Identification of Expected Outcomes in a Data Error Mixing Model with Multiplicative Mean Independence

Brent Kreider
Department of Economics
Iowa State University
bkreider@iastate.edu

John V. Pepper
Department of Economics
University of Virginia
jvpepper@virginia.edu

May 2009

Abstract. We consider the problem of identifying a mean outcome in corrupt sampling where the observed outcome is drawn from a mixture of the distribution of interest and another distribution. Relaxing the contaminated sampling assumption that the outcome is statistically independent of the mixing process, we assess the identifying power of an assumption that the conditional means of the distributions differ by a factor of proportionality. For binary outcomes, we consider the special case that all draws from the alternative distribution are erroneous. We illustrate how these models can inform researchers about illicit drug use in the presence of reporting errors.

Keywords: contaminated sampling, corrupt sampling, measurement error, partial identification, nonparametric bounds

JEL codes: C1, C10, C14, I12

The authors received valuable comments from an anonymous referee, the editor, an associate editor, Phil Cross, Francesca Molinari, Steve Stern, and seminar participants at Georgetown University, Iowa State University, and the University of Virginia. They also benefited from discussions at meetings of the Econometric Society and the Southern Economic Association. This research was supported in part by the Bankard Fund for Political Economy.

1 Introduction

Empirical analyses have long struggled with how to draw credible inferences in light of data errors that arise from a variety of sources and are often known to be extensive. In the 2001 Current Population Survey, for example, the wages of nearly a third of the workers are imputed (Hirsch and Schumacher, 2004) and validation studies consistently reveal large and systematic reporting errors even for variables one might think should be reported accurately (see, e.g., Bound *et al.* (2001)). Credible solutions to these data error problems, however, remain elusive. The assumptions of the nondifferential errors-in-variables models are often untenable (see, e.g., Bound *et al.* (2001) for discussion), and alternative models rely on parametric assumptions that can be difficult to justify in many applications. There is good reason, therefore, to consider alternative approaches.

Recently, a growing body of literature conceptualizes the data error problem using a mixture model in which the observed outcome distribution is a mixture of the unobserved distribution of interest, F, and another unobserved distribution, G (see, e.g., Horowitz and Manski (1995, HM henceforth), Lambert and Tierney (1997), Dominitz and Sherman (2004), Mullin (2005), and Kreider and Pepper (2007 and 2008)). In this environment, the "contaminated sampling" model pertains to the case in which data errors are known to be statistically independent of sample realizations from the population of interest. The more general "corrupted sampling" model pertains to the case that nothing is known about the pattern of data errors. Using nonparametric methods, HM derive sharp bounds on parameters of F under both corrupt and contaminated sampling for the case that the researcher has an upper bound on the fraction of draws that come from G.

In this paper, we study what can be inferred about the expected outcome given assumptions about how the mean of F varies with the mixing process. Specifically, we relax the statistical independence assumption embodied in the contamination model to instead consider the identifying power of mean independence and, most notably, a variant we call

"multiplicative mean independence." In the latter case, the conditional means are allowed to differ by a known or bounded factor of proportionality. Our approach is motivated by the observation that, in practice, corrupt sampling bounds tend to be frustratingly wide given the lack of structure on the measurement error process, while the contamination independence assumption is often untenable. For example, income nonresponse is thought to be related to income levels, and the accuracy of reported health status is thought to be related to true health status (e.g., Bound *et al.*, 2001). Likewise, in our empirical application described below, the misreporting of illicit drug use is thought to occur more frequently among users than nonusers. While the independence assumption is unlikely to hold in these examples, it seems reasonable to apply the multiplicative mean independence model developed in this paper.

We begin in Section 2 by studying the identifying power of the multiplicative mean independence model. Applying the contaminated sampling results in HM, we are able to partially identify the expected outcome for any distribution with a finite mean. We then illustrate the partial identification bounds under the important special case of a binary outcome distribution. In Section 3, we further consider the problem of identifying binary outcome distributions under additional restrictions. In this context, researchers using mixture models often assume (sometimes implicitly) that all draws from G are known to be inaccurate, as might be the case when mixing arises from response error. This response error mixture model provides a link between F and G that is especially informative for binary outcome distributions: in this case, realizations from G reveal precisely what the outcome of interest is not. Not surprisingly, imposing this additional assumption has substantial identifying power.

The parts of our analysis that focus on binary variables are related to Molinari (2008) who presents an alternative conceptualization of the data error problem for discrete outcome variables. In her "direct misclassification" approach, one focuses on assumptions related to

classification error rates instead of restrictions on the mixing process. For corrupt and contaminated samples, Molinari derives the same closed-form bounds provided in HM. While she does not consider multiplicative mean independence restrictions directly, in principle her computational methods can handle this type of restriction when considering binary outcome distributions. From a practical perspective, however, it is not clear how one would explicitly map our more general multiplicative mean independence assumption into exhaustive restrictions on misclassification probabilities. Moreover, we derive closed-form identification regions, tailored to our maintained assumptions, that are not available in her analysis.

In contexts where theory or validation data implies direct restrictions on misclassification probabilities, Molinari's framework provides a natural method for producing the associated identification regions. Our proposed framework is natural for cases in which a researcher has knowledge about conditional means. For example, it is straightforward in our framework to impose a restriction that the prevalence rate of illicit drug use is higher among inaccurate responders than among accurate responders. The mixing distribution framework is also well-suited for studying data problems in which corrupt responses do not necessarily constitute misclassifications. For example, cases in which the data are corrupted with imputations or proxy responses are better handled in a mixing framework that allows for the possibility that observations from G may be accurate.

In Section 4, we apply these methods to the problem of using self-reported surveys to infer the fraction of the noninstitutionalized population consuming illicit drugs. In this application, we find that a response error model with multiplicative mean independence is easy to motivate and can have substantial identifying power. Finally, we draw conclusions in Section 5.

Throughout, we simplify the exposition by leaving implicit any conditioning variables; one can condition our results on any observed covariates. The recent literature on partial identification has considered restrictions between covariates and the mixing distributions.

In particular, instrumental variable and verification assumptions have been shown to reduce the ambiguity resulting from data errors (see Lambert and Tierney (1997), Dominitz and Sherman (2004), and Kreider and Pepper (2007 and 2008)). Layering these assumptions on top of the multiplicative mean independence assumption will serve to narrow the bounds presented in this paper.

Finally, since our focus is on identification, we treat identified quantities as known. In the empirical section, we can consistently estimate the derived identification bounds by replacing population probabilities with their sample analogs. To account for sampling variability, a growing body of literature has developed procedures for drawing inferences in partially identified models (e.g., Imbens and Manski, 2004; Beresteanu and Molinari, 2008; Rosen, 2008; Stoye, forthcoming) that can be applied to the bounds derived in Proposition 1. Application of these approaches for the response error mixture model in Proposition 2, however, can be complicated because the bounds vary discontinuously. Molinari (2008) proposes a method of inference for similar problems that might be useful in this setting. We focus, however, on the question of identification.

2 Multiplicative Mean Independence

In this section, we define and characterize the identifying power of the multiplicative mean independence assumption. In Section 2.1, we introduce the notation and the basic question, and then we review some of the relevant findings from HM. In Section 2.2, we define the multiplicative independence assumption and derive bounds on the mean outcome that apply under this restriction. In Section 2.3, we consider the special case of a binary outcome where we can find closed form bounds.

2.1 The Mixture Model and Identification With Contaminated Sampling

To distinguish between the reported and true outcome distributions, let W be the outcome of interest and let Z indicate whether the observed outcome, X, comes from F or G. Assume that the means of X and W exist. Our interest is in learning $w \equiv E(W)$, but we only observe the outcome distribution $X = WZ + \widetilde{W}(1-Z)$ where $\widetilde{W}$ is the random variable drawn from the alternative distribution, G. An identification problem arises because knowledge of X alone does not reveal $E(W)$.

The mean outcome, however, can be partially identified under a variety of different restrictions on the mixing process. A common starting point in this literature is to assume a known lower bound v on the fraction of cases that are drawn from the distribution of interest, F:

Assumption 1. $z \geq v$ (1)

where $z \equiv P(Z = 1)$. This type of restriction is used in the literatures on robust statistics (Huber, 1981) and data errors with binary regressors (see, e.g., Bollinger, 1996 and Frazis and Loewenstein, 2003). A particular lower bound restriction may be informed by a validation study of a related population or the known fraction of responses that are imputed (see, e.g., HM, Kreider and Pepper (2007 and 2008); Dominitz and Sherman (2004)).

Let $X \in [k_0^x, k_1^x]$ and, for simplicity, be continuous. Given this restriction, HM (1995, Corollary 4.1) show that

$$vE(X|X \leq \tau_X(v)) + (1-v)k_0^x \leq w \leq vE(X|X > \tau_X(1-v)) + (1-v)k_1^x \tag{2}$$

where $\tau_X(\cdot)$ is the quantile function for the distribution of X. These bounds are easily generalized to allow for non-continuous outcome distributions (see HM, 1995 and Dominitz and Sherman, 2004). Notice that in this conservative "corrupted sampling" environment,

identification of $E(W)$ deteriorates rapidly with the allowed fraction of misclassifications, $1-v$.

Prior information can narrow these corrupt sampling bounds. A common assumption known to have identifying power is that the sampling process is contaminated, in which case the mixing process, Z, is independent of the outcome distribution of interest: $P(W) = P(W|Z)$. Given Assumption 1, HM (1995, Corollary 4.1) derive sharp bounds on the conditional mean, $E(W|Z=1)$:

$$E\left(X|X \leq \tau_X(v)\right) \leq E(W|Z=1) \leq E\left(X|X > \tau_X(1-v)\right). \tag{3}$$

Under contaminated sampling, these bounds also apply to the quantity of interest, $E(W)$, since $E(W) = E(W|Z=1)$.

Two features of the contaminated sampling bounds are worth highlighting. First, the contaminated sampling bounds in Equation (3) are weakly narrower than the corrupt sampling bounds in Equation (2). Second, these sharp bounds on $E(W)$ are informative even if the support of the distribution of X is unbounded. Thus, given Assumption 1, we can find meaningful bounds on the conditional expectation $E(W|Z=1)$ for any observed outcome distribution with a finite mean. In the next section, we use this result to derive sharp bounds on $E(W)$ under a generalization of the contaminated sampling assumption.

2.2 Identification With Multiplicative Mean Independence

Given our interest in the mean outcome, $E(W)$, one obvious way to relax the statistical independence restriction is to consider a mean independence restriction that $E(W) = E(W|Z)$. As we saw in Equation (3), this mean independence assumption is sufficient to derive the contaminated sampling bounds on w. In many empirical applications, however, both the statistical independence and mean independence assumptions may be untenable. It seems unlikely, for example, that the misreporting of illicit drug use is orthogonal to actual drug

use status or that the true income distribution is mean independent of whether responses are imputed or self-reported.

Our notion of *multiplicative mean independence* generalizes the mean independence restriction by allowing the two conditional means to differ by a factor of proportionality. That is, for $z < 1$

Assumption 2. $E(W|Z = 0) = \gamma E(W|Z = 1)$ (4)

for some known or bounded value of $\gamma \in [0, \infty)$. Under fully accurate reporting, $z = 1$, Assumption 2 provides no identifying information: $E(W) = E(X)$. In some cases, a particular value of γ may be informed by a validation study of a related population. Otherwise, one can often rule out values of γ less than 1 or values greater than 1. For example, the use of illicit drugs is thought to be at least as prevalent among inaccurate reporters as among accurate reporters. In this context, a model that imposes the restriction $\gamma \geq 1$ may be credible when the restriction $\gamma = 1$ is untenable.

Proposition 1 below provides sharp bounds on the expected outcome $E(W)$ in this more general setting. We begin by deriving $E(W)$ as a function of γ, $E(X)$, and the unobserved probability z. Using the law of iterated expectations, we see that

$$E(W) = E(W|Z = 1)z + E(W|Z = 0)(1 - z).$$

Then, given Assumption 2, we have

$$E(W) = E(W|Z = 1)\left[1 + (\gamma - 1)(1 - z)\right]. \tag{5}$$

Bounds on $E(W)$ follow directly from Equation (5). To see this, suppose the fraction observations drawn from the distribution of interest, z, is known. As we saw above in Equation (3), HM derive informative bounds on the unknown conditional expectation, $E(W|Z = 1)$, under Assumption 1. These HM bounds apply whenever the mean of the observed outcome exists, regardless of whether the support of the distribution is bounded. If, however, the

unobserved random variable W is known to lie within the bounded support, $[k_0, k_1]$, then Assumption 2 further restricts $E(W|Z=1) \in \left[\frac{k_0}{\gamma}, \frac{k_1}{\gamma}\right]$. If the support is either unknown or unbounded, let $k_0 = -\infty$ and/or $k_1 = \infty$. Then, given z, we have

$$\begin{aligned} LB(z) &\equiv \max\left\{\frac{k_0}{\gamma}, E(X|X \leq \tau_X(z))\right\}[1+(\gamma-1)(1-z)] \\ &\leq E(W) \leq \\ UB(z) &\equiv \min\left\{\frac{k_1}{\gamma}, E(X|X > \tau_X(1-z))\right\}[1+(\gamma-1)(1-z)]. \end{aligned} \tag{6}$$

Thus, for a known z, Equation (6) provides informative bounds on $E(W)$.

When z is unknown, bounds on $E(W)$ are found by taking the infimum of $LB(\cdot)$ and the supremum of $UB(\cdot)$ over all feasible values of z. This set of feasible values is restricted directly by Assumption 1 and indirectly by Assumption 2. Assumption 1 rules out all values of $z < v$ and Assumption 2 rules out values of $z \in [0,1)$ where the associated HM bounds on $E(W|Z=1)$ lie strictly outside $\left[\frac{k_0}{\gamma}, \frac{k_1}{\gamma}\right]$. That is, a conjectured rate for z is feasible only if

$$\frac{k_0}{\gamma} \leq E(X|X > \tau_X(1-z)) \text{ and } \frac{k_1}{\gamma} \geq E(X|X \leq \tau_X(z)). \tag{7}$$

Thus, z is restricted to exceed v and to satisfy the condition in Equation (7). Notice that if v does not satisfy the condition in Equation (7), then the monotonicity of the HM bounds with respect to z implies that there are no feasible values of $z < 1$. In this case, $z = 1$ and $E(W) = E(X)$.

Given these restrictions on feasible values on z, we have:

Proposition 1 (multiplicative mean independence). *Suppose Assumptions 1 and 2 hold with* γ and v *known. Let* Θ *be the set of feasible values of* z *(defined by Equations (1) and (7)). Then*

$$\inf_{z^* \in \Theta} LB(z^*) \leq E(W) \leq \sup_{z^* \in \Theta} UB(z^*).$$

If the conditions in Equation (7) are satisfied, the lower bound simplifies to $\min\{E(X), LB(v)\}$

for $\gamma \leq 1$ *and the upper bound simplifies to* $\max\{E(X), UB(v)\}$ *for* $\gamma \geq 1$. *If the conditions in Equation (7) are not satisfied,* $E(W) = E(X)$.

A proof of these closed form results is provided in the appendix. Notice that closed form results for the upper and lower bounds on $E(W)$ can be found for certain γ but not in general. In particular, both terms of the upper bound in Equation (6) monotonically decrease with z when $\gamma \geq 1$, and both terms of the lower bound increase with z when $\gamma \leq 1$. In these cases, closed form bounds can be found by evaluating Equation (6) at the smallest feasible value of z subject to the constraints implied by Assumption 2. When the two terms move in opposite directions, no closed form result applies for all distributions and all z. Finally, notice that when when $\gamma = 1$, the Proposition 1 bounds are identical to the HM contaminated sampling bounds on the expected outcome in Equation (3).

2.3 Illustration: Binary Outcome Distribution

For binary outcomes, where $k_0 = 0$ and $k_1 = 1$, the HM bounds in Equation (3) become $(p - \min\{1 - v, p\})/v \leq E(W|Z = 1) \leq (p - \max\{p - v, 0\})/v$, where $p \equiv E(X)$. Applying Proposition 1, we find:

Corollary 1. Let W and X be Bernoulli random variables. Then by Proposition 1,

$$\min\{p, LB(v)\} \leq w \leq \max\{p, \min\{UB(v), p + \gamma(1 - p)\}\}.$$

A proof is provided in the appendix. Notice that the lower bound attains the HM corrupt sampling lower bound when $\gamma = 0$. The upper bound attains the HM corrupt sampling upper bound when $\gamma = \frac{v}{p}$ (for $v \geq p$).

Figure 1 illustrates these identification regions under hypothetical combinations of $\{\gamma, p\}$ where the curves LB_{MMI} and UB_{MMI} trace out the Proposition 1 bounds on w as a function of v. A representative case when $\gamma < 1$ is presented in Case A, and a representative case when $\gamma > 1$ is presented in Case B. The diagonal lines converging at $w = p$ reflect the HM corrupt sampling bounds. For $v < 1$, the Proposition 1 bounds are always weakly more

informative than the corrupt sampling HM bounds, and they may even be informative when there is no prior information on the degree of accurate reporting ($v = 0$). Consider, for example, Case A where $\gamma = 0.8$ and $p = 0.3$. If $v = 0$, the outcome distribution must lie within $[0, 0.86]$. In contrast, under corruption the data reveal nothing about the outcome distribution until v exceeds 0.3. Both the corrupt and contaminated sampling lower bounds are uninformative unless over 70% of the responses are known to come from the distribution of interest.

3 The Response Error Mixing Model

While realizations from G are often referred to as *data errors* (see HM, 1995), the mixing model alone does not impose the restriction that each draw from G is erroneous. This feature allows for the possibility that draws from G come from a proxy that, for some realizations, provides a valid measure of the distribution of interest (e.g., when contamination arises from imputation). For binary outcomes discussed in Section 2.3, however, data errors are often conceptualized as a response error with false negative and positive reports. Thus, we also consider the identifying power of the following response error assumption:

Assumption 3. $P(W = X|Z = 0) = 0.$ (8)

We refer to Assumption 3 as the *response error mixture model* in that all draws from the alternative distribution are known to be erroneous. This assumption provides a link between F and G that is informative for discrete outcome distributions. Realizations from G reveal what the outcome of interest is not. Thus, for a binary outcome, Assumption 3 implies that $P(W = 1|Z = 0) = P(X = 0|Z = 0)$ so that $X = WZ + (1 - W)(1 - Z)$.

To derive analytic identification regions when combining Assumptions 1-3, our strategy is to (a) derive the outcome probability, w, as a function of γ, p, and the unobserved probability z, (b) translate restrictions on false positives and false negatives into restrictions on possible values of z, and then (c) identify w extrema over valid candidates of z.

3.1 The Outcome Distribution and The Accurate Reporting Rate

Using the law of total probability, decompose the observed outcome distribution to consider information embedded in the reported classifications:

$$p = P(X = 1|Z = 1)z + P(X = 1|Z = 0)\,(1 - z)\,.$$

It follows from Assumptions 2 and 3 that

$$p = P(W = 1|Z = 1)\,[(\gamma + 1)\,z - \gamma] + (1 - z) \tag{9}$$

so that we can write the prevalence rate among accurate reporters as

$$P(W = 1|Z = 1) = \frac{z - (1 - p)}{(\gamma + 1)\,z - \gamma} \quad \text{if } z \neq \frac{\gamma}{\gamma + 1}. \tag{10}$$

Substituting (10) into (5), we can now write the outcome probability as a function of the unknown accurate reporting probability, z:

$$w(z) = \frac{[z - (1 - p)]\,[\gamma + (1 - \gamma)\,z]}{(\gamma + 1)\,z - \gamma} \qquad \text{if } z \neq \frac{\gamma}{\gamma + 1}. \tag{11}$$

Notice that for a given $z \neq \frac{\gamma}{\gamma+1}$ and γ, the outcome distribution w is identified. In contrast, knowledge of z and γ does not identify $w(z)$ under Assumption 2 alone (see Equation (6)). When $z = \frac{\gamma}{\gamma+1}$, Equation (9) reveals that $p = \frac{1}{\gamma+1}$; this outcome is treated as a special case in Proposition 2 below.

While w is not identified when z is unknown, the outcome distribution can be bounded by considering $w(z)$ over the feasible range of z. There are two sources of restrictions on z. First, values of z less than v are ruled out by Assumption 1. Second, given values of γ and p, restrictions on false positive and false negative classifications constrain the possible values of z. For $z \neq \frac{\gamma}{\gamma+1}$, use Equation (10) and Assumptions 2 and 3 to write the fraction of false positives as

$$\theta^{+} = P(W = 0|Z = 0)(1 - z) = \frac{(z - \gamma p)\,(1 - z)}{(\gamma + 1)z - \gamma} \tag{12}$$

and the fraction of false negatives as

$$\theta^- = P(W = 1|Z = 0)(1 - z) = \frac{\gamma\left[z - (1 - p)\right](1 - z)}{(\gamma + 1)z - \gamma}. \tag{13}$$

The fraction of false positives cannot be negative, nor can it exceed the total fraction of positive classifications: $\theta^+ \in [0, p]$. Similarly, the fraction of false negatives cannot be negative, nor can it exceed the total fraction of negative classifications: $\theta^- \in [0, 1 - p]$. These constraints imply the following restrictions on the accurate reporting rate:

Lemma 1. *Given Assumptions 2 and 3, the accurate reporting rate is bounded as follows: When $\gamma = 0$, $z \geq 1 - p$. When $\gamma > 0$, $z \leq \min\left\{\max\left\{0, \frac{\gamma-(1-p)}{\gamma}\right\}, 1 - p, \gamma p\right\}$ for $z < \frac{\gamma}{\gamma+1}$ and $z \geq \max\left\{\frac{\gamma-(1-p)}{\gamma}, 1 - p, \min\{\gamma p, 1\}\right\}$ for $z > \frac{\gamma}{\gamma+1}$. The value $z = \frac{\gamma}{\gamma+1}$ is only feasible if $p = \frac{1}{\gamma+1}$.*

A proof is provided in the appendix.

To illustrate the restrictions on the accurate reporting rate, z, consider the case of contaminated sampling where $\gamma = 1$ and $p = 0.3$. Based on Assumptions 2 and 3 alone, Lemma 1 reveals that $z \in [0, 0.3] \cup [0.7, 1]$. A lower bound accurate reporting rate, v, provides additional information. As noted earlier, studies assessing measurement error in binary variables often assume $z > \frac{1}{2}$. In that case, Assumption 1 rules out any values of $z \leq \frac{1}{2}$, while Assumptions 2 and 3 rule out values between 0.3 and 0.7. Thus, Assumptions 1-3 imply that at least 70 percent of the data are correctly classified. The commonplace restriction that more than half the data are correctly classified, $z > \frac{1}{2}$, can be sharpened – sometimes quite substantially – under this response error mixture model.

3.2 Bounding the Outcome Distribution

Now that we can identify the set of feasible candidates for z, it remains to identify the possible range of w for each feasible value of z. To do this, we need to characterize the behavior of the function $w(z)$ across different values of γ and p. In particular, $w(z)$ is weakly concave in $\left(-\infty, \frac{\gamma}{\gamma+1}\right)$ and convex in $\left(\frac{\gamma}{\gamma+1}, \infty\right)$, or vice versa, depending on the values γ

and p. Moreover, the local extrema of $w(z)$, which play a role in defining the bounds, are sensitive to these parameters.

Before presenting the formal results characterizing $w(z)$, it is instructive to visualize the shape of this function across several different parameter values. Figures 2 and 3 depict the identification regions for different values of γ, p, and v. Figure 2 considers the special case $\gamma = 1$ (pure contaminated sampling) for the values $p = 0.3$ and $p = 0.7$. For comparison, we also depict the corrupt and contaminated sampling bounds. The horizontal axis now depicts the unknown fraction of draws, z, from the distribution of interest, F. Values of z lying between the vertical dotted lines are ruled out by Lemma 1.

Under the response error model assumption, $w(z)$ traces out the true prevalence rate as a function of z. When $z = 1$, the true prevalence rate equals the reported prevalence rate: $w = p$. At the other extreme when $z = 0$, all classifications are inaccurate and $w = 1 - p$. Most importantly, notice that the shape of the outcome distribution function $w(z)$ depends on the fraction of respondents reporting in the affirmative, p. When $p = 0.7$ (0.3), for example, the outcome distribution decreases (increases) in z for feasible values of z outside $[0.3, 0.7]$.

For $\gamma \neq 1$ in Figure 3, it is useful to study the behavior of $w(z)$ when (1) $\gamma - 1$ and $p - \frac{1}{\gamma+1}$ have the same sign, and (2) $\gamma - 1$ and $p - \frac{1}{\gamma+1}$ have opposite signs. The latter case is more complicated because $w(z)$ may exhibit local extrema within $z \in (0, 1)$. Figure 3A depicts the first case when both signs are negative: $\{\gamma \leq 1, p < \frac{1}{\gamma+1}\}$. As in Figure 2, $w(z)$ is monotonic in z for values of z not ruled out by Lemma 1. Specifically, the outcome distribution increases over contiguous feasible ranges of z. The figure is analogous for the case that both signs are positive, $\{\gamma \geq 1, p > \frac{1}{\gamma+1}\}$ (not shown), except the outcome distribution decreases in z. When $\{\gamma < 1, p > \frac{1}{\gamma+1}\}$ as depicted in Figure 3B, $w(z)$ is not monotonic in z and, as such, the bounds may be sensitive to interior extrema. Define z_1 and z_2 as the values of z that minimize and maximize, respectively, the function $w(z)$. Then for $\gamma < 1$ and $p > \frac{1}{\gamma+1}$, $w(z)$

is increasing within $(0, z_2]$ and decreasing within $\left[z_2, \frac{\gamma}{\gamma+1}\right)$, while decreasing within $\left(\frac{\gamma}{\gamma+1}, z_1\right]$ and increasing within $[z_1, 1)$. The figure is analogous for the case $\{\gamma > 1, p < \frac{1}{\gamma+1}\}$.

These characterizations about the shape of the outcome distribution $w(z)$ are summarized in the Appendix as Lemma 2. Most importantly, this lemma reveals that the function $w(z)$ is not monotonic in z when $(p - \frac{1}{\gamma+1})$ and $(\gamma - 1)$ have opposite signs, in which case $w(z)$ exhibits interior extrema lying within $[\min\{p, 1-p\}, \max\{p, 1-p\}]$; otherwise, $w(z)$ is monotonic within each of the two regions.

Given Lemmas 1 and 2, we can identify the set of feasible candidates for z (Lemma 1) and characterize the shape of the outcome distribution function, $w(z)$ (Lemma 2). It still remains to identify the possible range of w for each feasible value of z. For illustration, it is again instructive to begin with the representative Figures 2 and 3. When z is known to exceed some value v, we can bound the outcome distribution by characterizing all feasible values of w associated with $z \geq v$. Consider the special case $\gamma = 1$ (pure contaminated sampling) for the values $p = 0.3$ (Figure 2A) and suppose $v = 0.9$. Then w can take any value between $\frac{v-(1-p)}{2v-1} = 0.25$ and $p = 0.3$. For sufficiently small values of v, the identification regions under the response error model become disjoint. If $v = 0.2$, for example, then values of w between $\frac{v-(1-p)}{2v-1} = 0.83$ and 1 become possible in addition to the values between 0 and 0.3. The same procedures are used to bound the outcome distribution when γ does not equal 1, although there may be additional complications introduced by the local extrema. In particular, when the outcome distribution is not monotonic in z, the value of z associated with an extremum may lie in the interior of the feasible range.

To formalize these ideas, we combine the results in Lemmas 1 and 2 to derive sharp identification regions for w as a function of γ, p, and v:

Proposition 2 (response error mixture model with multiplicative mean independence). *Suppose Assumptions 1-3 hold. Define $P_k \equiv w(k)$ for $k \neq \frac{\gamma}{\gamma+1}$ and $P_k \equiv p$ otherwise. Then for $\gamma = 0$, w must lie within $[\max\{0, p-(1-v)\}, p]$. For $p = \frac{1}{\gamma+1}$, w must lie*

within $[\min\{P_v, p\}, \max\{P_v, p\}]$ *when* $v > \frac{\gamma}{\gamma+1}$ *and within* $\left[0,\ \max\left\{\min\left\{\frac{2}{\gamma+1}, \frac{2\gamma}{\gamma+1}\right\},\ \max\{P_v, p\}\right\}\right]$ *when* $v \leq \frac{\gamma}{\gamma+1}$. *Otherwise,* w *is constrained to lie within the following regions:*

Case A: $p < \frac{1}{\gamma+1}$

$$w \in \begin{cases} [0, P_{z_2}] \cup [P_{z_1}, \max\{P_v, P_\delta\}] & \text{if } v \leq z_1 \\ [0, P_{z_2}] \cup [P_v, P_\delta] & \text{if } z_1 < v \leq \delta \\ [0, P_{z_2}] & \text{if } \delta < v < 1-p \\ [\min\{p, P_v\}, P_{z_2}] & \text{if } 1-p \leq v < z_2 \\ [p, P_v] & \text{if } v \geq z_2 \end{cases}$$

Case B: $p > \frac{1}{\gamma+1}$

$$w \in \begin{cases} [0, P_{z_2}] \cup [P_{z_1}, \max\{p, P_\delta\}] & \text{if } v \leq z_2 \\ [0, P_v] \cup [P_{z_1}, \max\{p, P_\delta\}] & \text{if } z_2 < v \leq 1-p \\ [P_{z_1}, \max\{p, P_\delta\}] & \text{if } 1-p < v < \delta \\ [P_{z_1}, \max\{p, P_v\}] & \text{if } \delta \leq v < z_1 \\ [p, P_v] & \text{if } v \geq z_1 \end{cases}$$

where $\delta = \begin{cases} \max\left\{0, \frac{\gamma-(1-p)}{\gamma}\right\} & \text{if } 0 < \gamma \leq 1 \\ \min\{\gamma p, 1\} & \text{if } \gamma > 1, \end{cases}$

$$z_1 \equiv \begin{cases} 0 & \text{if } p < \frac{1}{\gamma+1} \text{ and } \gamma \leq 1 \\ \min\{\delta, \max\{0, z_a\}\} & \text{if } p < \frac{1}{\gamma+1} \text{ and } \gamma > 1 \\ \max\{\delta, \min\{1, z_a\}\} & \text{if } p > \frac{1}{\gamma+1} \text{ and } \gamma < 1 \\ 1 & \text{if } p > \frac{1}{\gamma+1} \text{ and } \gamma \geq 1, \end{cases}$$

$$z_2 \equiv \begin{cases} 1 & \text{if } p < \frac{1}{\gamma+1} \text{ and } \gamma \leq 1 \\ \min\{1, z_b\} & \text{if } p < \frac{1}{\gamma+1} \text{ and } \gamma > 1 \\ \max\{0, z_b\} & \text{if } p > \frac{1}{\gamma+1} \text{ and } \gamma < 1 \\ 0 & \text{if } p > \frac{1}{\gamma+1} \text{ and } \gamma \geq 1, \end{cases}$$

and (z_a, z_b) *are defined in Lemma 2 (see the appendix).*

A proof is provided in the appendix. If the researcher believes that γ lies in some range $[\gamma_L, \gamma_H]$, then the relevant identification regions are obtained by taking the union of the above regions across possible values of γ.

In the special case that $\gamma = 1$, the response error mixing model bounds in Proposition 2 simplify as follows:

Corollary 2: *Suppose $\gamma = 1$. When $p = \frac{1}{2}$, the prevalence rate $P(W = 1)$ equals $\frac{1}{2}$ for $v > \frac{1}{2}$ and is unconstrained otherwise. For $p \neq \frac{1}{2}$, $P(W = 1)$ is constrained to lie in the following regions:*

$$P(W=1) \in \begin{cases} [0,p] \cup \left[\frac{v-(1-p)}{2v-1}, 1\right] & \text{if } 0 \leq v \leq p \\ [0,p] & \text{if } p < v < 1-p \quad \text{for } p < \frac{1}{2} \\ \left[\frac{v-(1-p)}{2v-1}, p\right] & \text{if } 1-p \leq v \leq 1 \end{cases} \tag{14}$$

$$P(W=1) \in \begin{cases} \left[0, \frac{v-(1-p)}{2v-1}\right] \cup [p,1] & \text{if } 0 \leq v \leq 1-p \\ [p,1] & \text{if } 1-p < v < p \quad \text{for } p > \frac{1}{2} \\ \left[p, \frac{v-(1-p)}{2v-1}\right] & \text{if } p \leq v \leq 1. \end{cases}$$

Using different approaches, Molinari (2008) and Kreider (2007) independently derive these regions in the special case where $\gamma = 1$. Kreider and Pepper (2008) provide a simpler derivation that covers cases involving $\gamma = 1$ and $v > 0.5$.

There are three notable features of these bounds. First, they are tighter than the HM bounds under corrupt sampling. Consider, for example, the case where $\frac{1}{2} \leq p \leq v$ with $\gamma = 1$. Under corrupt sampling, the outcome distribution is known to exceed $p - (1 - v)$ whereas the lower bound increases to p under the response error mixing model. Second, for sufficiently low values of v the range of the identification region is not contiguous. Finally, in many situations the bounds are informative even when there is no prior information on the degree of accurate reporting ($v = 0$).

Consider the case depicted in Figure 2A where $\gamma = 1$ and $p = 0.3$. When $v = 0$, the prevalence rate w cannot lie within $(0.3, 0.7)$. In contrast, the data reveal nothing about w under corrupt or contaminated sampling when $v \leq 0.3$. Likewise, when $\gamma = 0.8$ and $p = 0.3$ (Figure 3A), the identification regions are informative even when $v = 0$. In that case, w must lie within $[0, 0.3] \cup [0.7, 0.825]$, implying that w cannot lie within $(0.3, 0.7)$ or within $(0.825, 1]$. In contrast, the Proposition 1 bounds only constrain w to lie within $[0, 0.86]$, and

the corrupt sampling bounds are uninformative. More generally, the response error mixture model bounds are considerably tighter than the Proposition 1 bounds across most values of v.

Overall, we find that the response error mixture model with multiplicative mean independence confers substantial identifying power. The Proposition 2 bounds are always more informative than the corrupt sampling bounds, are tighter than the Proposition 1 bounds across most values of v, and generally lie strictly inside the unit interval even when there is no information on the degree of accurate reporting (i.e., when $v = 0$).

4 Illustration

To illustrate the response error mixture model with multiplicative mean independence, we consider the problem of drawing inferences on the rate of illicit drug use in the presence of nonrandom reporting errors. Self-reported survey data on deviant behavior inevitably yield some inaccurate responses. Respondents concerned about the legality of their behavior may falsely deny consuming illicit drugs, while the desire to fit into a deviant culture or otherwise be defiant may lead some respondents to falsely claim to consume illicit drugs (see, e.g., Pepper, 2001).

To draw inferences on the prevalence of illicit drug use, we use self-reported data from the 2002 National Household Survey of Drug Use and Health (NHSDH). The top row in Table 1 displays the basic sample prevalence rates used in the analysis. In particular, 54% of 18-24 year-olds claimed to have consumed marijuana within their lifetimes with 30% reporting use during the last year. The corresponding rates for cocaine are 15% and 7% (Office of Applied Studies, 2003).

To draw inferences about true rates of illicit drug use in the U.S., one must combine these self-reports with assumptions about the nature and extent of reporting errors. There does, in fact, exist some information on response errors in drug use questionnaires. Harrison (1995),

for example, compares self-reported marijuana and cocaine use during the past three days to urinalysis test results for the same period among a sample of arrestees. That study reveals a 22% misreporting rate for marijuana consumption ($z = 0.777$) and a 27% misreporting rate for cocaine ($z = 0.730$). As expected, the outcome probability among misreporters is higher than the outcome probability among accurate reporters: γ equals 2.90 for marijuana and 2.61 for cocaine.

We use results from Harrison's (1995) validation study to help identify true rates of illicit drug use in the general population of young adults. In making inferences about true drug use rates, we consider the identifying power of several sets of assumptions. Throughout, we maintain the assumption that the accurate reporting rate, z, in the general noninstitutionalized population exceeds that obtained in the sample of arrestees studied by Harrison (1995). Presumably, arrestees have a relatively high incentive to misreport (Harrison, 1995; Pepper, 2001). Under this restriction alone, the HM corrupt sampling bounds reveal much uncertainty about the true drug use rates. For example, we only learn that between 32% and 76% of the young adult population has ever used marijuana.

When the lower bound accurate reporting rate is coupled with the HM contamination assumption (Proposition 1 with $\gamma = 1$), the bounds narrow considerably (Frame A). For lifetime marijuana use, for example, the bounds narrow from [32%, 76%] to [41%, 69%], a 36 percent reduction in the width of the bounds. When we additionally impose the Assumption 3 response error mixture model (Frame B), the lifetime marijuana use rate is nearly point-identified, lying in the narrow range [54%, 57%].

While powerful, the identifying assumption that drug use rates are identical among accurate and inaccurate reporters ($\gamma = 1$) seems implausible. More realistically, the rate of illicit drug use is higher among inaccurate reporters. Imposing the arguably innocuous assumption that $\gamma \geq 1$, we can identify that the lifetime rate of marijuana use lies within [54%, 76%], a 50 percent reduction in the range of uncertainty compared with the HM corrupt sampling

bounds. For cocaine use, the restriction $\gamma \geq 1$ confers no identifying power compared with the corrupt sampling case.

Table 1. Bounds on the Fractions of 18-24 Year-olds Using Marijuana and Cocaine

	Marijuana, Past Year	Marijuana, Lifetime	Cocaine, Past Year	Cocaine, Lifetime
Reported Use, $P(X=1)^a$	0.30	0.54	0.07	0.15
HM Corrupt Sampling $(z > z_v)^b$	$[0.08, 0.52]$	$[0.32, 0.76]$	$[0.0, 0.34]$	$[0.0, 0.42]$
A. Multiplicative Mean Independence				
HM Contaminated Sampling, $\gamma = 1$	$[0.10, 0.39]$	$[0.41, 0.69]$	$[0.0, 0.10]$	$[0.0, 0.21]$
Multiplicative, $\gamma \geq 1$	$[0.10, 0.52]$	$[0.41, 0.76]$	$[0.0, 0.34]$	$[0.0, 0.42]$
Multiplicative, $\gamma = \gamma_v$ ^{b}	$[0.14, 0.49]$	0.54	$[0.0, 0.14]$	$[0.0, 0.29]$
B. Response Error Mixture Model with Multiplicative Mean Independence				
HM Contaminated Sampling, $\gamma = 1$ + Assumption 3	$[0.14, 0.30]$	$[0.54, 0.57]$	$[0.0, 0.07]$	$[0.0, 0.15]$
Multiplicative, $\gamma \geq 1$ + Assumption 3	$[0.14, 0.52]$	$[0.54, 0.76]$	$[0.0, 0.34]$	$[0.0, 0.42]$
Multiplicative, $\gamma = \gamma_v$ ^{b} + Assumption 3	$[0.30, 0.43]$	0.54	$[0.0, 0.07]$	$[0.0, 0.15]$

[a] Office of Applied Studies, 2003
[b] z_v and γ_v are the relevant values of z and γ from the validation studies discussed in the text.

Instead, one might assume that values of γ found in Harrison's (1995) validation study of arrestees apply to the NHSDH sample. For cocaine consumption, imposing the value $\gamma = 2.6$ substantially reduces the range of uncertainty about its rate of use. The HM bounds only reveal that the rate of prior-year cocaine consumption lies between 0% and 34%. When $\gamma = 2.6$ under multiplicative mean independence Assumption 2, the upper bound falls to

14%, nearly a 60 percent reduction. This upper bound falls further to 7% – nearly an 80 percent reduction – after additionally imposing the response error mixture model Assumption 3. These results for $\gamma = 2.6$ are very close to those under the untenable pure contamination assumption with $\gamma = 1$. So, in this application, multiplicative mean independence has allowed us to substantially reduce uncertainty about the parameter without requiring an assumption that is nearly certain to be invalid.

A useful practical feature of Propositions 1 and 2 is that we can assess the sensitivity of the bounds to variation in (γ, z_v). After all, Harrison's estimates of misreporting may be in error due to sampling variability and because her validation study might not accurately reflect misreporting rates in the general population. In sensitivity analysis, we considered how the Proposition 1 and 2 bounds vary with $\gamma \in [0, 4]$ over the four different outcome measures.

The results are traced out in Figure 4 for marijuana use (figures for cocaine use available upon request). The most striking results involve the bounds on incidence of lifetime marijuana consumption. Under the response error Assumption 3, we see from Figure 4B that if $\gamma \geq 1.85$ the prevalence rate of lifetime use is identified to equal the self-reported rate of 0.54. Thus, if $\gamma = 2.9$, as revealed by Harrison, it follows that the prevalence rate among the general population is identified using data from the NHSDH. Moreover, this finding holds for all $z_v \in [0.70, 0.90]$. Being able to point-identify lifetime marijuana consumption follows from the Lemma 1 bounds which reveal that for the range of parameters that apply in this setting (i.e., $z > 0.5$, $\gamma \geq 2$, and $p > 0.5$) everyone reports accurately. In fact, many researchers believe that measures of lifetime use are much less prone to reporting errors than shorter run measures (e.g., Harrison, 1995). For cocaine use, the Proposition 1 upper bound increases slightly with γ whereas the Proposition 2 bounds do not vary. Likewise, when assessing how the bounds vary over $z_v \in [0.70, 0.90]$, we find that the Proposition 1 upper bound on the

probability of prior-year cocaine consumption falls from 15% to 9%, whereas the Proposition 2 upper bound does not vary.

5 Conclusion

In the contaminated sampling model studied by Horowitz and Manski (1995), the assumption that the outcome distribution is independent of the mixing process has substantial identifying power. In many applications, however, this independence assumption is untenable. Yet when the independence assumption is discarded, the resulting bounds tend to be frustratingly wide. In this paper, we introduce a general notion of a response error mixture model with multiplicative mean independence, with Propositions 1 and 2 characterizing the identifying power of these assumptions. Under these assumptions, we often find informative identification regions even when there is no prior information on the degree of accurate reporting. Moreover, we find that these assumptions can be easy to motivate and apply. Considering inference on the use of illicit drugs, our empirical illustration reveals that the multiplicative mean independence assumption can be credible and informative in environments where the pure contamination assumption is controversial.

Given the long-standing struggle to credibly address inferential problems that arise from response errors, we are hopeful that this nonparametric bounding framework can be usefully applied and extended. It is easy to think of variations on this theme that warrant study. For example, an interesting possibility might be to extend the idea of contaminated instruments used to evaluate treatment effects, as introduced by Hotz, Mullins, and Sanders (1997), to the case of multiplicative mean independence.

References

[1] Beresteanu, A., and Molinari, F. (2008), "Asymptotic Properties for a Class of Partially Identified Models," *Econometrica*, 76, 763-814.

[2] Bollinger, C. (1996), "Bounding Mean Regressions When a Binary Variable is Mismeasured," *Journal of Econometrics*, 73, 387-99.

[3] Bound, J., Brown, C., and Mathiowetz, N. (2001), "Measurement Error in Survey Data," In J. Heckman and E. Leamer (Eds.), Handbook of Econometrics, 5, Ch. 59, 3705-3843.

[4] Dominitz, J. and Sherman, R. (2004), "Sharp Bounds Under Contaminated or Corrupted Sampling With Verification, with an Application to Environmental Pollutant Data," *Journal of Agricultural, Biological, and Environmental Statistics*, 9, 319-338.

[5] Frazis, H. and Loewenstein, M. (2003), "Estimating Linear Regressions with Mismeasured, Possibly Endogenous, Binary Explanatory Variables," *Journal of Econometrics*, 117, 151-178.

[6] Harrison, L. D. (1995), "The Validity of Self-Reported Data on Drug Use," *Journal of Drug Issues*, 25, 91-111.

[7] Hirsch, B.T. and Schumacher, E.J. (2004), "Match Bias in Wage Gap Estimates Due to Earnings Imputation," *Journal of Labor Economics*, 22, 689-722.

[8] Horowitz, J. and Manski, C. (1995), "Identification and Robustness with Contaminated and Corrupted Data," *Econometrica* 63, 281-02.

[9] Huber, P. (1981), *Robust Statistics*. New York: Wiley.

[10] Hotz, J., Mullins, C. and Sanders, S. (1997), "Bounding Causal Effects Using Data from a Contaminated Natural Experiment: Analyzing the Effects of Teenage Childbearing," *Review of Economic Studies*, 64, 575-603.

[11] Imbens, G. and Manski, C. (2004), "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845-57.

[12] Kreider, B. and Pepper, J. (2007), "Disability and Employment: Reevaluating the Evidence in Light of Reporting Errors," *Journal of the American Statistical Association*, 102, 432-441.

[13] ____ and ____ (2008), "Inferring Disability Status from Corrupt Data," *Journal of Applied Econometrics*, 23, 329-49.

[14] Kreider, B. (2007), "Partially Identifying the Prevalence of Health Insurance Given Contaminated Sampling Response Error," Working Paper, Iowa State University.

[15] Lambert, D. and Tierney, L. (1997), "Nonparametric Maximum Likelihood Estimation from Samples with Irrelevant Data and Verification Bias," *Journal of the American Statistical Association*, 92, 937-944.

[16] Molinari, F. (2008), "Partial Identification of Probability Distributions with Misclassified Data," *Journal of Econometrics*, 144, 81-117.

[17] Mullin, C.H., (2005), "Identification and Estimation with Contaminated Data: When do Covariate Data Sharpen Inference?" *Journal of Econometrics,* 130, 253-272.

[18] Office of Applied Studies (2003), Results from the 2002 National Survey on Drug Use and Health: Summary of National Finding (DHHS Publication No. SMA 03-3836, NHSDA Series H-22). Rockville, MD: Substance Abuse and Mental Health Services Administration.

[19] Pepper, J.V. (2001), "How Do Response Problems Affect Survey Measurement of Trends in Drug Use?" in C. F. Manski, J. V. Pepper, and C. Petrie eds., Informing America's Policy on Illegal Drugs: What We Don't Know Keeps Hurting Us, National Academy Press, Washington, D.C., 321-48.

[20] Rosen, A.M. (2008), "Confidence Sets for Partially Identified Parameters that Satisfy a Finite Number of Moment Inequalities." *Journal of Econometrics*, 146, 107-117.

[21] Stoye, J. (forthcoming), "More on Confidence Intervals for Partially Identified Parameters." *Econometrica.*

6 Appendix

Proof of Proposition 1.

If $\gamma \geq 1$, both terms of the upper bound in Equation (6) monotonically decrease with the accurate reporting rate, z. Thus, a closed form representation can be found by evaluating the mean outcome at the lowest possible value of z subject to the constraint implied by Assumption 2 that $\frac{k_0}{\gamma} \leq E(W|Z=1) \leq \frac{k_1}{\gamma}$ for $z \in [v, 1)$. When the conditions in (7) are satisfied, we have $E(W|Z=1) \leq \min\left\{E\left(X|X > \tau_X(1-v)\right), \frac{k_1}{\gamma}\right\}$ using Equations (3) and (4). Thus, $E(W) \leq \max(E(X), UB(v))$. For the lower bound when $\gamma \geq 1$, there is no closed form bound that applies for all distributions (except when $\gamma = 1$) because the two terms in Equation (6) move in opposite directions with z.

Analogously, if $\gamma \leq 1$ then both terms of the lower bound in Equation (6) increase with z. A closed form representation can be found by evaluating the mean outcome at the lowest possible value of the accurate reporting rate subject to the constraint implied by Assumption 2 that $\frac{k_0}{\gamma} \leq E(W|Z=1) \leq \frac{k_1}{\gamma}$ for $z \in [v, 1)$. When the conditions in (7) are satisfied, we have $E(W|Z=1) \geq \max\left\{E\left(X|X \leq \tau_X(v)\right), \frac{k_0}{\gamma}\right\}$ using Equations (3) and (4). Thus, $E(W) \geq \min(E(X), LB(v))$. For the upper bound when $\gamma \leq 1$, there is no closed form bound that applies for all distributions (except when $\gamma = 1$) because the two terms in Equation (6) move in opposite directions with z.

Finally, if the conditions in (7) are not satisfied, $z \in [v, 1)$ is not feasible. In that case, $z = 1$ and $E(W) = E(X)$. $\square$

Corollary 1. For binary outcomes, $k_0 = 0$ and $k_1 = 1$. The HM bounds in Equation (3) become $\frac{p-\min\{1-v,p\}}{v} \leq P(W=1|Z=1) \leq \frac{p-\max\{p-v,0\}}{v}$, and the γ–constraint restrictions in (7) become $0 \leq \frac{p-\max\{p-v,0\}}{v}$ and $\frac{1}{\gamma} \geq \frac{p-\min\{1-v,p\}}{v}$. The first γ–constraint inequality is always satisfied. When the second inequality is also satisfied, we have $P(W=1|Z=1) \leq \min\left\{\frac{p-\max\{p-v,0\}}{v}, \frac{1}{\gamma}\right\}$ using the HM upper bound and Assumption 2 restriction that $\gamma P(W=1|Z=1) \leq 1$.

First consider the upper bound on $P(W = 1)$. The Equation (6) upper bound as a function of z is given by $UB(z) \equiv \min\left\{\frac{p-\max\{p-z,0\}}{z}, \frac{1}{\gamma}\right\}[1 + (\gamma - 1)(1 - z)]$.

Case (A): $\gamma \leq 1$. In this case, $UB(z) = \frac{p-\max\{p-z,0\}}{z}[1 + (\gamma - 1)(1 - z)]$, and the γ–constraint is satisfied since $\frac{1}{\gamma} \geq 1$ and $\frac{p-\min\{1-v,p\}}{v} \leq 1$. When $v \geq p$, we know $z \geq p$; then $UB(z) = \frac{p}{z}[1 + (\gamma - 1)(1 - z)]$, which is decreasing in z (the derivative is $-\gamma p/z^2$). Setting $z = \max\{p, v\} = v$, its minimum feasible value, the upper bound is $UB(v)$(which does not exceed 1 since $\frac{p}{v} \leq 1$ and $K \leq 1$ when $\gamma \leq 1$). When $v < p$, z may be either smaller or larger than p. In the former case, $UB(z) = 1 + (\gamma - 1)(1 - z)$ which is increasing in z. Setting $z = p$ yields $UB(p) = 1 + (\gamma - 1)(1 - p) = p + \gamma(1 - p)$ (which is no greater than 1 since $\gamma \leq 1$). In the latter case, $UB(z) = \frac{p}{z}[1 + (\gamma - 1)(1 - z)]$ which is decreasing in z. Setting $z = \max\{p, v\} = p$ again yields $UB(p) = p + \gamma(1 - p)$. Finally, if $v \geq p$, $UB(v) = \frac{p}{v}[1 + (\gamma - 1)(1 - v)] \leq p + \gamma(1 - p)$ and if $v < p$, $p + \gamma(1 - p) \leq UB(v)$.

Case (B): $\gamma > 1$. Since $\frac{1}{\gamma} \leq 1$, we can write $UB(z) = \min\left\{\frac{p}{z}, \frac{1}{\gamma}\right\}[1 + (\gamma - 1)(1 - z)]$ which is decreasing in z. That is, we can ignore the case that $\frac{p-\max\{p-z,0\}}{z} = 1$. From Proposition 1, we know that the upper bound equals $UB(v)$. In this case, when $\gamma \geq 1$, $UB(v) \leq 1 \leq p + \gamma(1 - p)$. Second, consider the lower bound when $\gamma > 1$. The function $LB(z) = \frac{p-\min\{1-z,p\}}{z}(1 + [1 + (\gamma - 1)(1 - z)])$ is concave in $z \in [0, 1]$, and bounded between 0 when $z \in [0, 1 - p]$ and p for $z = 1$. Moreover, $z \in [\frac{\gamma(1-p)}{\gamma-1}, 1)$ violate the restrictions in Equation (7); the HM lower bound $\frac{p-\min\{1-z,p\}}{z}$ exceeds $\frac{1}{\gamma}$. Thus, the lower bound is attained by setting $z = v$ if the restriction in (7) holds and $z = 1$ otherwise. $\square$

Proof of Lemma 1: The fraction of false positives is $\theta^+ = P(X=1, Z=0) = \frac{(z-\gamma p)(1-z)}{(\gamma+1)z-\gamma}$, and the fraction of false negatives is $\theta^- = P(X=0, Z=0) = \frac{\gamma[z-(1-p)](1-z)}{(\gamma+1)z-\gamma}$. If $\gamma = 0$, $w(z) = z-(1-p)$, so that $z \geq 1-p$. If $\gamma > 0$, we have three cases:

Case (i) $z < \frac{\gamma}{\gamma+1}$: Then (a) $\theta^+ \geq 0 \Rightarrow (z-\gamma p)(1-z) \leq 0 \Rightarrow z \leq \gamma p$; (b) $\theta^+ \leq p \Rightarrow \frac{(z-\gamma p)(1-z)}{(\gamma+1)z-\gamma} \leq p \Rightarrow 1-z-p \geq 0 \Rightarrow z \leq 1-p$; (c) $\theta^- \geq 0 \Rightarrow \gamma[z-(1-p)](1-z) \leq 0 \Rightarrow z \leq 1-p$; or (d) $\theta^- \leq 1-p \Rightarrow \frac{\gamma[z-(1-p)](1-z)}{(\gamma+1)z-\gamma} \leq 1-p \Leftrightarrow z(\gamma z-\gamma+1-p) \geq 0 \Rightarrow z \leq \max\left\{0, \frac{\gamma-(1-p)}{\gamma}\right\}$if $\gamma > 0$ and $z \geq 0$ if $\gamma = 0$.

Case (ii) $z > \frac{\gamma}{\gamma+1}$: Then (a) $\theta^+ \geq 0 \Rightarrow (z-\gamma p)(1-z) \geq 0 \Rightarrow \{z \geq \gamma p \text{ or } z = 1\} \Rightarrow z \geq \min\{\gamma p, 1\}$; (b) $\theta^+ \leq p \Rightarrow \frac{(z-\gamma p)(1-z)}{(\gamma+1)z-\gamma} \leq p \Rightarrow 1-z-p \leq 0 \Rightarrow z \geq 1-p$; or (c) $\theta^- \geq 0 \Rightarrow \frac{\gamma[z-(1-p)](1-z)}{(\gamma+1)z-\gamma} \geq 0 \Rightarrow \{\gamma = 0 \text{ or } z \geq 1-p\} \Rightarrow z \geq 1-p$ if $\gamma > 0$ and $z \geq 0$ if $\gamma = 0$. (d) $\theta^- \leq 1-p \Rightarrow \frac{\gamma(z-(1-p))(1-z)}{(\gamma+1)z-\gamma} \leq 1-p \Rightarrow \gamma z-\gamma+1-p \geq 0 \Rightarrow z \geq \frac{\gamma-(1-p)}{\gamma}$ if $\gamma > 0$ and $z \geq 0$ if $\gamma = 0$.

Case (*iii*) $z = \frac{\gamma}{\gamma+1}$: Then (9) implies that $p = 1-z = \frac{1}{\gamma+1}$. Thus, $z = \frac{\gamma}{\gamma+1} \Rightarrow p = \frac{1}{\gamma+1}$.

Combining these results leads to the stated restrictions on allowed values of z. □

Lemma 2.

(a) *For* $p < \frac{1}{\gamma+1}$ *and* $\gamma \leq 1$, $w(z)$ *is increasing within* $\left(-\infty, \frac{\gamma}{\gamma+1}\right)$ *and within* $\left(\frac{\gamma}{\gamma+1}, \infty\right)$. *Throughout, "increasing" means weakly increasing and "decreasing"means weakly decreasing. For* $p > \frac{1}{\gamma+1}$ *and* $\gamma \geq 1$, $w(z)$ *is decreasing within* $\left(-\infty, \frac{\gamma}{\gamma+1}\right)$ *and within* $\left(\frac{\gamma}{\gamma+1}, \infty\right)$. **(b)** *For* $p < \frac{1}{\gamma+1}$ *and* $\gamma > 1$, $w(z)$ *is decreasing within* $(-\infty, z_a]$ *and increasing within* $\left[z_a, \frac{\gamma}{\gamma+1}\right)$ *where* $w(z)$ *has zero slope at* $z_a \equiv \frac{\gamma}{\gamma+1} - \frac{1}{\gamma^2-1}\sqrt{2\gamma(\gamma^2-1)\left(\frac{1}{\gamma+1}-p\right)} \in \left(-\infty, \frac{\gamma}{\gamma+1}\right)$; $w(z)$ *is increasing within* $\left(\frac{\gamma}{\gamma+1}, z_b\right]$ *and decreasing within* $[z_b, \infty)$ *where* $w(z)$ *has zero slope at* $z_b \equiv \frac{\gamma}{\gamma+1} + \frac{1}{\gamma^2-1}\sqrt{2\gamma(\gamma^2-1)\left(\frac{1}{\gamma+1}-p\right)} \in \left(\frac{\gamma}{\gamma+1}, \infty\right)$. *Finally,* $w(z_a)$ *lies within* $(w(z_b), 1-p]$ *and* $w(z_b)$ *lies within* $[p, w(z_a))$.

(c) *For* $p > \frac{1}{\gamma+1}$ *and* $\gamma < 1$, $w(z)$ *is increasing within* $(-\infty, z_b]$ *and decreasing within* $\left[z_b, \frac{\gamma}{\gamma+1}\right)$ *where* $w(z)$ *has zero slope at* z_b; $w(z)$ *is decreasing within* $\left(\frac{\gamma}{\gamma+1}, z_a\right]$ *and increasing*

within $[z_a, \infty)$ *where* $w(z)$ *has zero slope at* z_b*. Finally,* $w(z_a)$ *lies within* $(w(z_b), p]$ *and* $w(z_b)$ *lies within* $[1-p, w(z_a))$*.*

Proof of Lemma 2: It is useful to begin by noting some facts about $w(z) = \frac{[\gamma+(1-\gamma)z][z-(1-p)]}{(\gamma+1)z-\gamma}$. Since $\frac{\partial w}{\partial z} = \frac{z^2(1-\gamma^2)+2z(\gamma^2-\gamma-p)+2\gamma(1-p-0.5\gamma)}{[(\gamma+1)z-\gamma]^2}$ and $\frac{\partial^2 w}{\partial z^2} = \frac{4\gamma[\gamma p-(1-p)]}{[(\gamma+1)z-\gamma]^3}$, it follows that: (f1) $\frac{\partial w}{\partial z}|_{z=0} \overset{s}{=} 2(1-p)-\gamma$ where $\overset{s}{=}$ means "has the same sign as;" (f2) $\frac{\partial w}{\partial z}|_{z=1} = 1-2p\gamma$; (f3) $\frac{\partial^2 w}{\partial z^2} \overset{s}{=} \frac{p-\frac{1}{\gamma+1}}{z-\frac{\gamma}{\gamma+1}}$; (f4) $\frac{\partial w}{\partial z} = 0$ at $z = z_a$ and $z = z_b$, which are real-valued if $p - \frac{1}{\gamma+1}$ and $\gamma - 1$ have opposite signs and imaginary otherwise; and (f5) $w(z_a) - w(z_b) = 4\sqrt{2}\frac{\sqrt{\gamma(\gamma-1)[1-p(\gamma+1)]}}{(1+\gamma)^2}$, which is real-valued positive if $p - \frac{1}{\gamma+1}$and $\gamma - 1$ have opposite signs and imaginary otherwise. Using (f1) and (f2), we see that the slope of $w(z)$ has the same sign as $2(1-p)-\gamma$ at $z = 0$ and has the same sign as $1 - 2\gamma p$ at $z = 1$. Using (f3), we learn that the second derivative has the same sign as $\left(p - \frac{1}{\gamma+1}\right) / \left(z - \frac{\gamma}{\gamma+1}\right)$ which reveals that $w(z)$ is convex (concave) if $p < (>) \frac{1}{\gamma+1}$ for $z < \frac{\gamma}{\gamma+1}$ and concave (convex) if $p < (>) \frac{1}{\gamma+1}$ for $z > \frac{\gamma}{\gamma+1}$.

For case (a), $p < \frac{1}{\gamma+1}$ and $\gamma \leq 1$ establish that the slope of $w(z)$ is positive at $z = 0$ and at $z = 1$ so that $w(z)$ is increasing-convex for $z < \frac{\gamma}{\gamma+1}$ and increasing-concave for $z > \frac{\gamma}{\gamma+1}$. Moreover, $p > \frac{1}{\gamma+1}$ and $\gamma \geq 1$ establishes that the slope of $w(z)$ is negative at $z = 0$ and at $z = 1$ so that $w(z)$ is decreasing-concave for $z < \frac{\gamma}{\gamma+1}$ and decreasing-convex for $z > \frac{\gamma}{\gamma+1}$.

For case (b), $p < \frac{1}{\gamma+1}$ and $\gamma > 1$ establishes that $w(z)$ is convex for $z < \frac{\gamma}{\gamma+1}$ with a local minimum at z_a and that $w(z)$ is concave for $z > \frac{\gamma}{\gamma+1}$ with a local maximum at z_b; $0 < \frac{\gamma}{\gamma+1}$ implies $w(z_a) \leq w(0) = 1-p$, while $\frac{\gamma}{\gamma+1} < 1$ implies $w(z_b) \geq w(1) = p$; and (f5) implies that $w(z_a) > w(z_b)$.

For case (c), $p > \frac{1}{\gamma+1}$ and $\gamma < 1$ establish that $w(z)$ is concave for $z < \frac{\gamma}{\gamma+1}$ with a local maximum at z_b and convex for $z > \frac{\gamma}{\gamma+1}$ with a local minimum at z_a; $0 \leq \frac{\gamma}{\gamma+1}$ implies $w(z_b) \geq w(0) = 1-p$, while $\frac{\gamma}{\gamma+1} < 1$ implies $w(z_a) \leq w(1) = p$; (f5) implies that $w(z_a) > w(z_b)$. $\square$

Proof of Proposition 2: For $\gamma = 0$, we have $w(z) = p - (1-z)$ so that w must lie within $[\max\{0, p-(1-v)\}, p]$. When $p = \frac{1}{\gamma+1}$ and $z = \frac{\gamma}{\gamma+1}$, (5) obtains $w = P(W =$

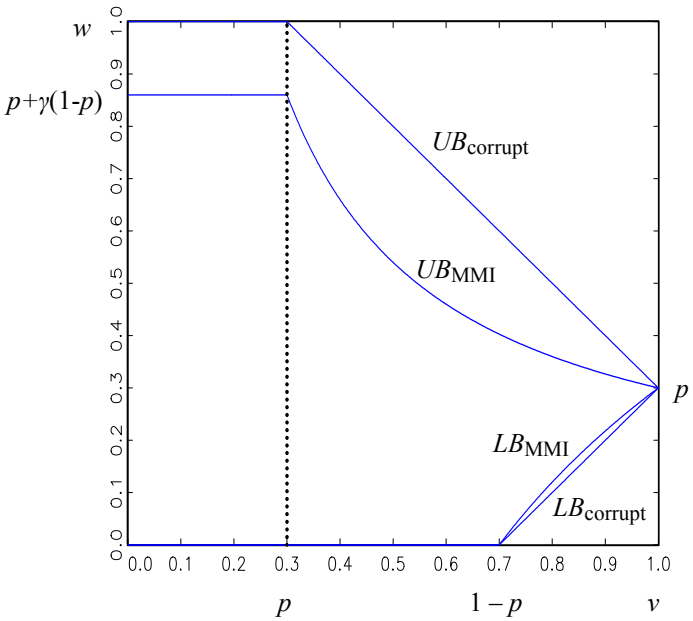
$1|Z = 1)\frac{2\gamma}{\gamma+1}$. By Assumption 2, $P(W = 1|Z = 1)$ cannot exceed $\min\left\{1, \frac{1}{\gamma}\right\}$. Varying $P(W = 1|Z = 1)$ within $\left[0, \min\left\{1, \frac{1}{\gamma}\right\}\right]$ reveals that $w \in \left[0, \min\left\{\frac{2}{\gamma+1}, \frac{2\gamma}{\gamma+1}\right\}\right]$ when $p = \frac{1}{\gamma+1}$ and $z = \frac{\gamma}{\gamma+1}$. When $p = \frac{1}{\gamma+1}$ and $z \neq \frac{\gamma}{\gamma+1}$, (11) reveals that $w(z) = 1 - p - (1 - 2p)\, z$. Since $w(z)$ is monotonic over z, we obtain $w \in [\min\{p, P_v\}, \max\{p, P_v\}]$ for $v > \frac{\gamma}{\gamma+1}$ and $w \in [\min\{p, P_v\}, \max\{p, P_v\}] \cup \left[0, \min\left\{\frac{2}{\gamma+1}, \frac{2\gamma}{\gamma+1}\right\}\right]$ for $v \leq \frac{\gamma}{\gamma+1}$. Otherwise, we have:

Case A: $p < \frac{1}{\gamma+1}$. Using Lemma 1, candidate values of z are confined to the space $z \in [v, 1] \cap \{S_A^1 \cup S_A^2\}$ where $S_A^1 = [0, \delta] \subset \left(-\infty, \frac{\gamma}{\gamma+1}\right)$ and $S_A^2 = [1 - p, 1] \subset \left(\frac{\gamma}{\gamma+1}, \infty\right)$. Using Lemma 2a and 2b, $w(z)$ is monotonically increasing within $\left(-\infty, \frac{\gamma}{\gamma+1}\right)$ and within $\left(\frac{\gamma}{\gamma+1}, \infty\right)$ if $\gamma \leq 1$. For $\gamma > 1$, $w(z)$ is declining within $(-\infty, z_a]$ before rising within $\left[z_a, \frac{\gamma}{\gamma+1}\right)$, and $w(z)$ is rising within $\left(\frac{\gamma}{\gamma+1}, z_b\right)$ before declining within (z_b, ∞). In the subset S_A^1, $w(z)$ declines from $w = 1 - p$ at $z = 0$ to its minimum value $w(z_1)$ at $z = z_1$ before rising to $w(\delta)$ at $z = \delta$ (where z_1 may be 0 or δ). In the subset S_A^2, $w(z)$ rises from $w = 0$ at $z = 1 - p$ to its maximum value $w = w(z_2)$ at $z = z_2$ and then declines to $w = p$ at $z = 1$. Lemma 2 establishes that any local extrema lie within $[\min\{p, 1 - p\}, \max\{p, 1 - p\}]$.

Case B: $p > \frac{1}{\gamma+1}$. Using Lemma 1, candidate values of z are confined to the space $z \in [v, 1] \cap \{S_B^1 \cup S_B^2\}$ where $S_B^1 = [0, 1 - p] \subset \left(-\infty, \frac{\gamma}{\gamma+1}\right)$ and $S_B^2 = [\delta, 1] \subset \left(\frac{\gamma}{\gamma+1}, \infty\right)$. Using Lemma 2a and 2c, $w(z)$ is monotonically decreasing within both $\left(-\infty, \frac{\gamma}{\gamma+1}\right)$ and $\left(\frac{\gamma}{\gamma+1}, \infty\right)$ if $\gamma \geq 1$. For $\gamma < 1$, $w(z)$ is increasing within $(-\infty, z_b]$ before rising within $\left[z_b, \frac{\gamma}{\gamma+1}\right)$ and is declining within $\left(\frac{\gamma}{\gamma+1}, z_a\right)$ before rising within (z_a, ∞). In the subset S_B^1, $w(z)$ increases from $w = 1 - p$ at $z = 0$ to its maximum value $w = w(z_2)$ at $z = z_2$ before declining to $w = 0$ at $z = 1 - p$. In the subset S_B^2, $w(z)$ declines from $w = w(\delta)$ at $z = \delta$ to its minimum value $w = w(z_1)$ at $z = z_1$ before rising to $w = p$ at $z = 1$. Lemma 2 establishes that any local extrema lie within $[\min\{p, 1 - p\}, \max\{p, 1 - p\}]$. $\square$

Figure 1. Multiplicative Mean Independence (MMI)

Case A: $\gamma < 1$: $\gamma = 0.8$, $p = 0.3$



Case B: $\gamma > 1$: $\gamma = 2$, $p = 0.3$

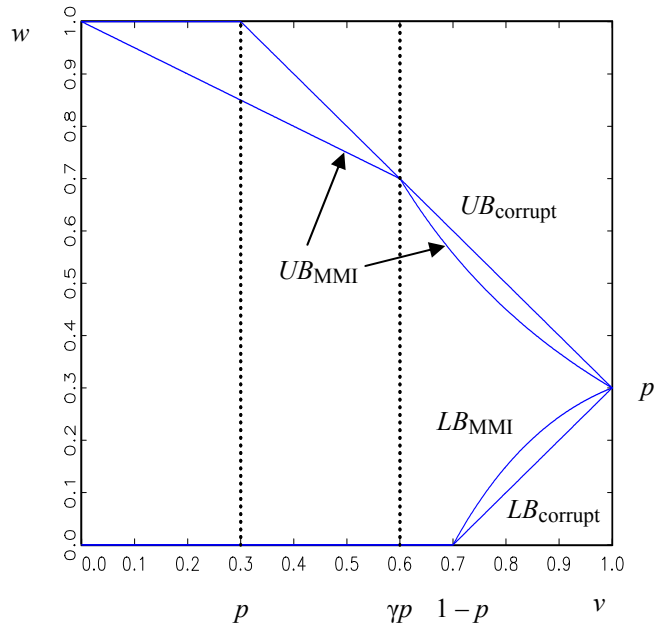
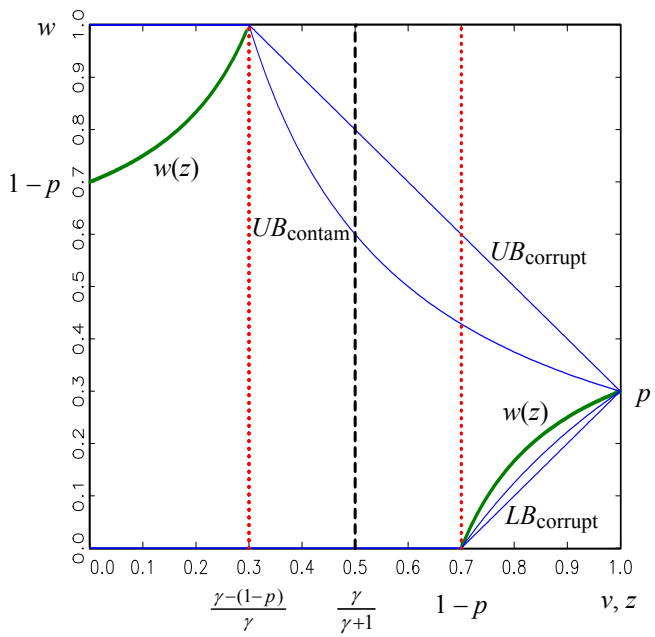
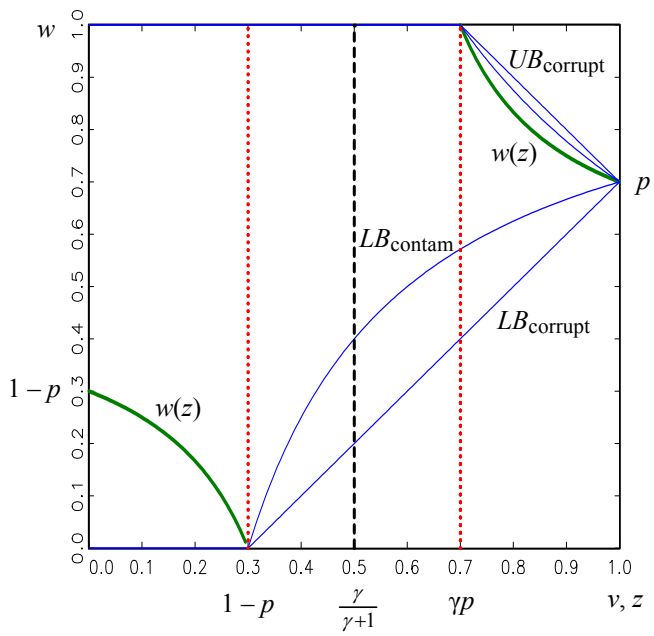


Figure 2. Response Error Mixture Model with Mean Independence

Case A: $\gamma = 1, p = 0.3 < ½$



Case B: $\gamma = 1, p = 0.7 > ½$



Note: $w(z)$ traces out w as a function of z, from which we can find bounds on w for a given v.

Figure 3. Response Error Mixture Model with Multiplicative Mean Independence

Case A: $\gamma < 1,\ p < \frac{1}{\gamma+1}$: $\gamma = 0.8,\ p = 0.3$

Case B: $\gamma < 1,\ p > \frac{1}{\gamma+1}$: $\gamma = 0.5,\ p = 0.7$

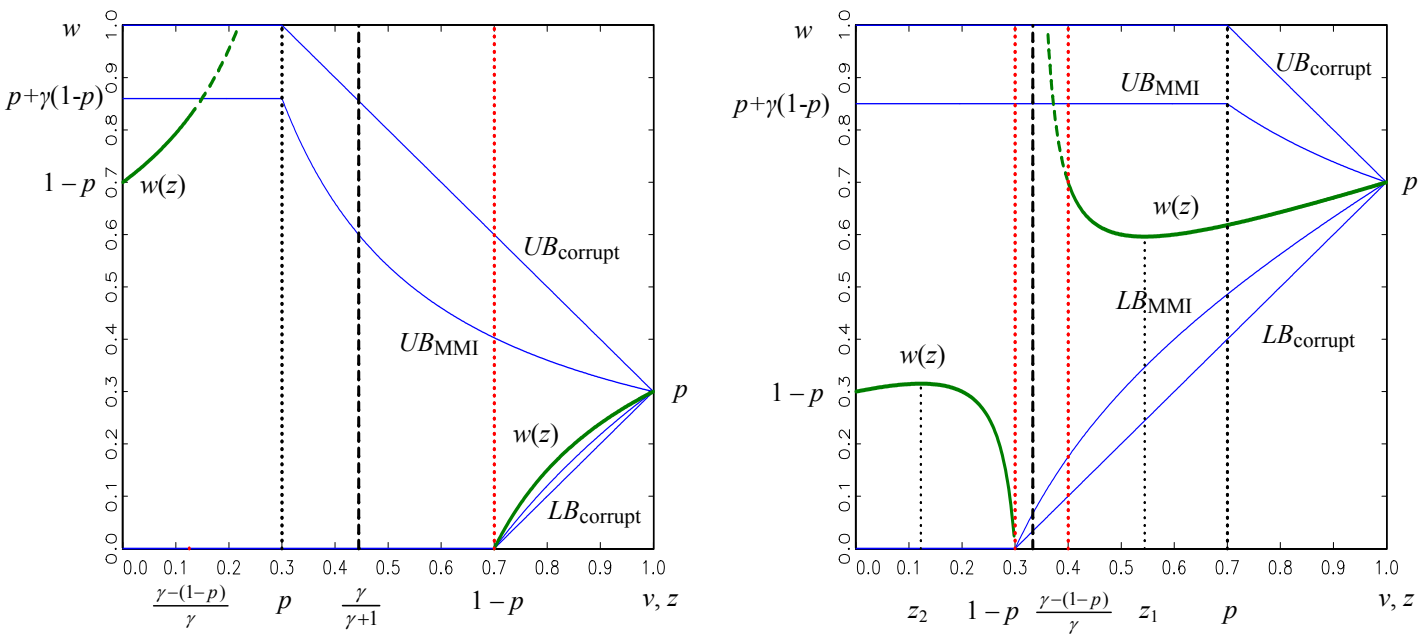


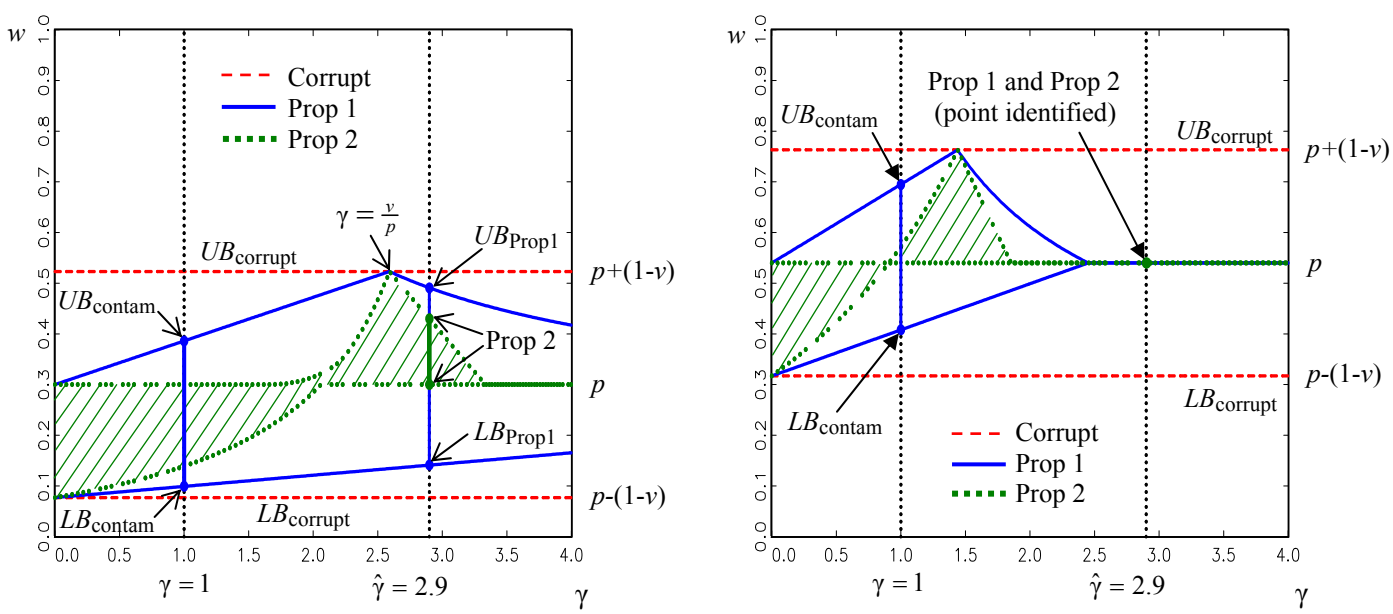
Figure 4. Bounds on Marijuana Use as a Function of γ

A. Marijuana use, past year

$p = 0.30,\ \ v = 0.777,\ \ \hat{\gamma} = 2.9$

B. Marijuana use, lifetime

$p = 0.54,\ \ v = 0.777,\ \ \hat{\gamma} = 2.9$



Note: These figures trace out sharp bounds on drug use prevalence rates as a function of γ. The Proposition 1 bounds evaluated at γ = 1 are equivalent to Horowitz and Manski's (1995) contaminated sampling bounds.