term in brackets, which measures the correlation between x_j and "reporting optimism" v. If X_j and V are positively correlated, then the second term above in $\partial_{x_j} P(R_i \leq r|x)$ will be positive, meaning that the observable relationship between X_j and R will be biased upwards by a positive non-causal term. If V_j and X were instead negatively correlated in this example, the bias would be in the other direction.[38]

B An illustrative example

Consider a population in which happiness is determined as by three things: i) one's income X_{1i} (measured in thousands of dollars); ii) whether they are married $X_{2i} \in \{0, 1\}$; and iii) an idiosyncratic error term U_i, according to a linear causal relationship:

$$H_i = \beta_1 \ln(X_{1i}) + \beta_2 X_{2i} + U_i \tag{29}$$

In this world money does not buy happiness: in fact, it has a slight negative effect with $\beta_1 = -0.1$. However, marriage does come with a substantial benefit to happiness: $\beta_2 = 1$.

Life satisfaction is measured by a binary question in which $R_i = 1$ indicates that individual i responded "yes" and $R_i = 0$ that they responded "no" to the question: "All things considered, are you satisfied with your life at present?" Without loss of generality,

[37] It is in principle possible for this bias term to be negative even if X_j is associated with more optimistic reporting functions: if U and V are correlated in such a way that conditional on X that those with more optimistic reporting functions tend to be less happy (this is difficult, but not impossible, to have happen while $\{X_j \perp\!\!\!\perp V\}|W$).

[38] Note that if $f_H(\tau(r) - v|x)$ and $\frac{\partial}{\partial_{x_j}} F(v|x)$ are "uncorrelated" over v in the sense that $\int \{f_H(\tau(r) - v|x) - \int f_H(\tau(r) - v'|x) \cdot dv'\} \cdot \left\{ \frac{\partial}{\partial_{x_j}} F(v|x) - \int \frac{\partial}{\partial_{x_j}} F(v'|x) \cdot dv' \right\} \cdot dv = 0$, then the density integrates to one and the non-causal term above becomes

$$\int f_H(\tau(r) - v|x) \cdot \left\{ -\frac{\partial}{\partial_{x_j}} F(v|x) \right\} \cdot dv = -\frac{\partial}{\partial_{x_j}} \mathbb{E}[V_i|X_i = x]$$

i.e. the bias from a failure of independence between X_j and reporting optimism V is simply the rate at which the mean of reporting optimism varies with X_j.

we know by Lemma 1 that individual reporting functions can be written

$$R_i = \mathbb{1}(H_i > \tau_{V_i})$$

Suppose that V_i takes two values in the population. Optimistic Reporters, indicated by $V_i = 1$ have a threshold $\tau_1 = -1$, and Pessimistic Reporters, indicated by $V_i = 0$, have $\tau_0 = 0$. While a Pessimistic Reporter requires H_i to be positive to indicate they are satisfied with life, Optimistic Reporters only require $H_i > -1$ to report that they are satisfied with life.

In line with Assumption EXOG, we will eventually assume that $(U_i, V_i) \perp\!\!\!\perp (X_{1i}, X_{2i})$, i.e. income and marital status are as good as randomly assigned. However, to investigate departures from this assumption, I introduce a parameter ρ that governs the correlation between income and "reporting optimism" V_i. In particular, the probability of being an Optimistic Reporter as a function of income is: $\mathbb{E}[V_i|X_{1i} = y] = \Phi(\rho \cdot \ln(y/50))$, where Φ is the normal CDF function. Thus if $\rho > 0$ the proportion of Optimistic Reporters is increasing with income: all among the richest are are, while none among the poorest are Optimistic Reporters.

I round out the remaining aspects of the DGP as follows:

- The distribution of income is log-normal: $ln(X_{1i}/50) \sim \mathcal{N}(0, 1)$ trimmed to the range $[20, 200]$, with X_{1i} in thousands of dollars. Trimming incomes below 20 avoids H_i tending towards infinity as $X_1 \downarrow 0$.
- Half of all individuals are married $\mathbb{E}[X_{2i}] = 0.5$, with $X_{2i} \perp\!\!\!\perp (U_i, V_i, X_{1i})$
- $U_i \sim \mathcal{N}(0, 1)$, and $U_i \perp\!\!\!\perp (V_i, X_i)$

Figure 10 shows kernel density estimates of the resulting distribution of H_i in the population, computed from a sample of $N = 10,000$. The threshold for Pessimistic Reporters $\tau_0 = 0$ (blue, dashed vertical line) and for Optimistic reporters $\tau_1 = -1$ (orange, dash-dot vertical line) fall close to the center of the distribution.

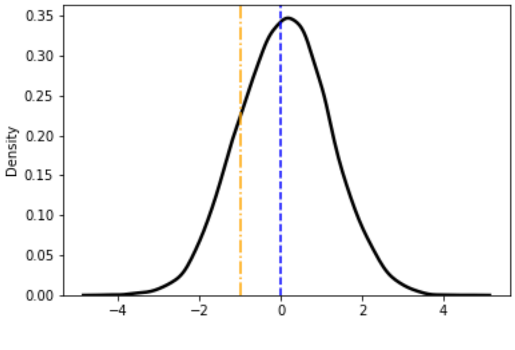
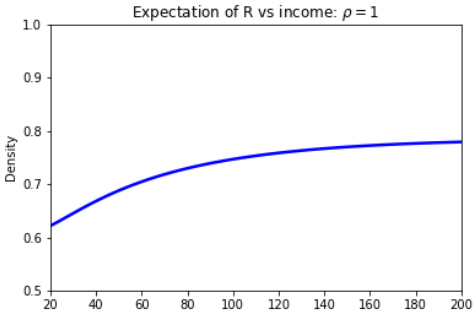


Figure 10: Distribution of H_i in the illustrative example. Vertical lines indicate the life satisfaction thresholds $\tau_0 = 0$ (blue, dashed) and $\tau_1 = -1$ (orange, dash-dot) for Pessimistic and Optimistic Reporters, respectively.

To illustrate the importance of Assumption EXOG, let us first consider ourselves in the shoes of an econometrician facing a DGP with $\rho = 1$. Since Optimistic Reporters tend to have higher incomes, this introduces a mechanical positive correlation between income and reported well-being, depicted in the left panel of Figure 11.[39] A regression of R on $\ln(X_1)$ and X_2 picks up this spurious correlation between R and X_1 that arises from reporting heterogeneity: the coefficient on log income reported in Column (1) is positive, despite $\beta < 0$. The ratio of estimates $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$ evaluates to 4.04, having opposite sign as the true value of $\beta_2/\beta_1 = -10$. Column (3) shows that if the econometrician did have access to direct observations of H, a simple OLS regression estimates β_1 and β_2, and hence their ratio, well–in line with Eq. (29).



	(1) R	(2) R	(3) H
Log Income	0.0765***		-0.106***
	(0.00699)		(0.0161)
Income		0.00155***	
		(0.000191)	
Married	0.309***	0.310***	1.015***
	(0.00861)	(0.00864)	(0.0199)
Constant	0.232***	0.698***	0.0162
	(0.0295)	(0.00482)	(0.0672)
$\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$	4.04	3.69	-9.55
Estimator	OLS	Local Linear	OLS
N	10000	10000	10000

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

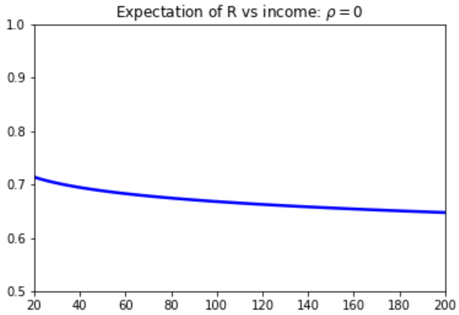
Figure 11: $\rho = 1$ case. Left panel depicts the conditional expectation function $\mathbb{E}\left\{\mathbb{E}[R_i|X_{1i} = y, X_{2i}]\right\}$ as a function of y (calculated from the known DGP as described in footnote 39), when $\rho = 1$. Regression results (right panel) of R_i on X_i reflect this spurious positive association between income and reported satisfaction, estimated on a simulated dataset of $10,000$ observations. Column (1) uses OLS of R on log-income and marriage, while Column (2) nonparametrically estimates the mean marginal effect of income and the mean effect of Marital (see text for details). Column (3) reports an (infeasible) direct regression of H_i on log-income and marriage, recovering consistent estimates of the true parameters $\beta_1 = -0.1$ and $\beta_2 = 1$.

Column (2) of Figure 11 shows that $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$ getting the wrong sign in Column (1) is not due to functional form misspecification in the OLS estimates. A nonparametric regression of R on income and marital status again captures a positive ratio, and of similar magnitude. Specifically, Column (2) reports the average derivative of $\mathbb{E}[R_i|X_{1i} = y, X_{2i}]$ with respect to y over the distribution of X_{1i} as the "coefficient" for income, and estimates the average difference $\mathbb{E}[R_i|X_{1i}, X_{2i} = 1] - \mathbb{E}[R_i|X_{1i}, X_{2i} = 0]$ as the "coefficient" for marital status. This is implemented using the kernel estimator of Li and Racine (2004), with bandwidth chosen by cross-validation. Standard errors are calculated using 500 bootstrap replications. I report $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$ com-

[39] This correlation can be computed explicitly: by the law of iterated expectations, we have that

$$\begin{aligned}\mathbb{E}[R_i|X_{1i} = y, X_{2i} = m] &= P(V_i = 0|X_{1i} = y) \cdot P(\beta_1 ln(y) + \beta_2 m + U_i > \tau_0) + P(V_i = 1|X_{1i} = y) \cdot P(\beta_1 ln(y) + \beta_2 m + U_i > \tau_1)\\ &= \Phi(\rho \cdot \ln(y/50)) \cdot \Phi(\beta_1 ln(y) + \beta_2 m - \tau_1) + (1 - \Phi(\rho \cdot \ln(y/50))) \cdot \Phi(\beta_1 ln(y) + \beta_2 m - \tau_0)\end{aligned}$$

puted by averaging the local ratio of effects across the empirical distribution of X_i: $\frac{1}{N}\sum_{i=1}^{N}\frac{\mathbb{E}[R_i|X_{1i},X_{2i}=1]-\mathbb{E}[R_i|X_{1i},X_{2i}=0]}{\partial_y \mathbb{E}[R_i|y,X_{2i}]|_{y=X_{1i}}}$.



	(1)	(2)	(3)
	R	R	R
Log Income	-0.0386***		
	(0.00726)		
Income		-0.000522***	-0.000520***
		(0.000103)	(0.000102)
Married	0.305***	0.306***	0.306***
	(0.00879)	(0.00892)	(0.00879)
Constant	0.688***	0.684***	0.567***
	(0.0304)	(0.00483)	(0.0100)
$\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$	-7.91	-12.13	-11.87
Estimator	OLS	Local Linear	OLS
N	10000	10000	10000

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 12: $\rho = 0$ case. Left panel depicts the conditional expectation function $\mathbb{E}\{\mathbb{E}[R_i|X_{1i} = y, X_{2i}]\}$ as a function of y as a function of y (calculated from the known DGP as described in footnote 39), when $\rho = 0$. Now assumption EXOG is satisfied and the observable relationship between R and income reflects sign of the true negative effect β_1. Right panel reports regression results of R_i on X_i on a simulated dataset of $10,000$ observations. Column (1) uses OLS on log income and married, and Column (2) again uses the nonparametric estimator described in the text for Figure 11. Column (3) compares this against OLS using income rather than the log of income as a regressor. For Column (3) $\hat{\beta}_{LogIncome}$ is computed as $\hat{\beta}_{Income} \cdot \hat{\mathbb{E}}[1/X_{1i}]$).

Figure 12 turns to the case of $\rho = 0$, in which Assumption EXOG is satisfied and hence the results of this paper apply. In the left panel, we see that the conditional expectation of R with respect to income is now decreasing in income, in line with the negative sign on β_1. The OLS regression of reported satisfaction on log income and marriage now yields $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome} = -7.91$, which has the correct sign but undershoots the true value of -10. This could be due to $w_{x,x'}/w_x < 1$, in the parlance of Section 5, but also could arise from misspecification of the functional form of $\mathbb{E}[R|X_{1i}, X_{2i}]$. Column (2) again implements the local linear regression method described above, returning estimates $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome} = -12.13$. These estimates are very similar to those obtained by an OLS regression of R on income (not in logs) as well as marital status. This underscores the fact that functional form assumptions regarding the effects of X on H to not translate unchanged into features of the observable correlations between X and R. In this case the causal relationship is linear in log income, while the observable one is linear in income. This distinction matters quantitatively in this case for assessing the relative contributions of income and marriage to well-being.

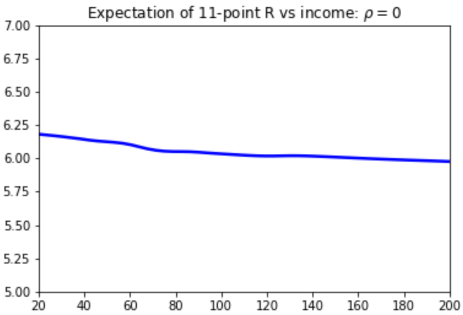
While the two DGPs reported above consider a binary R for simplicity, Figure 13 reports the $\rho = 0$ case with an 11-point scale for R. The DGP is unchanged from above except that now Pessimistic Reporters have linear reporting functions with

$$\tau_0(r) = -5 + r$$

while Optimistic Reporters have all thresholds shifted down by 1 relative to the Pessimists:

$$\tau_1(r) = -6 + r$$

Figure 13 again compares a linear regression of R on log-income and marital status (1) to a nonparametric (2) and linear regression (3) of R on income and marital status. In Column (1), the estimated ratio $\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$ is close in magnitude to -10 while the estimated ratios in Columns (2) and (3) are somewhat larger. This suggests that the Column (1) estimate being close to the truth is a coincidence of functional-form misspecification offsetting $w_{x,x'}/w_x > 1$ in line with Theorem 5. Indeed, comparing Columns (2) and (3) the CEF of R appears to again be approximately linear in income.



	(1) R	(2) R	(3) R
Log Income	-0.108*** (0.0194)		
Income		-0.00144*** (0.000264)	-0.00145*** (0.000271)
Married	1.021*** (0.0230)	1.021*** (0.0221)	1.021*** (0.0230)
Constant	6.018*** (0.0808)	6.094*** (0.0129)	5.682*** (0.0252)
$\hat{\beta}_{Married}/\hat{\beta}_{LogIncome}$	-9.50	-14.35	-14.25
Estimator	OLS	Local Linear	OLS
N	10000	10000	10000

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 13: $\rho = 0$ case. Left panel depicts a lowess regression of R on income, in the simulated dataset of $10,000$ observations.

C Extensions of the model

C.1 Using instrumental variables for identification

Suppose for that rather than making Assumption EXOG, we instead have a set of instrumental variables Z_i such that $\{Z_i \perp\!\!\!\perp (\eta_i, U_i, V_i)\}|W_i$. We assume each X_j for $j = 1 \dots J$ is continuously distributed, and Z_i contains a continuously distributed instrument corresponding tos each X_j, i.e.

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

Finally, for each $j = 1 \dots J$, suppose that $x_j(z, \eta)$ is strictly increasing in η_j. Let $\eta_i = (\eta_{1i}, \eta_{2i}, \dots \eta_{Ji})^T$.

Assumption INSTRUMENT (conditional independence of instruments).

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

Without loss of generality, we may take $\eta_{ji} \sim U[0,1]$ by redefining $\eta_{ji} = F_{X_j|Z,W}(X_{ji}|Z_i, W_i)$. The following result from Imbens and Newey (2009) implies that we can use η_i as a control variable in W_i, in the sense that

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

Thus, if η_i is simply included in the vector W_i to begin with, all of the results of the paper hold under the weaker assumption of INSTRUMENT rather than EXOG.

Theorem. *In the IV model from the last slide, with HONEST and REG (conditional on η):*

$$\nabla P(R_i \le r|X_i = x, \eta_i = \eta) = -\int dF_{V|\eta}(v|\eta) \cdot f_H(\tau_v(r)|x, v, \eta) \\ \cdot \mathbb{E}\left[\nabla_x h(x, U_i)|H_i = \tau_v(r), x, v, \eta\right]$$

Thus for $\mathcal{R} = \{0, 1, 2, \dots\}$

$$\frac{\frac{\partial}{\partial x_1} P(R_i \le r|x)}{\frac{\partial}{\partial x_2} P(R_i \le r|x)} = \frac{\int dF_{V|\eta}(v|\eta) \sum_r \cdot f_H(\tau_v(r)|x, v, \eta) \cdot \mathbb{E}\left[\partial x_1 h(x, U_i)|H_i = \tau_v(r), x, v, \eta\right]}{\int dF_{V|\eta}(v|\eta) \sum_r \cdot f_H(\tau_v(r)|x, v, \eta) \cdot \mathbb{E}\left[\partial x_2 h(x, U_i)|H_i = \tau_v(r), x, v, \eta\right]}$$

In the separable linear model, we can again identify β_1/β_2.

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 $\tilde{H}$ is a weakly increasing function of H, where H is an objectively-defined variable of ultimate interest to the researcher.

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

$$R_i = \tilde{r}_i(\tilde{H}_i) = \tilde{r}(\tilde{H}_i, S_i) \tag{30}$$

$$\tilde{H}_i = \tilde{h}_i(H_i) = \tilde{h}(H_i, T_i) \tag{31}$$

$$H_i = h_i(X_i) = h(X_i, U_i) \tag{32}$$

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 $r(\cdot, v) = \tilde{r}(\tilde{h}(\cdot, t), s)$ HONEST 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 accomodate 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. (15), despite individuals using $\tilde{H}$ rather than H to answer the survey question.

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. (3) 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. It will be useful to write

$$\frac{d}{dx_j} P(R_i \le r | X_i = x) = \int \int_{T(r)} \frac{d}{dx} f_H(h_1, h_2 | x) \cdot dh_1 dh_2 \tag{33}$$

where $T(r)$ is the set of (h_1, h_2) such that $r(h_1, h_2) \le r$. 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) show that a quantity like $\frac{d}{dx} f_H(h_1, h_2|x)$ can be rewritten as:

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

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 K:

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

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 $K \ge 1$ and again allowing reporting-function heterogeneity (satisfying EXOG), and multiple treatment variables, Eq. (33) becomes

$$\frac{d}{dx_j} P(R_i \le r | X_i = x) = \int dF_{V|W}(v|w) \int_{T_v(r)} \frac{d}{dx_j} f_H(\mathbf{h}|x) \cdot d\mathbf{h} \tag{34}$$

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

An application of the divergence theorem allows us to rewrite Eq. (33) 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 14 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 $T_v(r)$. Provided that

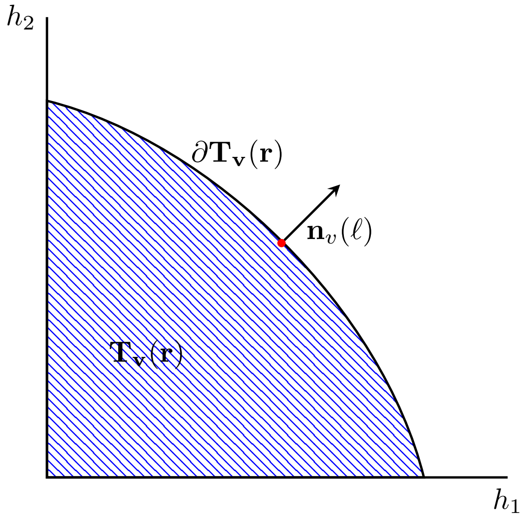


Figure 14: Components of $\hat{n}(\ell)$ are positive, by monotonicity of $h(h_1, h_2, v)$ w.r.t h_1 and h_2.

$r(\mathbf{h}, v)$ 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 14.

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

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

Suppose for the moment 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 any latent variable H_k, the effect of X_1 on

H_k is γ times as large as the effect of X_2 on H_k. Then:

$$\frac{\frac{d}{dx_1}P(R_i \le r|X_i = x)}{\frac{d}{dx_2}P(R_i \le r|X_i = x)} = \frac{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{11}\cdot \hat{n}_{v,1}(\ell) + \beta_{12}\cdot \hat{n}_{v,2}(\ell)\}\cdot d\ell\middle| X_i = x\right]}{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{21}\cdot \hat{n}_{v,1}(\ell) + \beta_{22}\cdot \hat{n}_{v,2}(\ell)\}\cdot d\ell\middle| X_i = x\right]}$$
$$= \frac{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\gamma\beta_{21}\cdot \hat{n}_{v,1}(\ell) + \gamma\beta_{22}\cdot \hat{n}_{v,2}(\ell)\}\cdot d\ell\middle| X_i = x\right]}{\mathbb{E}\left[\int_{\partial T_{V_i}(r)} \{\beta_{21}\cdot \hat{n}_{v,1}(\ell) + \beta_{22}\cdot \hat{n}_{v,2}(\ell)\}\cdot d\ell\middle| X_i = x\right]} = \gamma$$

D Idiosyncratic reporting

The main results in the paper assume both parts of Assumption EXOG: $\{X_{ji} \perp\!\!\!\perp V_i\} \mid W_i$ and $\{X_{ji} \perp\!\!\!\perp U_i\} \mid (W_i, V_i)$. These are both natural when there is idiosyncratic variation in X_i arising from an experiment or natural experiment, and reporting functions V_i are unaffected by X_i. However, if causal inference is not the researcher's goal, and the researcher simply wishes to document features of the joint distribution of H_i and X_i, we can let the function h simply represent the conditional quantile function of H as in Eq. Footnote 15 (with the definitions $U_i = (\theta_i, V_i)^T$ and $\theta_i := F_{H|XV}(H_i|X_i, V_i)$). In this case, the latter condition of EXOG holds automatically, since $\theta_i|(X_i, V_i) \sim Unif[0, 1]$, for all (X_i, V_i). Thus, to learn about the joint distribution of H_i and X_i, we only need to assume the first part of EXOG: that X_j is conditionally independent of reporting heterogeneity V (and not that it is independent of U and V *jointly*). In this case all results from the body of the paper still hold.

A stronger assumption that may be attractive in these contexts is that reporting heterogeneity V_i, rather than X_j, varies "idiosyncratically". I.e., we might assume:

Assumption IDR (idiosyncratic reporting). $\{V_i \perp\!\!\!\perp (U_i, X_{ji})\} \mid W_i$ *for each* $j = 1 \dots J$.

Assumption IDR may be an attractive alternative to Assumption EXOG introduced in Section 2.2, though neither assumption nests the other. EXOG aligns more with cases in which there is "selection-on-observables": Eq. (7) that $\{X_j \perp\!\!\!\perp (U, V)\}|W$ may follow naturally in settings in which the researcher has already argued for $\{X_j \perp\!\!\!\perp U\}|W$. Furthermore, EXOG allows U and V to be arbitrarily correlated, unlike IDR.

IDR leads to some alternative identification results to the ones in the body of this paper, for establishing features of the joint distribution of H and X. To this end, we need not make reference to any structural function $h(x, u)$ for happiness, and can take IDR as saying simply that $\{V_i \perp\!\!\!\perp (H_i, X_{ji})\} \mid W_i$. Note that this implication and IDR as stated above are equivalent under the mapping Eq. Footnote 15 that defines $h(x, u)$ as a conditional quantile function, without any causal interpretation.

The following result replaces the assumption of EXOG in Theorem 1 by IDR:

Corollary 3. *Under HONEST, IDR, and REG:*

$$\nabla_x \mathbb{E}[R_i|x] = \int dF_{V|W}(v|w) \cdot \sum_{r \in \mathcal{R}} f_H(\tau_v(r)|x) \cdot \nabla_x \, Q_{H|X}(\alpha|x)\big|_{\alpha = F_{H|X}(\tau_v(r)|x)}$$

Proof. The first steps of the proof to Theorem 1 applies $\{X_{ji} \perp\!\!\!\perp V_i\} \mid W_i$, HONEST and REG to obtain

$$\nabla_x \mathbb{E}[R_i|x] = \int dF_{V|W}(v|w) \cdot \sum_{r \in \mathcal{R}} f_H(\tau_v(r)|x, v) \cdot \nabla_x \, Q_{H|XV}(\alpha|x, v)\big|_{\alpha = F_{H|XV}(\tau_v(r)|x,v)}$$

Apply $\{V_i \perp\!\!\!\perp (H_i, X_{ji})\} \mid W_i$ again to arrive at the result. □

Another result that makes the alternative Assumption IDR rather than EXOG, but allows for discrete variation in X:

Proposition 7. *Given HONEST, IDR, and* $f_{\tau_V}(h) = \frac{d}{dh}P(h \leq \tau_{V_i}(r))$ *exists, then for any* $r \in \mathcal{R}$ *and* x, x' *that differ only in the first* J *components:*

$$P(R_i \leq r|x') - P(R_i \leq r|x) = \int_h \left\{F_{H|X}(h|x') - F_{H|X}(h|x)\right\} \cdot f_{\tau_V}(h)$$

Proof. See Appendix G. □

One consequence of Proposition 7 is that if $F_{H|X=x'}$ first order stochastically dominates $F_{H|X=x}$, i.e. that $F_{H|X}(h|x') \leq F_{H|X}(h|x)$ for all h, then under IDR this will be reflected in $F_{R|X=x'}$ first order stochastically dominating $F_{R|X=x}$. That is, the idiosyncratic reporting function transformations preserve this ranking of conditional distributions, in aggregate. This generalizes results found in Schröder and Yitzhaki (2017), Bond and Lang (2019) and Kaiser and Vendrik (2022), which assume a common reporting function across individuals. Note that the existence of $\frac{d}{dh}P(h \leq \tau_{V_i}(r))$ requires that for any response r and happiness level h, there individuals in the population with thresholds for r very close to h.

An alternative to Proposition 7 considers the conditional mean rather than the conditional CDF of R_i:

Proposition 8. *Given HONEST and IDR, for* x, x' *that differ only in the first* J *components:*

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

where $\bar{r}'_{x',x}(u) := \int dF_{V|W}(v|w) \cdot \frac{r(Q_{H|X=x'}(u),v) - r(Q_{H|X=x}(u),v)}{Q_{H|X=x'}(u) - Q_{H|X=x}(u)}$.

Proof. See Appendix G. To prove it quickly in the common reporting function case of Eq. (**??**), note that left-continuity of r implies that $Q_{r(H)|X=x'}(u) = r(Q_{H|X=x'}(u))$ (Hosseini, 2010). Then use that $\mathbb{E}[R|X = x] = \int_0^1 Q_{R|X=x}(u) \cdot du$. □

Note that Proposition 8 provides a generalization of the expression

$$\mathbb{E}[H_i|X_i = x'] - \mathbb{E}[H_i|X_i = x] = \int_0^1 \left\{ Q_{H|X=x'}(u) - Q_{H|X=x}(u) \right\} du,$$

which reveals how an (infeasible) comparison of means of H_i between x and x' aggregates over conditional quantile differences. It also generalizes Eq. (**??**) to the case of heterogeneous reporting functions.

To finish this section, I connect Proposition 8 to an identification result of Kaiser and Vendrik (2022) (henceforth KV) for the case of a common reporting function. This returns discussion to the full model of the paper in which h represents a structural/causal relationship between H and X, though the results can also be interpreted with $h(x, u)$ reflecting the quantiles of $H|X$.

It is clear from the expression in Proposition 8 that if the distribution $H|X = x'$ stochastically dominates $H|X = x$, or vice versa, then the direction of this stochastic dominance will be reflected in $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$. However, stochastic dominance between $H|X = x'$ and $H|X = x$ cannot be verified empirically, since H is unobserved. KV introduce an identification condition that implies that the signs of components of β in a linear structural model $h(x, u) = x^T\beta + u$ are identified. The first portion of the condition is that $R|X = x'$ stochastically dominates $R|X = x$ or vice versa, which can be empirically verified since R is observable.

The second part of the condition, which they call Assumption A2, is that we can write $H_i = f(R_i) + \xi$ for some increasing function f such that $E[\xi_i|R_i, X_i] = \mathbb{E}[\xi_i|R_i]$. As a leading sufficient condition for A2, they consider that $\mathbb{E}[H_i|R_i = r, X_i = x] = \mathbb{E}[H_i|R_i = r]$ in which case we may take $f(r)$ to be the unknown function $\mathbb{E}[H_i|R_i = r]$. This condition holds if happiness among individuals reporting a given category r is mean independent of observables X. In general this may be hard to motivate on theoretical grounds, since X can be expected to shift the distribution of H between adjacent thresholds $\tau(r + 1) - \tau(r)$. KV implement a clever empirical test using data in which respondents rated life satisfaction *both* on 100 point and 10-point scales, and are nevertheless unable to reject A2 for most of the X considered.

The "constant density" approximation introduced in Section 4 represents one case in which we might expect $\mathbb{E}[H_i|R_i = r, X_i = x] = \mathbb{E}[H_i|R_i = r]$ to hold, at least for interior categories r. Recall that this approximation treats $f_H(h|\Delta, x, v)$ as constant for all h between $\tau_v(r) - \Delta$ and $\tau_v(r)$. Given that the linear model considered by KV implies homogeneous Δ_i, we can drop the conditioning on Δ_i in the density. In line with a common reporting function, let us also drop V_i for the moment. If the density of $H_i|X_i$ is constant within each interval $[\tau(r), \tau(r + 1)]$, A2 would be satisfied for interior response categories r with $f(r)$ defined as $\mathbb{E}[H_i|R_i = r]$, since then $\mathbb{E}[H_i|R_i = r, X_i = x] = \int_{\tau(r-1)}^{\tau(r)} f_H(h|x)dx = \frac{\tau(r)-\tau(r-1)}{2}$, which depends on r but not on x. Now with reporting heterogeneity and EXOG, this becomes $\mathbb{E}[(\tau_{V_i}(r) - \tau_{V_i}(r-1))/2]$, which still does not depend on x.

Nevertheless, the constant density assumption remains hard to motivate unless treatment effects Δ_i are small relative to the scale over which the conditional density of H has significant curvature. And to integrate to unity, it cannot be constant everywhere on it's support, suggesting that KV's A2 requires separate justification and the highest and lowest categories.

E What would be identified with a smooth reporting function

It is informative to compare the implications of Theorem 1 to what would be identified if H_i were itself directly observable in the data. 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 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 15 depicts two examples of reporting functions on this continuum of responses. While the example on the left side of Figure 15 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 HONEST can only fail to exist only at isolated points in $\mathcal{H}$.[40] Provided that H_i is continuously distributed, it therefore does not affect results to treat $r'(h, v)$ as defined for all h.

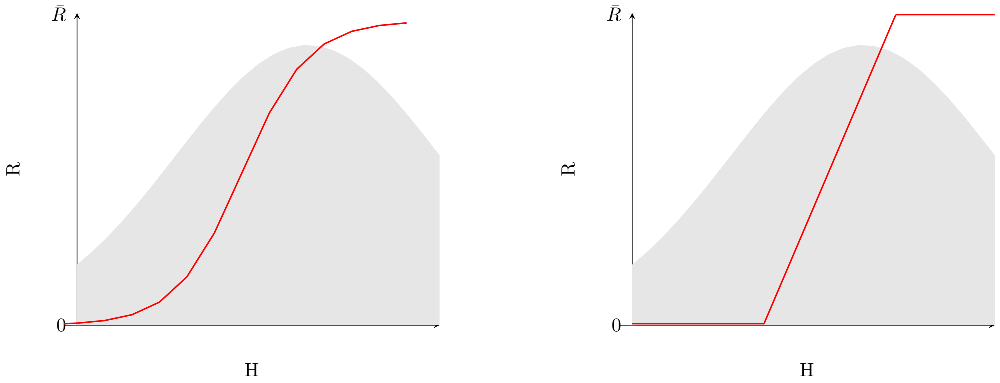


Figure 15: Example of two "continuous" reporting functions, with the density of H depicted in gray.

With "smooth reporting", we have the following analog of Theorem 1:

Proposition 9. *Assume HONEST, EXOG and REG for at least one j, with $\mathcal{R}$ a convex*

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

subset of $\mathbb{R}$. *Then:*

$$\nabla_x \mathbb{E}[R_i|X_i = x] = \int dF_{V|W}(v|w) \int dh \cdot r'(h,v) \cdot f_H(h|x,v) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|h,x,v\right] \tag{35}$$

provided the "boundary condition": $\lim_{h\to\pm\infty} f_H(h|x,v)\mathbb{E}\left[\frac{\partial}{\partial x_j}h(x,U_i)|H_i = h,x,v\right] = 0$, *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.*

Proof. See Appendix G □

The proof of Proposition 9 makes uses 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 Theorem 1.

We can compare this expression to what would be recovered by the infeasible regression of H_i on X_i, i.e:

$$\nabla_x \mathbb{E}[H_i|X_i = x] = \int dF_{V|W}(v|w) \int dh \cdot 1 \cdot f_H(h|x,v) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|H_i = h,x,v\right] \tag{36a}$$

And with integer categories $\mathcal{R}$, using Theorem 1:

$$\nabla_x \mathbb{E}[R_i|X_i = x] = \int dF_{V|W}(v|w) \sum_r f_H(\tau_v(r)|x,v) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|H_i = \tau_v(r),x,v\right] \tag{36b}$$

These three expressions differ only in what multiplies $f_H(h|x,v) \cdot \mathbb{E}\left[\nabla_x h(x,U_i)|h,x,v\right]$ for various values of h. Relative to (36a), (35) 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 (36b) to (35) 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 $\tau_v(r)$.[41]

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 9 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). In Appendix G.7, I give a formal definition of this "dense response limit" corresponding to an integer

[41] In the case of linear reporting functions with a continuous response space, Proposition 9 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. (35) reduces to $\partial_{x_1}\mathbb{E}[R_i|x] = \beta_1 \cdot \int dF_{V|W}(v|w) \cdot \frac{\bar{R}}{\mu(v)-\ell(v)} \cdot P(0 < R_i < \bar{R}|x,v)$ 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 $\partial_{x_1}\mathbb{E}[R_i|X_i = x] = \beta_1 \cdot P(0 < R_i < R|X_i = x)$.

response space $\mathcal{R} = \left\{0, 1, \ldots, \bar{R}\right\}$, which proves useful in the foregoing analysis. Appeal to this limit is indicated by the symbol $\overset{R}{\to}$ in the results of Section 5.

F Additional tables and figures

Δ	# r's	1	10	11	1000
-0.5	-1.77878	1.019952	1.006977	0.997881	1.018028
-0.1	-0.36785	1.016885	0.998345	0.996135	1.000664
0.1	0.367972	1.008979	1.004651	0.997364	1.001079
0.25	0.912569	1.012629	1.010883	1.009681	1.004132
0.5	1.779339	1.009398	1.020753	1.035096	1.017904
1	3.230443	1.136782	1.086727	1.036715	1.061607
5	5.013323	0.493566	0.579100	0.526786	0.544236
1/NB		2.369596	1.828073	1.841236	1.873973

Table 1: $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number $|\mathcal{V}| \in \{1, 10, 11, 1000\}$ of reporting functions. All cells take $\bar{R} = 11$ response categories, and the column labeled $\#r's$ reports the average number of these 11 thresholds crossed by the value of Δ corresponding to that row, averaged over the distribution of $H_i|X_i = x$. $H_i|X_i = x$ is standard normal, and in all cases thresholds are sampled as depicted in Figure 5.

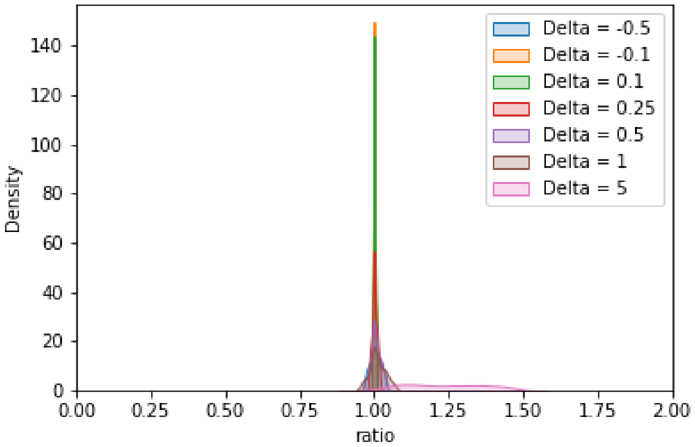


Figure 16: The distribution of $1 + \delta_{\Delta,x,V_i}$ across V_i is depicted across alternative values of Δ_i, with $H_i|X_i = x$ an equal mixture of $\mathcal{N}(-2, 1)$ and $\mathcal{N}(2, 1)$, $\bar{R} = 100$, and 1000 reporting functions with thresholds sampled as depicted in Figure 7.

Δ	# r's	1	10	11	1000
-0.5	-0.65973	0.987486	1.008402	1.002409	1.002822
-0.1	-0.13228	0.998206	0.997224	1.000729	1.000005
0.1	0.132622	0.998407	1.000369	0.998732	1.000106
0.25	0.330621	0.999557	1.000884	1.001690	1.000275
0.5	0.660180	0.998541	1.001786	1.003425	1.003705
1	1.296189	1.000055	1.006245	1.020919	1.009504
5	4.851162	1.408659	1.182369	1.230598	1.225963
1/NB		1.785322	1.425248	1.507230	1.483049

Table 2: $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number $|\mathcal{V}| \in \{1, 10, 11, 1000\}$ of reporting functions. All cells take $\bar{R} = 11$ response categories, and the column labeled $\#r's$ reports the average number of these 11 thresholds crossed by the value of Δ corresponding to that row, averaged over the distribution of $H_i|X_i = x$. $H_i|X_i = x$ is an equal mixture of $\mathcal{N}(-1/2, 1)$ and $\mathcal{N}(1/2, 1)$, and in all cases thresholds are sampled as depicted in Figure 7.

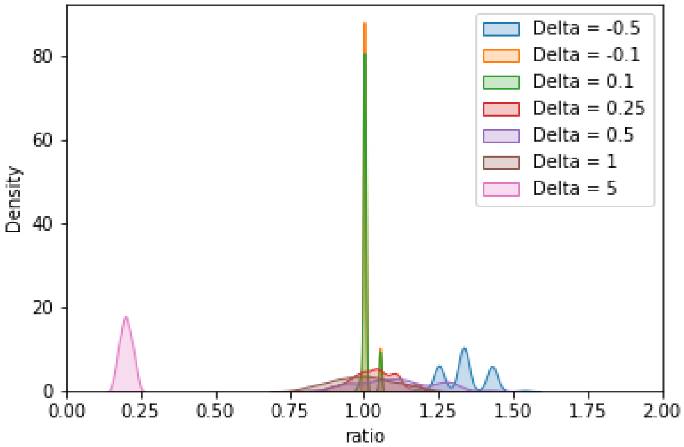


Figure 17: The distribution of $1 + \delta_{\Delta,x,V_i}$ across V_i is depicted across alternative values of Δ_i, with $H_i|X_i = x$ uniform on $[0, 1]$, $\bar{R} = 100$, and 1000 reporting functions with thresholds sampled as depicted in Figure 7.

Δ	# r's	1	10	11	1000
-0.5	-5.0	1.333333	1.351351	1.301775	1.332001
-0.1	-1.0	0.999999	1.005025	1.0	1.005277
0.1	0.999075	0.999999	1.005025	1.004566	1.004095
0.25	2.412154	1.003553	1.015483	1.036788	1.031220
0.5	4.131344	1.063924	1.021872	1.078504	1.101178
1	4.976521	1.074331	1.067193	0.989457	0.995304
5	4.979371	0.195783	0.198180	0.212324	0.199174
1/NB		1.219642	1.400211	1.309034	1.357831

Table 3: $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number $|\mathcal{V}| \in \{1, 10, 11, 1000\}$ of reporting functions. All cells take $\bar{R} = 11$ response categories, and the column labelled $\#r's$ reports the average number of these 11 thresholds crossed by the value of Δ corresponding to that row, averaged over the distribution of $H_i|X_i = x$. $H_i|X_i = x$ is uniform $[0, 1]$, and in all cases thresholds are sampled as depicted in Figure 8.

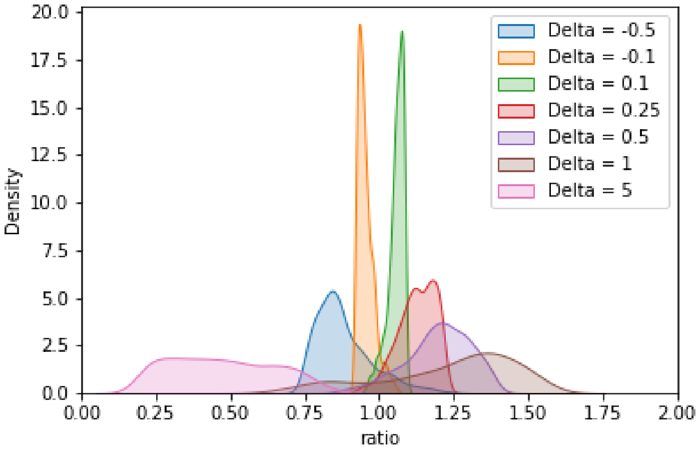
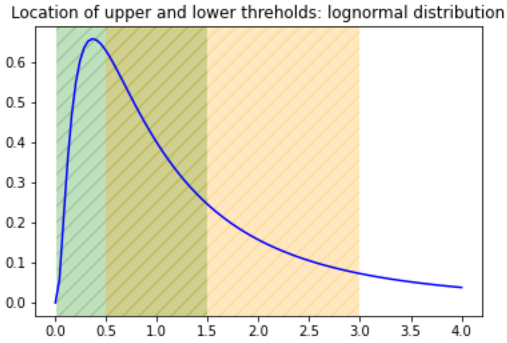


Figure 18: The distribution of $1 + \delta_{\Delta,x,V_i}$ across V_i is depicted across alternative values of Δ_i, with $H_i|X_i = x$ a standard log-normal, $\bar{R} = 100$, and 1000 reporting functions with thresholds sampled as depicted in Figure 7.

Δ	# r's	1	10	11	1000
-0.5	-1.70479	0.875985	0.850173	0.919761	0.891200
-0.1	-0.39632	0.938366	0.951939	0.949562	0.954190
0.1	0.412033	1.073444	1.045219	1.062470	1.052181
0.25	1.033962	1.120209	1.077792	1.136499	1.118010
0.5	1.980888	1.311599	1.181132	1.178890	1.185238
1	3.438382	1.122504	1.187908	1.187129	1.196697
5	4.659613	0.190009	0.427811	0.435993	0.439618
1/NB		1.799321	1.459198	1.573983	1.445131

Table 4: $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number $|\mathcal{V}| \in \{1, 10, 11, 1000\}$ of reporting functions. All cells take $\bar{R} = 11$ response categories, and the column labeled $\#r's$ reports the average number of these 11 thresholds crossed by the value of Δ corresponding to that row, averaged over the distribution of $H_i|X_i = x$. $H_i|X_i = x$ is standard log-normal, and in all cases thresholds are sampled as depicted in Figure 9.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.697930	0.730712	0.746278	0.755180
-0.1	0.914523	0.921389	0.924831	0.929371
0.1	1.092894	1.085627	1.079682	1.072299
0.25	1.242410	1.222273	1.206747	1.194887
0.5	1.461151	1.413333	1.394014	1.366348
1	1.729324	1.596417	1.559468	1.534750
5	0.630123	0.627392	0.608452	0.583946
1/NB	2.333187	0.562761	-235.774	-1.75283

Figure 19: $H_i|X_i = x$ is standard lognormal, and 1000 reporting functions are drawn from $\ell(v) \sim U[.01, 1.5]$, $\mu(v) \sim U[0.5, 3]$. Thus, the highest threshold μ for some individuals is lower than the lowest threshold ℓ is for other individuals. 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 $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

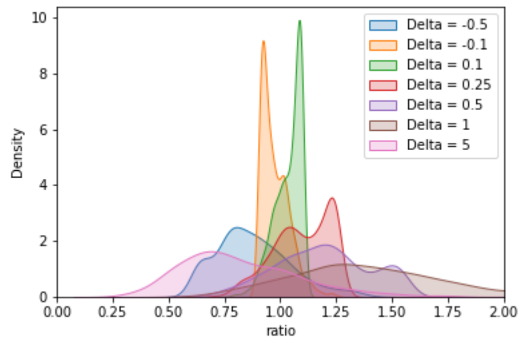
Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.608198	0.830388	0.862295	0.908220
-0.1	0.904246	0.924110	0.952734	0.984422
0.1	1.103991	1.083320	1.054811	1.022938
0.25	1.275106	1.205307	1.116844	1.061898
0.5	1.594517	1.299465	1.193268	1.132321
1	1.955425	1.359822	1.262187	1.213427
5	0.529326	0.462219	0.454500	0.458494
1/NB	1.341849	1.341849	1.341849	1.341849

Figure 20: $H_i|X_i = x$ is standard log-normal, and all individuals have the same linear reporting function with $\mu(v) = 0.1$ and $\ell(v) = 0.2$. Table reports values of $w_{x,x'}/\frac{1}{2}(w_x + w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$.

G Proofs

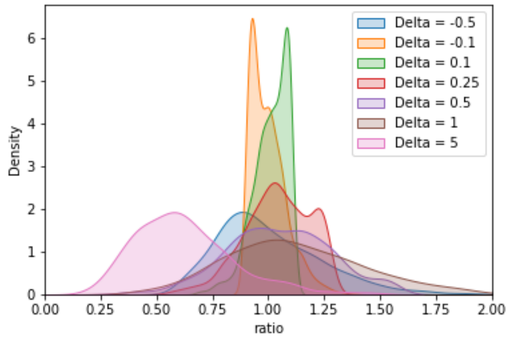
G.1 Proof of Proposition 1

Fix any v. First we show that if (5) holds for all r then Assumption HONEST holds. Indeed, suppose that for some $h' > h$ we had $r(h', v) < r(h, v)$. Substituting $r = r(h', v)$



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.888923	0.878441	0.873254	0.874889
-0.1	0.963337	0.981338	0.976410	0.979380
0.1	1.030927	1.026978	1.029188	1.026451
0.25	1.047773	1.071765	1.082974	1.074864
0.5	1.126755	1.175856	1.166967	1.167957
1	1.364791	1.335619	1.334358	1.325175
5	0.756363	0.773198	0.776693	0.767673
1/NB	1.341849	1.341849	1.341849	1.341849

Figure 21: The distribution of $1+\delta_{\Delta,x,V_i}$ across 1000 reporting functions (left), and values of $w_{x,x'}/\frac{1}{2}(w_x+w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$ (right), for $H_i|X_i = x$ following a log normal distribution with all thresholds sampled individually from a uniform distribution on $[0.1, 3]$. Thus, thresholds are not equally spaced within individual reporting functions.



Δ	$\bar{R}$=2	$\bar{R}$=5	$\bar{R}$=11	$\bar{R}$=100
-0.5	0.981985	0.976757	0.980952	0.986851
-0.1	0.985272	0.990361	0.994975	0.993520
0.1	0.996470	1.006833	1.007525	1.007262
0.25	1.011348	1.028414	1.029626	1.017586
0.5	1.043327	1.044149	1.030733	1.043710
1	1.062959	1.096241	1.084608	1.075300
5	0.606502	0.594470	0.585655	0.588043
1/NB	1.341849	1.341849	1.341849	1.341849

Figure 22: The distribution of $1+\delta_{\Delta,x,V_i}$ across 1000 reporting functions (left), and values of $w_{x,x'}/\frac{1}{2}(w_x+w_{x'})$ as a function of Δ and the number of response categories $\bar{R}$ (right), for $H_i|X_i = x$ following a log normal distribution with all thresholds sampled individually from a normal distribution with mean 2 and variance 1. Thus, thresholds are not equally spaced within individual reporting functions.

into (5), 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 (5) for $h = \tau_v(r^*)$, because $r(h,v) > r \implies h > \tau_v(r)$.

Now we will show that if Assumption HONEST holds then (5) 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 (5), without even invoking Assumption HONEST. 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 HONEST: $h \leq \tau_v(r) \implies r(h,v) \leq r(\tau_v(r),v)$. 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 Theorem 1

By the law of iterated expectations, Lemma 1, and then $\{X_{ji} \perp\!\!\!\perp V_i\}\,|W_i$:

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

By totally differentiating $Q_{H|XV}(F_{H|XV}(h|x,v)|x,v) = h$ with respect to x_j:

$$\frac{\partial}{\partial x_j} P(H_i \leq h|X_i = x, V_i = v) = -f_H(h|x,v)\, \frac{\partial}{\partial x_j} Q_{H|XV}(\alpha|x,v)\Big|_{\alpha = F_{H|XV}(h|x,v)}$$

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

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

Using that $\{X_{ji} \perp\!\!\!\perp U_i\}|(V_i, W_i)$, the theorem of Hoderlein and Mammen (2007) implies that

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

Therefore:

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

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).

One can also establish the final result of Theorem 1 by applying Theorem 2 and letting $x' \to x$ (see Footnote 27).

G.3 Proof of Proposition 1

Consider a binary covariate $A_i \in \{0,1\}$

$$\frac{\frac{d}{dx_j}\mathbb{E}[A_i \cdot R_i|X_i = x]}{\frac{d}{dx_j}\mathbb{E}[R_i|X_i = x]} = \frac{\frac{d}{dx_j}P(A_i = 1|X_i = x) \cdot \mathbb{E}[R_i|X_i = x, A_i = 1]}{\frac{d}{dx_j}\mathbb{E}[R_i|X_i = x]}$$

For any x and x':

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x]}{\mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x]} \\
&\qquad = \int dF_{V|W}(v|w) \cdot \frac{\mathbb{E}[A_i \cdot \mathbb{1}(H_i \le \tau_v(r))|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(H_i \le \tau_v(r))|X_i = x]}{\mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x]} \\
&\qquad = \frac{1}{denom}\int dF_{V|W}(v|w) \cdot \mathbb{E}[A_i \cdot \mathbb{1}(H_i \le \tau_v(r))|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(H_i \le \tau_v(r))|X_i = x] \\
&\qquad = \frac{1}{denom}\int dF_{V|W}(v|w) \cdot \mathbb{E}[A_i \cdot \mathbb{1}(h(x,U_i) \le \tau_v(r))|w] - \mathbb{E}[A_i \cdot \mathbb{1}(h(x',U_i) \le \tau_v(r))|w] \\
&\qquad = \frac{1}{denom}\int dF_{V|W}(v|w) \cdot \mathbb{E}[A_i \cdot \mathbb{1}(h(x,U_i) \le \tau_v(r) < h(x',U_i))|w] - \mathbb{E}[A_i \cdot \mathbb{1}(h(x',U_i) \le
\end{aligned}$$

where $denom = \mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[\mathbb{1}(R_i \le r)|X_i = x]$. If all units have the same sign of treatment effect (assume positive WLOG) then:

$$\begin{aligned}
&\frac{\mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x]}{P(R_i \le r|X_i = x') - P(R_i \le r|X_i = x)} \\
&\qquad = \frac{\int dF_{V|W}(v|w) \cdot \mathbb{E}[A_i \cdot \mathbb{1}(h(x,U_i) \le \tau_v(r) < h(x',U_i))|w]}{\int dF_{V|W}(v|w) \cdot \mathbb{E}[\mathbb{1}(h(x,U_i) \le \tau_v(r) < h(x',U_i))|w]} \\
&= \frac{\mathbb{E}[A_i \cdot \mathbb{1}(h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i))|W_i = w]}{P(h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i)|W_i = w)} \\
&= \frac{P(h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i)|W_i = w) \cdot \mathbb{E}[A_i|h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i), W_i = w]}{P(h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i)|W_i = w)} \\
&= \mathbb{E}[A_i|h(x,U_i) \le \tau_{V_i}(r) < h(x',U_i), W_i = w]
\end{aligned}$$

Letting x' differ from x only in component j and taking the limit as $x' \to x$ where

$x_j' = x_j + \delta$:

$$\begin{aligned}
\frac{\partial_{x_j}\mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x]}{\partial_{x_j}P(R_i \le r|X_i = x)} &= \lim_{\delta \downarrow 0} \frac{\mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x]}{\delta} \\
&\quad \cdot \lim_{\delta \downarrow 0} \frac{\delta}{P(R_i \le r|X_i = x') - P(R_i \le r|X_i = x)} \\
&= \lim_{\delta \downarrow 0} \frac{\mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x'] - \mathbb{E}[A_i \cdot \mathbb{1}(R_i \le r)|X_i = x]}{P(R_i \le r|X_i = x') - P(R_i \le r|X_i = x)} \\
&= \lim_{\delta \downarrow 0} \mathbb{E}[A_i|h(x, U_i) \le \tau_{V_i}(r) < h(x', U_i), W_i = w] \\
&= \mathbb{E}[A_i|h(x, U_i) = \tau_{V_i}(r), W_i = w]
\end{aligned}$$

G.4 Proof of Proposition 2

Given IDR and (12), we have

$$\partial_{x_1}\mathbb{E}[R_i|X_i = x] = \mathbb{E}\left[w_x(V_i)|H_i \in \tau_{V_i}, X_i = x\right] \cdot \mathbb{E}\left[\partial_{x_1}h(x, U_i)|H_i \in \tau_{V_i}, X_i = x\right]$$

So the RHS of (13) becomes: $\mathbb{E}\left[\partial_{x_1}h(x, U_i)\right| H_i \in \tau_{V_i}, x\left] / \mathbb{E}\left[\partial_{x_2}h(x, U_i)\right| H_i \in \tau_{V_i}, x\right]$. Now, using $Cov\left(\frac{\partial_{x_1}h(x,U_i)}{\partial_{x_2}h(x,U_i)}, \partial_{x_2}h(x, U_i)\middle| H_i \in \tau_{V_i}, x\right) \le 0$,

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

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

G.5 Proof of Proposition 3

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. (15).

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:

$$(\partial_{x_k} \log g(x))^{-1} = \frac{g(x)}{\partial_{x_k} g(x)} = 1 + \sum_{j \neq k} \frac{\partial_{x_j} g(x)}{\partial_{x_k} g(x)} \cdot x_j$$

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

$$g(x) = e^{\int_{x_j^*}^{x_j} \left(1+\sum_{j\neq k} \frac{\partial_{x_j} \mathbb{E}[R_i|(x_1,\dots x_{j-1},t,0,\dots,0)]}{\partial_{x_k} \mathbb{E}[R_i|(x_1,\dots x_{j-1},t,0,\dots,0)]} \cdot x_j \right)^{-1} dt} \tag{37}$$

G.6 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 $P(R_i \leq r|x') - P(R_i \leq r|x)$. 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 $P(R_i \leq r|x') - P(R_i \leq r|x)$. The RHS of Theorem 2 can be written as

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

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

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

EXOG

$$
\begin{aligned}
&P(R_i \le r|X_i = x') - P(R_i \le r|X_i = x)\\
&= \int dF_{UV|X}(u,v|x') \cdot \mathbb{1}(r(h(x',u),v) \le r) - \int dF_{UV|X}(u,v|x) \cdot \mathbb{1}(r(h(x,u),v) \le r)\\
&= \int dF_{UV|X}(u,v|x') \cdot \mathbb{1}(h(x',u) \le \tau_v(r)) - \int dF_{UV|X}(u,v|x) \cdot \mathbb{1}(h(x,u) \le \tau_v(r))\\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \le \tau_v(r)|X_i = x', V_i = v) - P(h(x,U_i) \le \tau_v(r)|X_i = x, V_i = v)\}\\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \le \tau_v(r)|X_i = x, V_i = v) - P(h(x,U_i) \le \tau_v(r)|X_i = x, V_i = v)\}\\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \le \tau_v(r) \text{ but not } h(x,U_i) \le \tau_v(r)|X_i = x, V_i = v)\\
&\qquad\qquad - P(h(x,U_i) \le \tau_v(r) \text{ but not } h(x',U_i) \le \tau_v(r)|X_i = x, V_i = v)\}\\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x',U_i) \le \tau_v(r) < h(x,U_i)|x,v) - P(h(x,U_i) \le \tau_v(r) < h(x',U_i)|x,v)\}\\
&= \int dF_{V|W}(v|w) \cdot \{P(h(x,U_i) \in (\tau_v(r), \tau_v(r) - \Delta_i]|x,v) - P(h(x,U_i) \in (\tau_v(r) - \Delta_i, \tau_v(r)]|x,v)\}\\
&= \int dF_{V|W}(v|w) \cdot \{P(H_i \in (\tau_v(r), \tau_v(r) - \Delta_i]|x,v) - P(H_i \in (\tau_v(r) - \Delta_i, \tau_v(r)]|x,v)\}\\
&= -\int dF_{V|W}(v|w) \cdot \int d\Delta \int_{\tau_v(r)-\Delta}^{\tau_v(r)} dy \cdot f_H(\Delta, y|X_i = x, V_i = v)\\
&= -\int dF_{V|W}(v|w) \cdot \int d\Delta \cdot f_H(\Delta|x,v) \int_{\tau_v(r)-\Delta}^{\tau_v(r)} dy \cdot f_H(h|\Delta, x, v)\\
&= -\int dF_{V|W}(v|w) \cdot \int d\Delta \cdot \bar{f}(\Delta, \tau_v(r), x, v) \cdot f_H(\Delta|x,v) \cdot \Delta
\end{aligned}
$$

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

G.7 The dense response limit: definition

Consider a fixed range $[0, \bar{R}]$ of responses for some integer $\bar{R}$, and a sequence of response spaces $\mathcal{R}_n = \{0, 1/n, 2/n, \dots, n\bar{R}/n\}$. For a fixed value of reporting heterogeneity v, consider a sequence of reporting functions $r_n(\cdot, v)$ indexed by n.[42] 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 $r(\cdot, v)$ in the dense response limit, denoted as $r_n(\cdot, v) \overset{R}{\to} r(\cdot, v)$, if:*

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

[42] Note that $\mathcal{R}_1$ is the set of integers from 1 to $\bar{R}$. However we need not see $n = 1$ as the “first” n in our sequence; we may instead begin the sequence with fractional n, for each divisor of $\bar{R}$. For example if $\bar{R}$ is even then $\mathcal{R}_{0.5} = \{0, 2, 4, \dots \bar{R}\}$.

for any sequence of $r_n \in \mathcal{R}_n$ *such that* $\lim_{n\to\infty} r_n = r$ *for some* $r \in [0, \bar{R}]$. *For any functional of all response functions* $\theta(\{r_n(\cdot, v)\}_{v\in\mathcal{V}})$, *let* $\theta(r_n) \overset{R}{\to} \Theta$ *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 $[0, \bar{R}]$.

G.8 Proof of Proposition 4

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

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

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

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

Now given the assumption that Δ_i and $\bar{r}'(H_i, \Delta_i, V_i)$ are uncorrelated conditional on $X_i = x$. Then

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

G.9 Proof of Proposition 5

Starting with Proposition 4, observe that $\bar{r}'(y,\Delta,v) := \frac{1}{\Delta}\int_y^{y+\Delta} r'(h,v)dh$ is equal to

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

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 $f_H(y|\Delta_i=\Delta, X_i=x, V_i=v)$ and $F(y)$ the corresponding conditional CDF. Let us keep v also implicit in both ℓ and μ. If we let θ denote the quantity $\frac{1}{r'(v)}\int dy\cdot\bar{r}'(y,\Delta,v)$ for a fixed Δ, then:

$$\theta = \begin{cases} [F(\ell)-F(\ell-\Delta)]\mathbb{E}\left[\left.\frac{H_i-(\ell-\Delta)}{\Delta}\right| H_i\in[\ell-\Delta,\ell]\right] + F(\mu-\Delta) & \\ \qquad -F(\ell) + [F(\mu)-F(\mu-\Delta)]\mathbb{E}\left[\left.\frac{\mu-H_i}{\Delta}\right| H_i\in[\mu-\Delta,\mu]\right] & \text{if } \Delta>0 \\ [F(\ell+|\Delta|)-F(\ell)]\mathbb{E}\left[\left.\frac{H_i-\ell}{|\Delta|}\right| H_i\in[\ell,\ell+|\Delta|]\right] + F(\mu) & \\ \qquad -F(\ell+|\Delta|) + [F(\mu+|\Delta|)-F(\mu)]\mathbb{E}\left[\left.\frac{\mu+\Delta-H_i}{|\Delta|}\right| H_i\in[\mu,\mu+|\Delta|]\right] & \text{if } \Delta<0 \end{cases} \tag{38}$$

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

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

A lower bound on the weight $w_{x,x'}$ on causal effects in $\mathbb{E}[R_i|X_i=x'] - \mathbb{E}[R_i|X_i=x]$ can thus given by averaging over V_i (c.f. Proposition 4):

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

Note that this exactly the same as the average between the weights w_x and $w_{x'}$ corresponding to using continuous variation at $X_i = x$ and $X_i = x'$, respectively. For example (c.f. Eq. 35):

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

This leads to the lower bound of $w_{x,x'}/(\frac{1}{2}w_x + \frac{1}{2}w_{x'}) \geq 1$ in Proposition 4.

Now, to obtain an upper bound, notice that an upper bound on θ occurs if we imagine putting all of the mass in each of the interval conditional expectations in (38) to the right in the intervals that depend on ℓ, and at the left end for the intervals that depend on μ. Then:

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

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

$$\theta \leq F(\mu(v)|\Delta,x,v) - F(\ell(v)|\Delta,x,v) + F(\mu(v)|\Delta,x',v) - F(\ell(v)|\Delta,x',v) \qquad (39)$$

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 $w_{x,x'}$ on causal effects in $\mathbb{E}[R_i|X_i = x'] - \mathbb{E}[R_i|X_i = x]$ is:

$$\begin{aligned}
w_{x,x'} \geq \int dF_{V|W}(v|w) \cdot \int d\Delta \cdot f_H(\Delta|x,v) \cdot \{&F(\mu(v)|\Delta,x',v) - F(\ell(v)|\Delta,x',v) \\
&+ F(\mu(v)|\Delta,x,v) - F(\ell(v)|\Delta,x,v)\}
\end{aligned}$$

leading to the upper bound of $w_{x,x'}/(\frac{1}{2}w_x + \frac{1}{2}w_{x'}) \leq 2$ in Proposition 4.

Now consider the final condition in Proposition 4. That $w_{x,x'}/w_x \geq 1/2$ follows from the above since $F(\mu(v)|\Delta,x',v) - F(\ell(v)|\Delta,x',v) \geq 0$ for all Δ, x, v. For the upper bound we have

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

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

G.10 Proof of Proposition 7

Using integration by parts and IDR:

$$
\begin{aligned}
P(R_i \leq r|X_i = x) &= \int_h P(r(h,V_i) \leq r|H_i = h, X_i = x)\cdot dF_{H|X}(h|x) \\
&= \int_h P(r(h,V_i) \leq r)\cdot dF_{H|X}(h|x) \\
&= F_{H|X}(h|x)P(r(h,V_i) \leq r)\big|_h - \int_h F_{H|X}(h|x)\cdot \frac{d}{dh}P(r(h,V_i) \leq r)dh
\end{aligned}
$$

This implies that

$$
\begin{aligned}
P(R_i \leq r|x') - P(R_i \leq r|x) &= -\int_h \left\{F_{H|X}(h|x') - F_{H|X}(h|x)\right\}\cdot \frac{d}{dh}P(r(h,V_i) \leq r) \\
&= -\int_h \left\{F_{H|X}(h|x') - F_{H|X}(h|x)\right\}\cdot \frac{d}{dh}P(h \leq \tau_{V_i}(r)) \\
&= \int_h \left\{F_{H|X}(h|x') - F_{H|X}(h|x)\right\}\cdot f_{\tau_V}(h)
\end{aligned}
$$

since the first term does not depend on x.

G.11 Proof of Proposition 8

The following sequence of steps uses the law of iterations, then IDR, then $\mathbb{E}[A_i|X_i = x] = \int_0^1 Q_{A|X=x}(u)\cdot du$ for any random variable A, and finally that $Q_{r(H,v)|X=x'}(u) =$

$r(Q_{H|X=x'}(u), v)$ since $r(\cdot, v)$ is weakly increasing and left-continuous for all v (Hosseini, 2010):

$$\begin{aligned}
\mathbb{E}[R_i|x'] - \mathbb{E}[R_i|x] &= \int dF_{V|W}(v|w) \cdot \mathbb{E}[r(H_i,v)|X_i = x', V_i = v] - \mathbb{E}[r(H_i,v)|X_i = x, V_i = v] \\
&= \int dF_{V|W}(v|w) \cdot \{\mathbb{E}[r(H_i,v)|X_i = x'] - \mathbb{E}[r(H_i,v)|X_i = x]\} \\
&= \int dF_{V|W}(v|w) \cdot \int_0^1 \left\{Q_{r(H,v)|X=x'}(u) - Q_{r(H,v)|X=x}(u)\right\} du \\
&= \int dF_{V|W}(v|w) \cdot \int_0^1 \left\{r(Q_{H|X=x'}(u), v) - r(Q_{H|X=x}(u), v)\right\} du \\
&= \int_0^1 \left[\int dF_{V|W}(v|w) \left\{r(Q_{H|X=x'}(u), v) - r(Q_{H|X=x}(u), v)\right\}\right] du \\
&= \int_0^1 \bar{r}'_{x',x}(u) \cdot \left\{Q_{H|X=x'}(u) - Q_{H|X=x}(u)\right\} du
\end{aligned}$$

where the interchange of integrals is warranted provided that each of $\mathbb{E}[R_i|x']$ and $\mathbb{E}[R|x]$ are finite, because

$$\begin{aligned}
&\int dF_{V|W}(v|w) \cdot \int_0^1 \left|r(Q_{H|X=x'}(u), v) - r(Q_{H|X=x}(u), v)\right| du \\
&\qquad \leq \int dF_{V|W}(v|w) \cdot \int_0^1 \left|r(Q_{H|X=x'}(u), v)\right| + \left|r(Q_{H|X=x}(u), v)\right| du \\
&\qquad = \mathbb{E}[|R_i||x'] - \mathbb{E}[|R_i||x] < \infty
\end{aligned}$$

G.12 Proof of Proposition 9

By the law of iterated expectations:

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

where $f_H(h|x,v)$ is the density of H_i conditional on $X_i = x, V_i = v$. Thus, using REG to move the derivative inside the integral:

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

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

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

Now use integration by parts, applying the assumed boundary condition eliminates the first term and

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

establishing the result.