2018-03

Pooling stated and revealed preference data in the presence of RP endogeneity

Pooling Stated and Revealed Preference Data in the Presence of RP Endogeneity

Corresponding Author:
John Paul Helveston[a]
jhelvy@bu.edu

Coauthors:
Elea McDonnell Feit[b]
efeit@drexel.edu

Jeremy Michalek[c]
jmichalek@cmu.edu

January 15, 2018

a Institute for Sustainable Energy
Boston University
595 Commonwealth Ave.
Boston, MA 02215, USA

b LeBow College of Business
Drexel University
3220 Market St.
Philadelphia, PA 19104, USA

c Department of Mechanical Engineering and
Department of Engineering and Public Policy
Carnegie Mellon University
5000 Forbes Ave.
Pittsburgh, PA 15213, USA

Abstract

Pooled discrete choice models combine revealed preference (RP) data and stated preference (SP) data to exploit advantages of each. SP data is often treated with suspicion because consumers may respond differently in a hypothetical survey context than they do in the marketplace. However, models built on RP data can suffer from endogeneity bias when attributes that drive consumer choices are unobserved by the modeler and correlated with observed variables. Using a synthetic data experiment, we test the performance of pooled RP-SP models in recovering the preference parameters that generated the market data under conditions that choice modelers are likely to face, including (1) when there is potential for endogeneity problems in the RP data, such as omitted variable bias, and (2) when consumer willingness to pay for attributes may differ from the survey context to the market context. We identify situations where pooling RP and SP data does and does not mitigate each data source's respective weaknesses. We also show that the likelihood ratio test, which has been widely used to determine whether pooling is statistically justifiable, (1) can fail to identify the case where SP context preference differences and RP endogeneity bias shift the parameter estimates of both models in the same direction and magnitude and (2) is unreliable when the product attributes are fixed within a small number of choice sets, which is typical of automotive RP data. Our findings offer new insights into when pooling data sources may or may not be advisable for accurately estimating market preference parameters, including consideration of the conditions and context under which the data were generated as well as the relative balance of information between data sources.

Keywords: endogeneity; discrete choice modeling; data enrichment; choice data combination; pooled models; revealed preference; stated preference; stated choice.

Highlights

1. We identify cases where pooling RP and SP data can improve or worsen parameter recovery.
2. The likelihood ratio test can falsely reject pooling when RP data describes a small number of choice sets.
3. We propose a method for computing the information balance in pooled models.
4. We offer new insights for pooling SP and RP data under potential RP endogeneity.

1 Introduction

Discrete choice modeling is an established area of research in transportation, econometrics, psychology, and marketing (Louviere et al., 2000; Train, 2009). The typical approach uses a random utility framework to develop a model of the processes presumed to have generated observed consumer choices, parameterized as a function of product attributes. The parameters of the model are then estimated from observed consumer choices. Although previously developed modeling techniques take many forms, the data for such models typically come from one of two sources: "Revealed Preference" (RP) data and "Stated Preference" (SP) data. RP data records the details of actual purchases made in the marketplace, and SP data is collected in controlled survey experiments where respondents rate, rank, or make choices from a set of hypothetical products controlled by the researcher (Louviere et al., 2000).

These two data sources have different strengths and weaknesses. RP data have the face validity of reflecting actual market choices, but they only provide information on existing products and past behavior. In addition, models estimated on RP data can often suffer from econometric problems, such as multicollinearity, which can result in poorly identified parameter estimates, or an endogeneity problem from omitted variables (attributes observed by the consumer but not by the modeler), which can result in biased parameter estimates if they influence choice and are correlated with observed attributes. Further, in many RP settings, such as the automotive market, modelers observe multiple choices from only one set of products, and, in the presence of omitted variables, this results in correlated errors that violate the IID assumption of the multinomial logit model, leading to inaccurate inference.

SP data can be designed to avoid these problems and can include hypothetical products and attributes, allowing researchers to explore options not currently available in the market (Carson and Groves, 2007; Ding et al., 2005). However, hypothetical consumer choices in a survey context may not be consistent with purchase choices in a market context. As a result, while model parameters estimated on SP data may reflect statistically unbiased estimates of preferences in the survey context, they may still be inconsistent with consumers' preferences as realized in the marketplace.

Recognizing their complementary strengths and weaknesses, modelers have developed methods for using both data sources in model estimation (c.f. Louviere et al., 2000, Ch. 8). In this study, we focus on the pooled RP-SP model where model parameters are informed by both data sets.[1] The pooled model assumes that preferences for attributes common to both RP and SP data sets can be modeled with common parameters while allowing for attributes that are only observed in one data set to be informed by that data set. Figure 1 provides a conceptual diagram of the RP, SP, and pooled models.

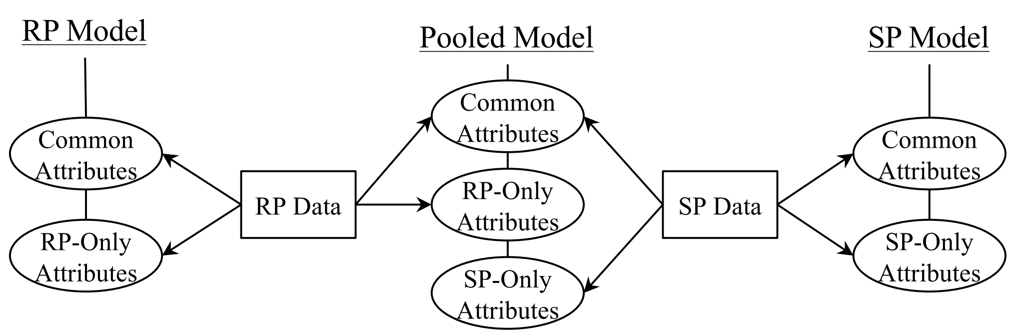


Figure 1: Conceptual diagram for RP, SP, and pooled models. Boxes indicate data, elipses indicate model parameters, and arrows indicate which data inform which parameter estimates. "Common Attributes" are those observed in both RP and SP contexts.

[1]While the sequential model provides another approach for using RP and SP data, attribute coefficients are only informed by the SP data while the RP data is used to calibrate predictions to actual market shares (Swait et al., 1994).

The pooled RP-SP model has been used in numerous studies to overcome some of the limitations of RP data, such as adding information on attributes or alternatives that do not exist in RP data and improving statistical properties by adding variation from SP data to attributes that are collinear in RP data. For instance, SP data might include new automotive technologies that do not currently exist or combine attributes in ways not currently seen in the marketplace (e.g. an inexpensive electric car with a 500 mile driving range). Pooled models have also been shown to achieve better predictive performance than RP-only or SP-only models in some applications (Brownstone et al., 2000; Feit et al., 2010; Hensher and Bradley, 1993; Swait et al., 1994; Swait and Andrews, 2003).

Each data source has potential issues that may be difficult to measure in practice, and the effect these issues have on pooled models has not been fully characterized in the prior literature. In this study, we focus on the effect of RP endogeneity from omitted variables on pooled RP-SP models. Although endogeneity in RP data has been studied extensively (Berry et al., 1995; Villas-Boas and Winer, 1999; Guevara, 2015), methods to correct endogeneity bias are often inefficient (Rossi, 2014; Haaf et al., 2016). Prior papers on RP-SP pooling have largely ignored the potential for omitted variable bias in the RP data. Using a synthetic data experiment, we show that pooling RP and SP data in the presence of RP endogeneity can lead to pooled model parameter estimates that are far from the true RP parameters and can render the frequently-used likelihood ratio test inappropriate for assessing whether the pooling assumption is justified.

We also illustrate how other characteristics of RP data, such as having multiple choice observations from the same choice set, can result in low informativeness about estimated parameters. SP data, by contrast, is often highly-informative; as a result, pooling RP and SP data can lead to parameter estimates (and substantive conclusions) that are primarily driven by the SP data. While this is not a problem if the SP choice process is the same as the RP choice process, it does call into question the value of pooling versus simply using the SP estimates. Little guidance is available in the literature on gauging the contribution of each data source to the pooled estimate. To remedy this, we introduce a new metric that researchers can use to assess the balance of information between RP and SP data.

In the next section, we provide some necessary background information on the pooled RP-SP model, including a precise description of the model and a summary of the literature. Section 3 describes the setup of the simulation study, and Section 4 reports the findings. In Section 5 we examine different sensitivity cases of the main results. We then discuss several important considerations for pooled models in Section 6, including a proposed metric for the information balance between RP and SP data, the potential to use endogeneity corrections in pooled models, and some limitations of our analysis. Finally, Section 7 concludes with a summary of our findings for understanding under what conditions pooling data sources may or may not be advisable for accurately estimating market preference parameters.

2 Background

2.1 The Random Utility Model

The pooled RP-SP model is based on the random utility model—a well-established probability model that can be estimated from observed consumer choices (Louviere et al., 2000; Train, 2009). Random utility models assume that consumers faced with a choice among a set of alternatives choose the alternative j that has the greatest utility u_j. The utility is modeled as $u_j = v_j + \varepsilon_j$, where v_j is a function of the vector of observed attributes of the product—$v_j = f(\mathbf{x}_j)$ representing the portion of utility explained by the attributes—and ε_j is a random variable representing the portion of utility unobservable to the modeler. Utility u_j is thus a random variable, and identification of the alternative with the highest utility is probabilistically assessed. In addition, utility only has relative, rather than absolute, value.

Consider the following utility model:

$$u_j^* = \boldsymbol{\beta}^{*'} \mathbf{x}_j - \alpha^* p_j + \varepsilon_j^*, \qquad \varepsilon_j^* \sim \text{Gumbel}\left(0, \sigma^2 \frac{\pi^2}{6}\right) \tag{1}$$

where α^* is the coefficient for price p_j, $\boldsymbol{\beta}^*$ is the vector of coefficients for non-price attributes $\mathbf{x}_j$, and the error term, ε_j^*, is an IID random variable with a Gumbel extreme value distribution of mean zero and variance $\sigma^2(\pi^2/6)$.[2] There exists an infinite set of combinations of values for α^*, $\boldsymbol{\beta}^*$ and σ that produce the same choice probabilities, and a model with all of these parameters is not identifiable. To specify an identifiable model, there are two commonly used normalization approaches. The first approach, referred to as the "preference space" model (Train and Weeks, 2005), normalizes the scale of the error term by dividing equation (1) by σ:

$$\left(\frac{u_j^*}{\sigma}\right) = \left(\frac{\boldsymbol{\beta}^*}{\sigma}\right)' \mathbf{x}_j - \left(\frac{\alpha^*}{\sigma}\right) p_j + \left(\frac{\varepsilon_j^*}{\sigma}\right), \qquad \left(\frac{\varepsilon_j^*}{\sigma}\right) \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{2}$$

Renaming the scaled utility parameters as $\boldsymbol{\beta} = (\boldsymbol{\beta}^*/\sigma)$ and $\alpha = (\alpha^*/\sigma)$, the standardized error term as $\varepsilon_j = (\varepsilon_j^*/\sigma)$, and the scaled utility as $u_j = (u_j^*/\sigma)$, equation (2) can be written as the usual parameterization of the multinomial logit model (MNL) in the preference space:

$$u_j = \boldsymbol{\beta}' \mathbf{x}_j - \alpha p_j + \varepsilon_j \qquad \varepsilon_j \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{3}$$

where $\boldsymbol{\beta}$ and α are the model parameters. The vector $\boldsymbol{\beta}$ represents the utility obtained from changes in each non-price attribute (relative to the standardized scale of the error term), and α represents utility obtained from price reductions (relative to the standardized scale of the error term). Since the error in equation (3) follows the standard extreme value distribution, the probability that a consumer will choose option j from the choice set $\mathcal{J}_c$ follows the familiar multinomial logit form (cf. Train, 2009):

$$P_{jc} = \frac{\exp\left(\boldsymbol{\beta}' \mathbf{x}_j - \alpha p_j\right)}{\sum\limits_{k \in \mathcal{J}_c} \exp\left(\boldsymbol{\beta}' \mathbf{x}_k - \alpha p_k\right)}, \qquad \forall c \in \{1, 2, 3, \ldots C\}, \quad j \in \mathcal{J}_c, \tag{4}$$

where c indexes a set of C choice sets. Model coefficients in equation (3) are measured in units of *utility* per unit attribute, which is only meaningful relative to other terms in the model and cannot be directly interpreted independently of the rest of the model or directly compared to other models.

The alternative normalization approach, referred to as the "willingness-to-pay (WTP) space" (Train and Weeks, 2005), is to normalize equation (1) by α^* instead of σ:

$$\left(\frac{u_j^*}{\alpha^*}\right) = \left(\frac{\boldsymbol{\beta}^*}{\alpha^*}\right)' \mathbf{x}_j - p_j + \left(\frac{\varepsilon_j^*}{\alpha^*}\right), \qquad \left(\frac{\varepsilon_j^*}{\alpha^*}\right) \sim \text{Gumbel}\left(0, \frac{\sigma^2}{(\alpha^*)^2} \frac{\pi^2}{6}\right) \tag{5}$$

The error term ε_j^* in equation (5) is scaled by $\lambda^2 = \sigma^2/(\alpha^*)^2$, and we can rewrite equation (5) multiplying both sides by $\lambda = (\alpha^*/\sigma)$ and renaming $u_j = (\lambda u_j^*/\alpha^*)$, $\boldsymbol{\beta} = (\boldsymbol{\beta}^*/\alpha^*)$, and $\varepsilon_j = (\lambda \varepsilon_j^*/\alpha^*)$, resulting in the parameterization of the multinomial logit model in the WTP space:

$$u_j = \lambda\left(\boldsymbol{\beta}' \mathbf{x}_j - p_j\right) + \varepsilon_j \qquad \varepsilon_j \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{6}$$

[2]Models that allow $\boldsymbol{\beta}^*$ to vary across consumers facilitate more flexible substitution patterns and can sometimes lead to better model fit (Brownstone and Train, 1999; McFadden and Train, 2000), but the key points in our analysis do not depend on heterogeneity, so we do not introduce it here.

where λ and $\boldsymbol{\beta}$ are the model parameters. $\boldsymbol{\beta}$ represents the importance of each non-price attribute relative to price (i.e. the WTP)[3], and λ represents the scale of the deterministic portion of utility relative to the standardized scale of the random error term (i.e. a "signal to noise ratio"). The probability that a consumer will choose option j from the choice set $\mathcal{J}_c$ again follows the familiar multinomial logit form but in the WTP space:

$$P_{jc} = \frac{\exp\left[\lambda\left(\boldsymbol{\beta}'\mathbf{x}_j - p_j\right)\right]}{\sum\limits_{k \in \mathcal{J}_c} \exp\left[\lambda\left(\boldsymbol{\beta}'\mathbf{x}_k - p_k\right)\right]}, \qquad \forall c \in \{1, 2, 3, \ldots C\}, \quad j \in \mathcal{J}_c, \tag{7}$$

The WTP-space parameterization was originally proposed in the context of heterogeneous models, where specifying a normal distribution for the preference space parameters is not equivalent to specifying a normal distribution for WTP-space parameters (Train and Weeks, 2005; Sonnier et al., 2007). In the homogeneous models we study here, the two parameterizations are equivalent, but the WTP-space has the advantage that the coefficients in $\boldsymbol{\beta}$ can be directly compared across different models irrespective of scale differences and are measured in a directly interpretable unit (currency). This is particularly convenient for pooling RP and SP data when the goal is to compare the value of attributes (which should be measured relative to currency) across SP, RP and pooled models. For these reasons, we adopt the WTP-space parameterization for the remainder of the main text, and we offer in supplemental information a comparison of our main results in the preference space.

2.2 The Pooled Model in the WTP Space

Our description of the pooled utility model is based on the description in Chapter 8 of Louviere et al. (2000) but transformed into the WTP space. We begin by specifying the data-generating utility for products in the SP data set, which follows the form of equation (6):

$$u_j^{\mathrm{S}} = \lambda^{\mathrm{S}}\left(\boldsymbol{\beta}^{\mathrm{S}'}\mathbf{x}_j + \boldsymbol{\gamma}^{\mathrm{S}'}\mathbf{y}_j^{\mathrm{S}} - p_j\right) + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{8}$$

where S denotes SP and we separate out the attributes into two vectors: $\mathbf{x}_j$ represents attributes that are common between the RP and SP data sets, and $\mathbf{y}_j^{\mathrm{S}}$ represents those attributes observed in the SP data but not in the RP data. The coefficients for these attribute vectors are $\boldsymbol{\beta}^{\mathrm{S}}$ and $\boldsymbol{\gamma}^{\mathrm{S}}$, respectively.

Similarly, we define the data-generating utility function for the RP data as:

$$u_j^{\mathrm{R}} = \lambda^{\mathrm{R}}\left(\boldsymbol{\beta}^{\mathrm{R}'}\mathbf{x}_j + \boldsymbol{\gamma}^{\mathrm{R}'}\mathbf{y}_j^{\mathrm{R}} + \boldsymbol{\zeta}'\mathbf{z}_j - p_j\right) + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{9}$$

where R denotes RP and we again separate the attributes into two vectors: $\mathbf{x}_j$ are the common attributes, and $\mathbf{y}_j^{\mathrm{R}}$ are the attributes observed in the RP data but not the SP data. We also include a vector of attributes $\mathbf{z}_j$ that affect consumer choice in the RP context but are unobserved by the modeler and have coefficients $\boldsymbol{\zeta}$. We do not include a $\mathbf{z}$ term in equation (8) because all attributes observed by the respondent in the controlled SP setting are, by design, also observed by the modeler.

Since the $\boldsymbol{\zeta}'\mathbf{z}_j$ term in equation (9) represents attributes unobserved by the modeler, they are omitted in model specifications used for estimation and are absorbed in the error term. For example, in the automobile context a vehicle's styling is known to be important to consumer choice in the RP setting, but because styling is difficult to codify it is typically omitted from estimated models. In the next subsection, we discuss the conditions under which this is reasonable and when it could create an endogeneity problem.

[3] Although prevalent in the literature, the term "willingness-to-pay" is potentially misleading, as it suggests a sharp threshold above which all customers will not buy even though purchase probability varies continuously with price in the MNL model. A better name might be "equivalence price," i.e. the prices at which an alternative with a particular feature has equal utility as one without the feature.

To estimate an SP-only model, the modeler assumes the model specification

$$u_j^{\mathrm{S}} = \hat{\lambda}^{\mathrm{S}} \left(\hat{\boldsymbol{\beta}}^{\mathrm{S}'} \mathbf{x}_j + \hat{\boldsymbol{\gamma}}^{\mathrm{S}'} \mathbf{y}_j^{\mathrm{S}} - p_j \right) + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel} \left(0, \frac{\pi^2}{6} \right) \tag{10}$$

and estimates the parameters $\hat{\boldsymbol{\beta}}^{\mathrm{S}}$, $\hat{\boldsymbol{\gamma}}^{\mathrm{S}}$ and $\hat{\lambda}^{\mathrm{S}}$ using the SP data. The hats indicate that the parameters are estimated rather than the true data-generating parameters in equation (8). To estimate an RP-only model, the modeler assumes the model specification

$$u_j^{\mathrm{S}} = \hat{\lambda}^{\mathrm{R}} \left(\hat{\boldsymbol{\beta}}^{\mathrm{R}'} \mathbf{x}_j + \hat{\boldsymbol{\gamma}}^{\mathrm{R}'} \mathbf{y}_j^{\mathrm{R}} - p_j \right) + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel} \left(0, \frac{\pi^2}{6} \right) \tag{11}$$

and estimates the parameters $\hat{\boldsymbol{\beta}}^{\mathrm{R}}$, $\hat{\boldsymbol{\gamma}}^{\mathrm{R}}$ and $\hat{\lambda}^{\mathrm{R}}$ using the RP data. Note that the $\boldsymbol{\zeta}'\mathbf{z}_j$ term in equation (9) is omitted from the RP utility specification used in estimation as it is unobserved by the modeler, so the model in equation (11) is intentionally misspecified in a way that mimics a potential misspecification faced in practical settings. If any of the elements in $\mathbf{z}_j$ are correlated with any of the elements in $\mathbf{x}$ or $\mathbf{y}^{\mathrm{R}}$, then the resulting parameter estimates may be biased unless the modeler controls for the endogeneity from the omitted variables.

In a pooled model, it is assumed that $\boldsymbol{\beta}^{\mathrm{R}} = \boldsymbol{\beta}^{\mathrm{S}} = \boldsymbol{\beta}$ for common attributes, and the pooled utility specification is given by

$$u_j^{\mathrm{S}} = \hat{\lambda}^{\mathrm{S}} \left(\hat{\boldsymbol{\beta}}' \mathbf{x}_j + \hat{\boldsymbol{\gamma}}^{\mathrm{S}'} \mathbf{y}_j^{\mathrm{S}} - p_j \right) + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel} \left(0, \frac{\pi^2}{6} \right) \tag{12}$$

$$u_j^{\mathrm{R}} = \hat{\lambda}^{\mathrm{R}} \left(\hat{\boldsymbol{\beta}}' \mathbf{x}_j + \hat{\boldsymbol{\gamma}}^{\mathrm{R}'} \mathbf{y}_j^{\mathrm{R}} - p_j \right) + \varepsilon_j^{\mathrm{R}} \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel} \left(0, \frac{\pi^2}{6} \right) \tag{13}$$

where $\hat{\boldsymbol{\beta}}$ is modeled as a vector of parameters common to the two utility models. Differences between the two data sets in the relative scale of the error term are accommodated through the estimated scale terms $\hat{\lambda}^{\mathrm{R}}$ and $\hat{\lambda}^{\mathrm{S}}$.

The key assumption in the pooled model is that $\hat{\boldsymbol{\beta}}$ is the same across RP and SP contexts; if this assumption is false for any of the attributes, then $\hat{\boldsymbol{\beta}}$ will not be an unbiased estimate of the true effect of the attribute in the RP context, $\boldsymbol{\beta}^{\mathrm{R}}$, which could lead to erroneous conclusions about how consumers will react to changes in attributes and prices in the market.

To examine if pooling is justifiable, the prior literature has used visual tests, such as plotting RP model coefficients versus SP model coefficient estimates for common attributes, and statistical tests, such as the likelihood ratio (LR) test for pooling (Swait and Louviere, 1993). The LR test checks whether there is enough evidence in the data to reject the hypothesis that the two data sources were generated from models with the same parameters ($\boldsymbol{\beta}^{\mathrm{R}} = \boldsymbol{\beta}^{\mathrm{S}}$). The test statistic is computed as $-2[(L^{\mathrm{R}} + L^{\mathrm{S}}) - L^{\mathrm{P}}]$, where L^{R} and L^{S} are the log-likelihood values from the separately estimated RP and SP models, respectively, and L^{P} is the log-likelihood of the pooled model. If pooling is rejected for some but not all attributes, then it is recommended that those parameters be moved into $\mathbf{y}^{\mathrm{R}}$ and $\mathbf{y}^{\mathrm{S}}$ so that separate parameters for those attributes may be estimated while still pooling the remaining parameters (Louviere et al., 2000). The LR test has been widely used to justify pooling assumptions in much of the RP-SP pooling literature (Adamowicz et al., 1994, 1997; Ben-Akiva and Morikawa, 1990; Bhat and Castelar, 2002; Birol et al., 2006; Brownstone et al., 2000; Hensher and Bradley, 1993; Hensher et al., 1999; Mark and Swait, 2004).

2.3 Advantages and Disadvantages of RP and SP Data

RP and SP data each have different strengths and weaknesses, but the literature suggests neither is a "gold standard." RP data reflect real purchases where money is exchanged for goods and/or services. As a result, estimated model parameters using RP data are generally believed to reflect consumers' preferences in the real-world market context. However, a number of common issues with RP data can result in biased parameter estimates. In particular, the problem of endogeneity, which occurs when an observed variable is correlated with the error term, leads to biased estimates.[4]

When the number and nature of attributes in a choice situation is sufficiently large and complex (for instance, automobile choice or housing choice), any discrete choice model will omit some unobserved information about attributes that influence choice; that is, there are important variables $\mathbf{z}$ that affect choice in equation (9), and these variables are not observed by the modeler and are omitted in model estimation. This misspecification will lead to biased parameter estimates if any of the observed attributes are correlated with the unobserved attributes and proper measures are not taken to account for the endogeneity (Wooldridge, 2010; Train, 2009). A particularly common concern in the literature on RP models is price endogeneity, which occurs if an unobserved attribute that influences choice (e.g. the "style" of a vehicle) also influences the price that manufacturers set (Berry et al., 1995). A number of past studies have developed methods to correct endogeneity bias in choice models (see Guevara, 2015, for a review). It has also been argued that many of these "fixes" can generate more problems than they solve (Rossi, 2014; Haaf et al., 2016).

Other concerns with RP data include measurement error (particularly in measuring the attributes and prices faced by decision makers in the market), low attribute variation, and multicollinearity (e.g. *price* and *size* are often positively correlated in the automobile market). Modelers also may not observe the consideration sets consumers face, and the very nature of RP data excludes information about products or attributes that do not yet exist in the market. Further, an important feature of RP data is that it often includes multiple choices from the same set of fixed alternatives with common unobserved attributes. As we will show, this produces correlations among these choices that violate the IID assumption of the multinomial logit model, resulting in inconsistent maximum likelihood estimators and inaccurate standard errors. For example, automobile choice models are often estimated using RP data from a small number of model years where unobserved features of the products in the market are fixed for all consumers. Thus, while RP data has the face validity of reflecting real market choices, the modeler will be limited in the set of parameters she can estimate and must consider whether to attempt to correct for a potential endogeneity (if possible). Determining whether or not estimated model parameters are unbiased is often a judgment call based on the modeler's understanding of how the data were generated and the availability of data.

In contrast, SP data is collected in controlled survey experiments, allowing the researcher to avoid many of the concerns that arise in RP data by controlling the observable attributes, designing the survey to avoid attribute correlations, and avoiding the presence of unobserved attributes that influence consumer choice. In addition, SP data can provide information about products or attributes that are not yet available in the market. However, it is well-known that the context of a choice situation can alter choice behavior. Carson and Groves (2007) illustrate that different incentives provided by particular response formats on surveys can induce "strategic behavior" in respondents, and unless the collected information is "incentive compatible" with the real-world incentives, respondents may fail to to reveal their true preferences. Ding et al. (2005) show a similar result in which models estimated from "incentive-aligned" choices (where respondents were required to actually purchase one of their chosen product profiles) made predictions that were more consistent with observed market choices. Additionally, the salience of each attribute in a survey, where attributes are often listed together in a side-by-side comparison, may differ from the salience of those same attributes in a market context (Hardt et al., 2017; Hensher, 2010).

Since the survey context of SP data is different from the real-world market context of RP data, WTP parameters from the SP context may differ from the corresponding WTP parameters in the RP context (Beck et al., 2016). We will refer to this as a "context difference" between the

[4] Endogeneity is typically irrelevant in SP data since SP surveys are controlled experiments where the researcher observes all attributes influencing choice.

SP and RP data. The term "hypothetical bias" has also been used to describe the SP parameters that do not align with RP parameters (e.g. Hensher, 2010), but to avoid confusion we reserve the term "bias" to refer to statistical bias.[5]

The literature is unclear about the likelihood that respondents will overreact versus underreact to product features in the survey context relative to the market context. For example, one might guess that respondents tend to be less sensitive to price in hypothetical tasks where they do not spend real money, but they also may overreact in some cases by, for example, paying more attention to price in the survey context than they do in the market context. Thus, while there are arguably fewer econometric challenges to recovering parameters that reflect survey behavior, the modeler typically remains uncertain about the degree to which those parameters are comparable to the corresponding parameters in the market context.

In short, RP and SP data often have opposing uncertainties: in the RP (SP) context, modelers are often more (less) certain that the observed choices reflect true market preferences but less (more) certain whether parameters estimated from those data are statistically biased. Whether or not any of these issues are cause for concern in any specific application depends on how the data were generated and collected. In the absence of definitive empirical evidence, the degree of concern about the potential presence of these issues is typically determined subjectively by the modeler. Table 1 provides a summary of the different advantages and disadvantages of RP and SP data.

Table 1: Advantages and Disadvantages of Stated and Revealed Preference Data

	Stated Preference (SP)	Revealed Preference (RP)
Advantages	• Include information on products and attributes not yet in market • Controlled experiment	• Reflects choices from real market
Disadvantages	• Potential difference in survey vs. market choice behavior	• Potential for omitted variable bias • Low attribute variation • Measurement error • Multicollinearity among explanatory variables • Missing information about consideration sets and product availability • No information on products and attributes that do not yet appear in the marketplace • Correlated errors among within-market choices

2.4 Literature on Pooled RP-SP Models

The pooled RP-SP model was originally proposed by Morikawa (1989) and has since been used in numerous studies to overcome some of the limitations of RP data, such as including information

[5] While past literature has used the term "hypothetical bias" to broadly describe *any* discrepancies between preferences revealed by actual market behavior and those from hypothetical experiments (Beck et al., 2016), there is still no widely accepted behavioral theory that explains the phenomenon (Loomis, 2011).

on attributes or alternatives that do not exist in RP data (Axsen et al., 2009; Ben-Akiva and Morikawa, 1990; Birol et al., 2006; Brownstone et al., 2000; Dissanayake and Morikawa, 2003; Hensher and Bradley, 1993; Polydoropoulou and Ben-Akiva, 2001; Swait and Andrews, 2003) and improving statistical properties of the data by adding variation to highly collinear attributes in RP data (Adamowicz et al., 1994, 1997; Ben-Akiva et al., 1994; Brownstone et al., 2000; Feit et al., 2010; Hensher et al., 1999; Mark and Swait, 2004). Other studies also suggest that RP data can help "ground" SP data in reality since the survey context of SP data may not reflect reality (Axsen et al., 2009; Ben-Akiva and Morikawa, 1990; Bhat and Castelar, 2002; Birol et al., 2006; Brownstone et al., 2000; Brownstone and Small, 2005; Dissanayake and Morikawa, 2003; Feit et al., 2010; Hensher et al., 1999; Huang et al., 1997; Mark and Swait, 2008; Polydoropoulou and Ben-Akiva, 2001; Swait et al., 1994). Finally, several studies have found that pooled models have achieved better predictive performance over separately-estimated RP or SP models on within-sample or hold-out tests across multiple domains (Adamowicz et al., 1997; Axsen et al., 2009; Börjesson, 2008; Brownstone et al., 2000; Feit et al., 2010; Hensher and Bradley, 1993; Mark and Swait, 2004; Swait et al., 1994; Swait and Andrews, 2003). Table 5 in the supplemental information provides a full summary of the different motivations for pooling in the 29 studies we reviewed.

Despite the breadth of the pooling literature, nearly all of the pooled RP-SP studies have implicitly assumed away endogeneity in the RP data and do not attempt to assess endogeneity bias. This is perhaps an understandable accident of history, since pooled models were introduced prior to the emphasis on endogeneity in the RP literature and the two literatures have developed somewhat independently of one another. However, given the prevalence of concern for endogeneity in non-pooling studies that use RP data, its omission in the pooled RP-SP literature today is surprising. In our review of 29 pooled RP-SP studies, only one study (von Haefen and Phaneuf, 2008) considers the presence of endogeneity in the RP data, but results were inconclusive due to limitations in the data used in that study (i.e. pooling was rejected).

In this study, we use a synthetic data experiment to explore the outcomes of pooled models under different conditions of endogeneity in the RP data and differences in attribute preferences between the RP and SP contexts. Our analysis focuses on the pooled WTP estimate $\hat{\boldsymbol{\beta}}$ relative to the true value of $\boldsymbol{\beta}^{\mathrm{R}}$ in the RP data. Any difference between those two is a direct consequence of endogeneity and / or context differences.

3 Synthetic Data Study

The goal of our synthetic data study is to characterize how (1) endogeneity in the RP data and (2) differences between consumer choice behavior in the SP versus RP context can affect the ability of pooled models to recover the true data-generating parameters of the market context. We first generate multiple sets of RP and SP data under different conditions of endogeneity and context differences using known parameters and then estimate a series of RP, SP, and pooled RP-SP models using those data. We then compare the resulting parameter estimates to the true RP market parameters. While features of the synthetic data generating process are inspired by typical RP and SP data for consumer automobile choices, we parameterized the data generating process in a way that allows us to explore data that would be representative of other products, such as consumer package goods, in Section 5.

3.1 Simulating the Data

We generate data based on a simplified example of the data-generating functions (the "true" utility models) in equations (8) and (9), where utility depends on price p, a single non-price attribute x that is common to each data source, and a single RP attribute z unobserved by the modeler that is used to control the level of price endogeneity in the RP data, depending on it's correlation with price. While this is a simple case, it provides a clean illustration of key issues for pooling in the presence of endogeneity in the RP data and context differences between the

RP and SP data. We explore more complex data sets with more than one pooled parameter later.

SP Data Generation. We simulate the SP data to approximate a typical choice-based conjoint survey. The true SP data-generating model is given by

$$u_j^{\mathrm{S}} = \lambda^{\mathrm{S}} \left(\beta^{\mathrm{S}} x_j - p_j\right) + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{14}$$

where each parameter is a scalar. Each of C^{S} choice sets (each representing a choice-based conjoint question) is generated by randomly choosing A^{S} alternatives from the full factorial design using 5 levels (-2, -1, 0, 1, 2) for each attribute. As would be typical, each randomized choice question is prohibited from containing any two identical alternatives. Choices are simulated by taking draws from the multinomial distribution defined by the choice probabilities P_{jc}^{S}:

$$P_{jc}^{\mathrm{S}} = \frac{\exp\left[\lambda^{\mathrm{S}}\left(\beta^{\mathrm{S}} x_j - p_j\right)\right]}{\sum\limits_{k \in \mathcal{J}_c^{\mathrm{S}}} \exp\left[\lambda^{\mathrm{S}}\left(\beta^{\mathrm{S}} x_k - p_k\right)\right]}, \qquad \forall c \in \left\{1, 2, 3, \ldots C^{\mathrm{S}}\right\}, \quad \forall j \in \mathcal{J}_c^{\mathrm{S}}, \tag{15}$$

where C^{S} is the total number of SP choice sets. The total number of choice observations is $N^{\mathrm{S}} = n^{\mathrm{S}} C^{\mathrm{S}}$ where n^{S} is the number of choice observations per SP choice set. For the SP data, we take just one choice observation draw from each choice set ($n^{\mathrm{S}} = 1$) to represent a randomized conjoint survey design.[6]

To introduce context differences between the SP data and the RP data, we compute β^{S} by adjusting β^{R} by a fixed scalar δ:

$$\beta^{\mathrm{S}} = \delta \beta^{\mathrm{R}}, \tag{16}$$

where β^{R} is the RP WTP parameter in equation (17) below. Parameterizing the context difference as a function of δ allows us to explore cases where the SP WTP for changes in attribute x is understated ($\delta < 1$), overstated ($\delta > 1$), or consistent with the market context ($\delta = 1$).[7] In addition to differences in WTP induced by δ, respondents could have different sensitivities to price between the survey and market contexts, which would affect all WTP parameters. For example, if a respondent values money less in the survey setting, then the SP WTP parameters for all the non-price attributes will be lower than the corresponding RP WTP parameters.

RP Data Generation. We simulate the RP data to approximate typical automotive market data (c.f. Axsen et al., 2009; Feit et al., 2010; Haaf et al., 2016). The true RP data-generating model is given by

$$u_j^{\mathrm{R}} = \lambda^{\mathrm{R}} \left(\beta^{\mathrm{R}} x_j + \zeta z_j - p_j\right) + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right), \tag{17}$$

where each parameter is a scalar. Each of C^{R} choice sets (each representing a particular fixed set of alternatives for which multiple choices are observed, i.e., a "market") is generated by randomly drawing A^{R} vectors of attributes $[p, x, z]'$ from a multivariate normal distribution with mean $\boldsymbol{\mu} = [0, 0, 0]'$ and variance-covariance matrix given by

$$\boldsymbol{\Sigma}^{\mathrm{R}} = \begin{bmatrix} 1 & \rho_{px} & \rho_{pz} \\ \rho_{px} & 1 & \rho_{xz} \\ \rho_{pz} & \rho_{xz} & 1 \end{bmatrix} \tag{18}$$

[6] Note that pencil-and-paper conjoint studies often collect multiple observations from the same choice set (i.e. $n^s > 1$), however randomizing the choice questions for each user is more common in today's online conjoint studies.

[7] Since we are focused on potential bias in the estimates of the WTP parameters, we do not introduce other context-specific differences that could also exist, such as scale differences, which we consider in our sensitivity analysis in Section 5.

where the relative correlations between attributes are given by the ρ terms. The correlation between p and z allows us to induce an endogeneity by choosing $\rho_{pz} \neq 0$ such that price becomes endogenous when z is omitted from the estimated model. Within this framework we can also explore cases where there is collinearity by setting $\rho_{px} \neq 0$. In our simulation studies, we examine cases with and without price endogeneity, and we also run sensitivity cases of different collinearity levels (see Table 2).

Choices are simulated by taking draws from the multinomial distribution defined by the choice probabilities P^{R}_{jc}:

$$P^{\mathrm{R}}_{jc} = \frac{\exp\left[\lambda^{\mathrm{R}}\left(\beta^{\mathrm{R}} x_j + \zeta z_j - p_j\right)\right]}{\sum\limits_{k \in \mathcal{J}^{\mathrm{R}}_c} \exp\left[\lambda^{\mathrm{R}}\left(\beta^{\mathrm{R}} x_k + \zeta z_k - p_k\right)\right]}, \qquad \forall c \in \left\{1, 2, 3, \dots C^{\mathrm{R}}\right\}, \quad \forall j \in \mathcal{J}^{\mathrm{R}}_c, \tag{19}$$

Note that the ζz_j term in the utility model is included when the data is simulated. (The term is only omitted later when estimating RP and pooled models). In the RP data, we allow for multiple choice observation draws from each choice set ($n^{\mathrm{R}} > 1$) to represent multiple consumer purchases among the same set of alternatives. C^{R} is the total number of RP choice sets, and there are a total of $N^{\mathrm{R}} = n^{\mathrm{R}} C^{\mathrm{R}}$ choice observations. Note that the unobserved z_j as well as the observed attributes x_j and p_j are fixed across multiple draws for each choice set.

We consider cases where $n^{\mathrm{R}} > 1$ because RP data often involves multiple choice observations from the same choice set. For example, in a given month, the set of vehicles available to US consumers is generally fixed, and thousands of customers are observed choosing from the same set of alternatives. Similarly, in package goods categories most large retailers provide the same set of options at the same prices across large geographic markets. This limits attribute variation in the data and will typically reduce the information about the parameters per choice observation, *ceteris paribus*. It also means that unobservables z are common across all observations from within the same choice set (as would be the case in most real markets), and when z is absorbed into the error term, the errors become non-IID and the multinomial logit model is misspecified.

3.2 Simulation Parameters

The goal of our simulation study is to understand how characteristics of the data affect estimates of WTP in the pooled model. We explore several features of the data including the size of the data, the presence of endogeneity in the RP data, the presence of context differences between the SP and RP data, and the number of choice sets in the RP data.

The values for all simulation parameters are summarized in Table 2. For our base case, we generate RP and SP data sets that are less typical of real data sets but have desirable properties for isolating the effects of RP endogeneity (controlled by ρ_{pz}) and context differences (controlled by δ) on pooled estimates of WTP. We examine the effects of more typical, less ideal conditions in a sensitivity analysis in Section 5. In our base case, we chose parameters that result in data sets that have an approximately equal contribution between the RP and SP data to the pooled parameter estimates. We achieve this by comparing the elements of the expected Fisher information matrix of each data set computed at the pooled estimates. The balance of information between the data sets is influenced by the number of observations, attributes, and alternatives in each data set as well as the level of attribute variation in the RP data, which increases with more choice sets (Huber and Zwerina, 1996).[8]

In our base case, we set the number of RP choice sets equal to the number of RP observations ($C^{\mathrm{R}} = N^{\mathrm{R}}$) so that only one choice is observed for each set of alternatives ($n^{\mathrm{R}} = 1$). This optimistic case ensures that the RP data have sufficient variation in the observed attributes and prevents correlations in the RP error term, allowing us to focus on the effects of ρ_{pz} and δ under ideal conditions. In Section 5, we extend these results by examining cases where more choices are observed for each choice set ($n^{\mathrm{R}} > 1$ and $C^{\mathrm{R}} < N^{\mathrm{R}}$) up to the opposite extreme where all

[8] Section 6.1 as well as Section 8.4 in the supplemental information provide further details on relationships between data set characteristics and the Fisher information.

choice observations are from one choice set ($n^{\mathrm{R}} = N^{\mathrm{R}}$ and $C^{\mathrm{R}} = 1$). We also consider a case where a vector of common attributes $\mathbf{x}_j$ is pooled in place of the single common attribute x_j used in the base case. Finally, we conduct an extensive parametric study examining the effects of wide ranges of values for each parameter in our synthetic data study (see Table 6 in the supplemental information).

Table 2: Parameters used to generate synthetic RP and SP data.

Parameter	Base Case	Sensitivity Range	Description
λ^{R}	1	[0.1, 5]	Scale of RP error term
λ^{S}	1	[0.1, 5]	Scale of SP error term
β^{R}	1	[0.5, 2]	WTP coefficient for attribute x in RP context
β^{S}	$\delta\beta^{\mathrm{R}}$	*	WTP coefficient for attribute x in SP context
δ	1	[0.5, 2]	Ratio of WTP for attribute x in the SP relative to RP context
ζ	1	[-1, 1.5]	WTP coefficient for unobserved attribute z in RP context
ρ_{px}	0	[0, 0.5]	Correlation between *price* and x in RP context
N^{R}	1,000	[500, 5,000]	Total number of RP choice observations per simulation
N^{S}	1,500	[500, 5,000]	Total number of SP choice observations per simulation
C^{R}	N^{R}	[1, N^{R}]	Number of RP choice sets per simulation
C^{S}	N^{S}	–	Number of SP choice sets per simulation
n^{R}	$N^{\mathrm{R}}/C^{\mathrm{R}}$	*	Number of RP choice observations per choice set
n^{S}	1	–	Number of SP choice observations per choice set
A^{R}	15	[3, 100]	Number of alternatives per RP market
A^{S}	3	[2, 10]	Number of alternatives per SP choice question

* Sensitivity range determined by other parameters

3.3 Test Cases

To test the effect of RP endogeneity (controlled by ρ_{pz}) and SP context differences (controlled by δ) on the ability of the pooled model to recover the true WTP parameter for x in the market context, β^{R}, we generate sets of RP and SP data using 6 different cases for ρ_{pz} and δ, shown in Table 3. If $\rho_{pz} = 0$ and $\delta = 1$ (case 2), then the pooled model assumption that $\beta^{\mathrm{R}} = \beta^{\mathrm{S}}$ is valid, and unbiased estimates of the true parameters can be recovered by estimating a pooled model. However, any deviation from these conditions implies that the pooling restriction is a misspecification and could result in estimates of $\hat{\beta}$ that differ from the true parameter β^{R}. By estimating the pooled model with the synthetic data, we can explore how well the pooled model recovers the true RP parameter under conditions of misspecification that frequently occur in practice.

Table 3: Test cases for generating synthetic RP and SP data

Case	ρ_{pz}	δ	Interpretation	
			RP Data	SP Data
1: "SP WTP Understated"	0	0.5	No price endogeneity	WTP for x understated
2: "Ideal"	0	1	No price endogeneity	No context differences
3: "SP WTP Overstated"	0	2	No price endogeneity	WTP for x overstated
4: "Two Wrongs Make a Right"	0.5	0.5	Price endogenous	WTP for x understated
5: "RP Price Endogenous"	0.5	1	Price endogenous	No context differences
6: "Wrong In The Same Way"	0.5	2	Price endogenous	WTP for x overstated

3.4 Model Estimation

We generate choice sets and choice observations for the pair of RP and SP data-generating functions described in Section 3.1 using the parameters in Section 3.2. For each pair of simulated RP and SP data, we estimate the parameters $\hat{\beta}$, $\hat{\lambda}^{\mathrm{S}}$, and $\hat{\lambda}^{\mathrm{R}}$ in the pooled RP-SP model:

$$u_j^{\mathrm{S}} = \hat{\lambda}^{\mathrm{S}} \left(\hat{\beta} x_j - p_j\right) + \varepsilon_j^{\mathrm{S}} \tag{20}$$

$$u_j^{\mathrm{R}} = \hat{\lambda}^{\mathrm{R}} \left(\hat{\beta} x_j - p_j\right) + \varepsilon_j^{\mathrm{R}} \tag{21}$$

where $\hat{\beta}$ is common between the two models and the hats on the parameters indicate that they are estimated. Note that the attribute z from the true data-generating process in equation (17) is omitted from the RP utility specification in equation (21) used for estimation because z is unobserved by the modeler; as a result, price will be endogenous in the RP data when $\rho_{pz} \neq 0$, and both the RP-only and pooled models will have biased coefficient estimates.

Assuming the error terms are distributed IID Gumbel, then the choice probabilities for each data source are given by the multinomial logit fraction such that [9]

$$\hat{P}_{jc}^{\mathrm{S}} = \frac{\exp\left[\hat{\lambda}^{\mathrm{S}}\left(\hat{\beta}x_j - p_j\right)\right]}{\sum\limits_{k \in \mathcal{J}_c^{\mathrm{S}}} \exp\left[\hat{\lambda}^{\mathrm{S}}\left(\hat{\beta}x_k - p_k\right)\right]}, \qquad \forall j \in \mathcal{J}_c^{\mathrm{S}} \tag{22}$$

$$\hat{P}_{jc}^{\mathrm{R}} = \frac{\exp\left[\hat{\lambda}^{\mathrm{R}}\left(\hat{\beta}x_j - p_j\right)\right]}{\sum\limits_{k \in \mathcal{J}_c^{\mathrm{R}}} \exp\left[\hat{\lambda}^{\mathrm{R}}\left(\hat{\beta}x_k - p_k\right)\right]}, \qquad \forall j \in \mathcal{J}_c^{\mathrm{R}} \tag{23}$$

The parameters in equations (21) and (20) are estimated by maximizing the pooled log-likelihood:

$$L = \sum_{c=1}^{C^{\mathrm{S}}} \sum_{i \in \mathcal{N}_c^{\mathrm{S}}} \sum_{j \in \mathcal{J}_c^{\mathrm{S}}} y_{ijc} \ln \hat{P}_{jc}^{\mathrm{S}} + \sum_{c=1}^{C^{\mathrm{R}}} \sum_{i \in \mathcal{N}_c^{\mathrm{R}}} \sum_{j \in \mathcal{J}_c^{\mathrm{R}}} y_{ijc} \ln \hat{P}_{jc}^{\mathrm{R}} \tag{24}$$

where $y_{ijc} = 1$ if alternative j is chosen from choice set c in choice observation i and $y_{ijc} = 0$ otherwise; $\mathcal{N}_c^{\mathrm{S}}$ and $\mathcal{N}_c^{\mathrm{R}}$ are the sets of choice observations for each choice set $\mathcal{J}_c^{\mathrm{S}}$ and $\mathcal{J}_c^{\mathrm{R}}$,

[9] This model is intentionally misspecified relative to the data generating processes specified by equations (14) and (17). Due to the omitted variable z_j, the effective error term in the RP data generating process is not Gumbel; rather, it is the sum of a Gumbel random variable and a normal random variable. More importantly, if we observe multiple RP choices from among the same set of alternatives (with the same omitted z_j), then the error term is no longer IID. We explore the implications of this misspecification in our sensitivity analysis in Section 5.

respectively; and the probabilities $\hat{P}^{\mathrm{S}}_{jc}$ and $\hat{P}^{\mathrm{R}}_{jc}$ in equation (24) are defined in equations (22) and (23). In our base case, $|\mathcal{N}^{\mathrm{S}}_{c}| = |\mathcal{N}^{\mathrm{R}}_{c}| = 1$.

Since the WTP utility function is nonlinear in parameters and the log likelihood function is nonconcave, we use a randomized multi-start algorithm to search for a global solution. In each of 10 iterations, we maximize the log-likelihood using a different set of random starting points drawn from a uniform distribution between -1 and 1 and store the result, and we report the solution with the greatest log-likelihood. For our base case simulation experiment, the algorithm converged to the same solution in all 10 multi-start iterations for 91% of the simulations, and the maximum difference in the log-likelihood across all 10 multi-start iterations was less than 0.001 for 99% of the simulations.

4 Results

We simulate 1,000 sets of RP and SP data for each of the six test cases in Table 3 and then use them to estimate RP, SP, and pooled models. Figure 2 shows the ratio between the estimated and true model WTP coefficients for each test case using the base case parameters in Table 2. The plots show the distribution of the ratio across all 1,000 simulations using a logarithmic y-axis for comparing ratios. Likelihood ratio tests were computed at the 10% significance level.

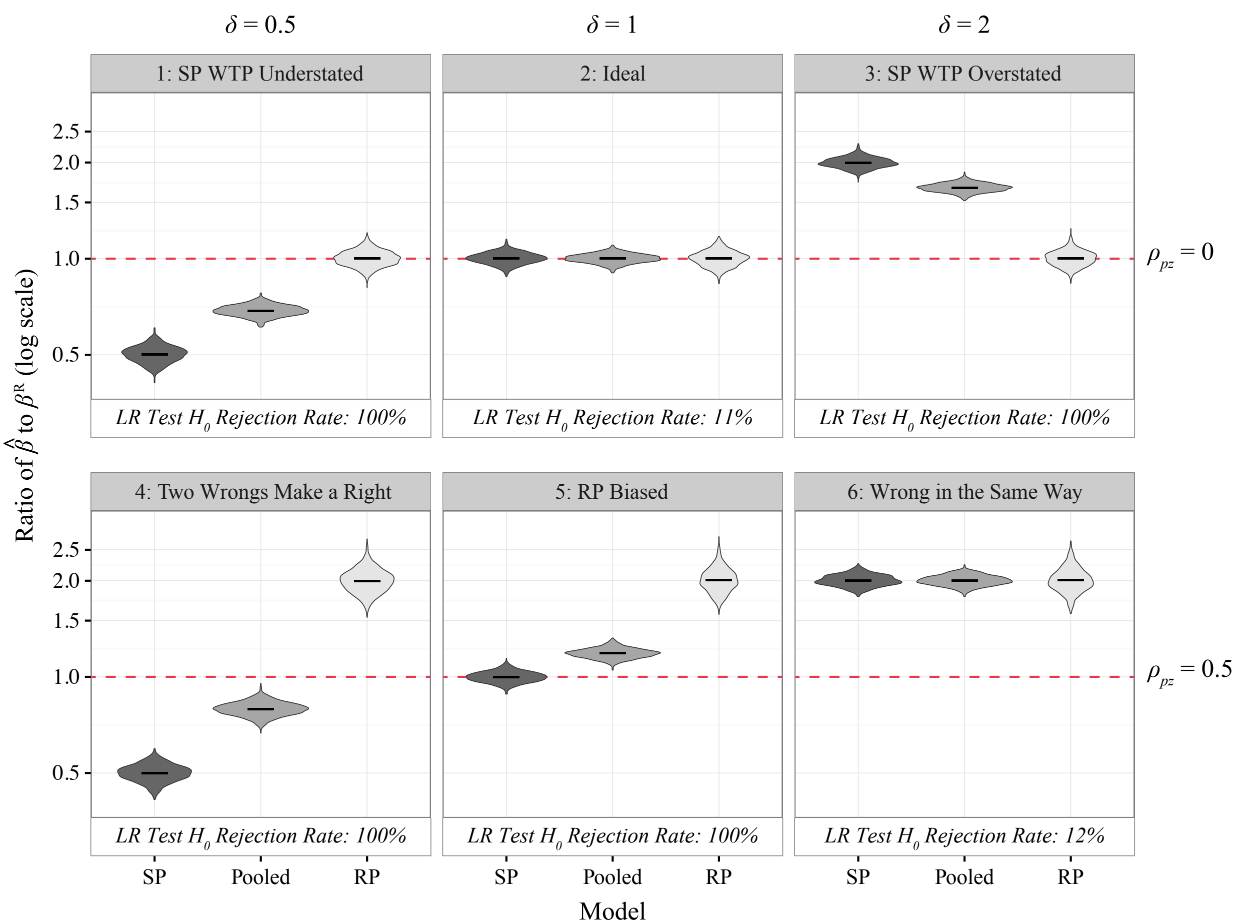


Figure 2: SP, RP, and pooled model results of the ratio of $\hat{\beta}$ to β^{R} for each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. The sampling variation is visualized as a kernel density plot. For each data set, the LR test of pooling is conducted at 10% significance.

In cases 1-3, price is not endogenous in the RP data, and the difference in the estimated SP WTP for x varies from underestimating β^{R} in case 1 when $\delta = 0.5$ to overestimating β^{R} in case 3 when $\delta = 2$. As a result, the pooled estimate of $\hat{\beta}$ is biased away from β^{R} in cases 1 and 3 but unbiased in case 2. Cases 1 and 3 illustrate the well-known problem that can arise if RP and SP data are pooled when there are substantial context differences. As expected, the LR test successfully detects these differences and rejects pooling. When pooling is appropriate (Case 2), the LR test rejects pooling at close to the nominal rejection rate of 10%.

In cases 4-6, price is endogenous in the RP data, which has different implications depending on the value of δ. In case 4 the upward bias on $\hat{\beta}$ created by the price endogeneity in the RP data is partially balanced by the context difference in the opposite direction for the SP data, and the pooled estimate of $\hat{\beta}$ is actually less biased than either the RP or SP estimates (hence the name "Two Wrongs Make a Right"). In case 5, the SP WTP matches that of the RP context, but the RP-only estimate of the WTP parameter is biased due to endogeneity. Pooling the RP data with SP data that do not suffer from context differences helps mitigate the endogeneity bias. As expected, the LR test rejects pooling in every simulation in these two cases since the SP and RP parameter estimates are significantly different from one another, even though in both cases the pooled model helps mitigate the RP endogeneity bias.

In case 6, the price endogeneity in the RP data and context difference in the SP data where $\delta = 2$ have similar effects on $\hat{\beta}$ such that the bias is roughly the same magnitude and direction; as a result, the pooled estimate of $\hat{\beta}$ is approximately the same as those in the RP and SP models (hence the name "Wrong in the Same Way"). In this case, the LR test accepts pooling at close to the nominal confidence of 10% since the RP and SP parameters are nearly the same, but the pooled model does not mitigate the biases.

The values of ρ_{pz} and δ in our base case simulations also affect the scale parameters $\hat{\lambda}^{\mathrm{R}}$ and $\hat{\lambda}^{\mathrm{S}}$ (see Figure 8 in the supplemental information). Depending on the case, the modeler could make false conclusions about both consumer WTP for attributes ($\hat{\beta}$) as well as how consistent consumer choices appear (the "signal-to-noise ratio") in the RP and SP contexts ($\hat{\lambda}^{\mathrm{R}}$ and $\hat{\lambda}^{\mathrm{S}}$).

5 Sensitivity Analysis

The parameters chosen for our base case in Table 2 were chosen to isolate the effects of endogeneity and context differences on parameter recovery. However, some characteristics of the data are idealized and not typical in real data sets (e.g. 1,000 independent RP choice sets). To confirm that our findings generalize beyond the base case, we examine a case with fewer RP choice sets as well as a case with a vector of five common attributes in $\mathbf{x}$ that can be pooled. We also conduct an extensive parametric study by running simulations across the full sensitivity range of each parameter in Table 2. For each parameter, we compare the WTP results from the RP, SP, and pooled models as in Figure 2. Figures for each sensitivity case are provided in the supplemental information.

5.1 Fewer Independent RP Choice Sets

Typical RP data sets often have very few choice sets (in some cases only one). Within each choice set, the products on offer are held fixed, which limits the variation in the attributes. To examine the effect of fewer choice sets on pooling with endogeneity bias, we run a simulation using the same parameters in Table 2 except we use only one RP choice set ($C^{\mathrm{R}} = 1$), resulting in all observed choices being made from the same set of products with the same unobserved attributes z_j. Figure 3 shows the results from 1,000 simulated RP and SP data sets for these simulations.

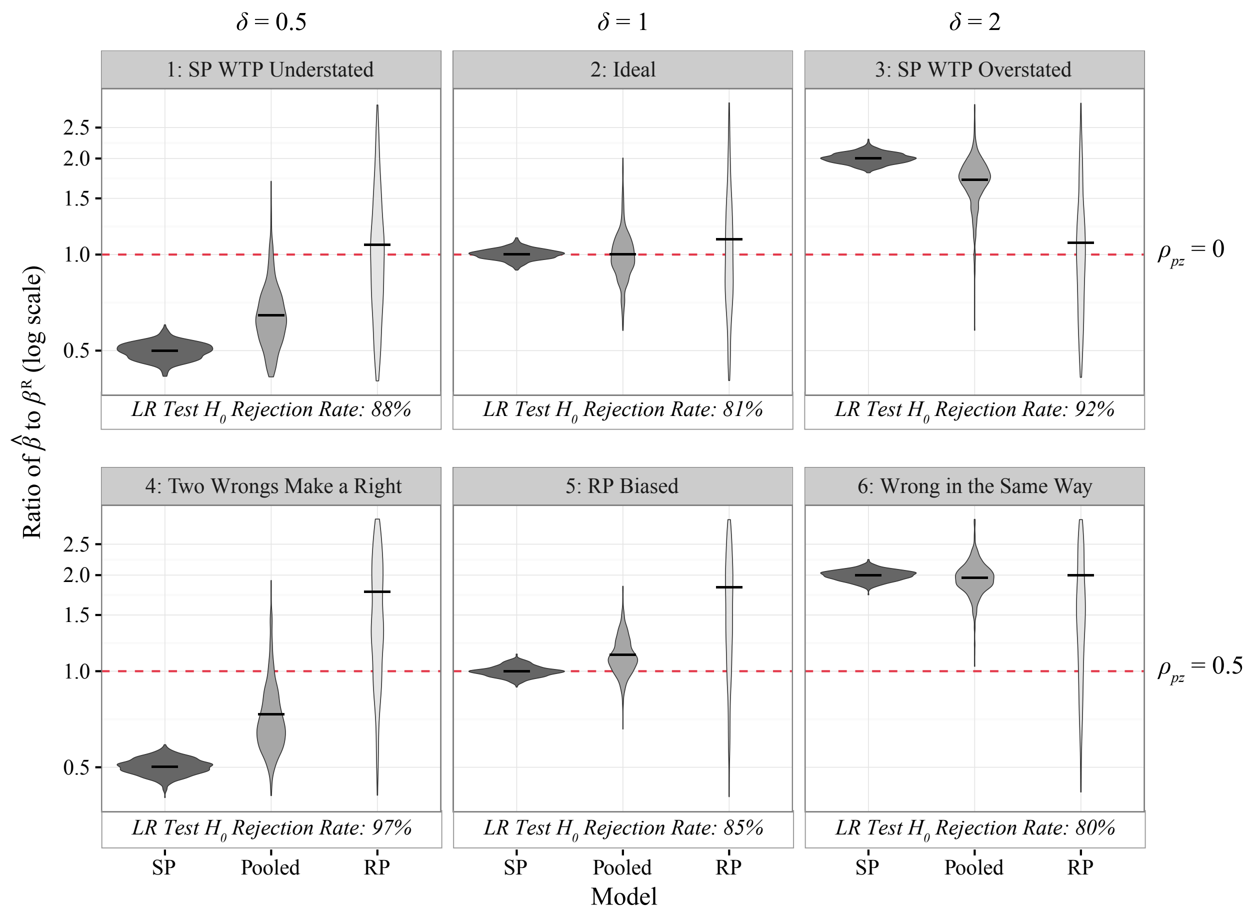


Figure 3: SP, RP, and pooled model results of the ratio of $\hat{\beta}$ to β^{R} for each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2 except with $C^{\mathrm{R}} = 1$. The sampling variation is visualized as a kernel density plot. For each data set, the LR test of pooling is conducted at 10% significance.

Results are similar to those from the base case simulations except that the sampling variance for the RP model is much greater. With only one choice set, the RP attribute variation is so low that $\hat{\beta}$ may only be weakly identified from the data, and the estimates for any given data set may be quite far from the true parameter.[10] However, because the SP model has a far lower sampling variance, the pooled model sampling variance is substantially improved over that of the RP model. The ideal case 2 illustrates why SP data is often used to augment deficient RP data and indeed is a strong motivation for pooling.

In addition, since there is only one choice set and a single omitted variable that is common to all choice observations, the IID assumption of the multinomial logit model is not satisfied, which has important consequences.[11] For example, if the IID error assumption is not met, the sampling error no longer has the expected variance, and standard errors estimated from the curvature of the misspecified likelihood function will be wrong. The error term dependence

[10] In comparing the results of our multistart algorithm, we found that for 95% of the simulations the optimizer converged to the same solution from 10 different random starting points, suggesting the model is indeed identified even with only 1 RP choice set. In the other 5% of simulations, the optimizer still converged to a discrete set of local minima (as opposed to all different minima), suggesting that the model is still identified even in these cases.

[11] While this structure is a result of our simulation design, it can also potentially occur in real market data where multiple choices are observed from a set of fixed alternatives with fixed attributes, such as in automotive sales data for a single model year where vehicle attributes are fixed.

also affects the sampling distribution of the chi-squared LR test statistic, resulting in an over-rejection of pooling in cases where the data generating process is actually the same in both data sets.

We see this over-rejection of the LR test in cases 2 and 6, which are much higher than the expected 10%. In case 2 where the true parameters are actually the same and pooling should be accepted, the LR test still rejects pooling in 81% of the simulations. We find similarly high rejection rates in case 6 when the parameter biases are in the same direction and have similar magnitudes. Figure 4 shows how the LR test H_0 rejection rate approaches the nominal 10% for cases 2 and 6 as the number of RP choice sets increases from 1 to 1,000, holding N^{R} fixed at 1,000. In cases 1, 3, 4 and 5, we can also see that the power of the test to detect parameter differences increases as the number of independent RP choice sets increases.

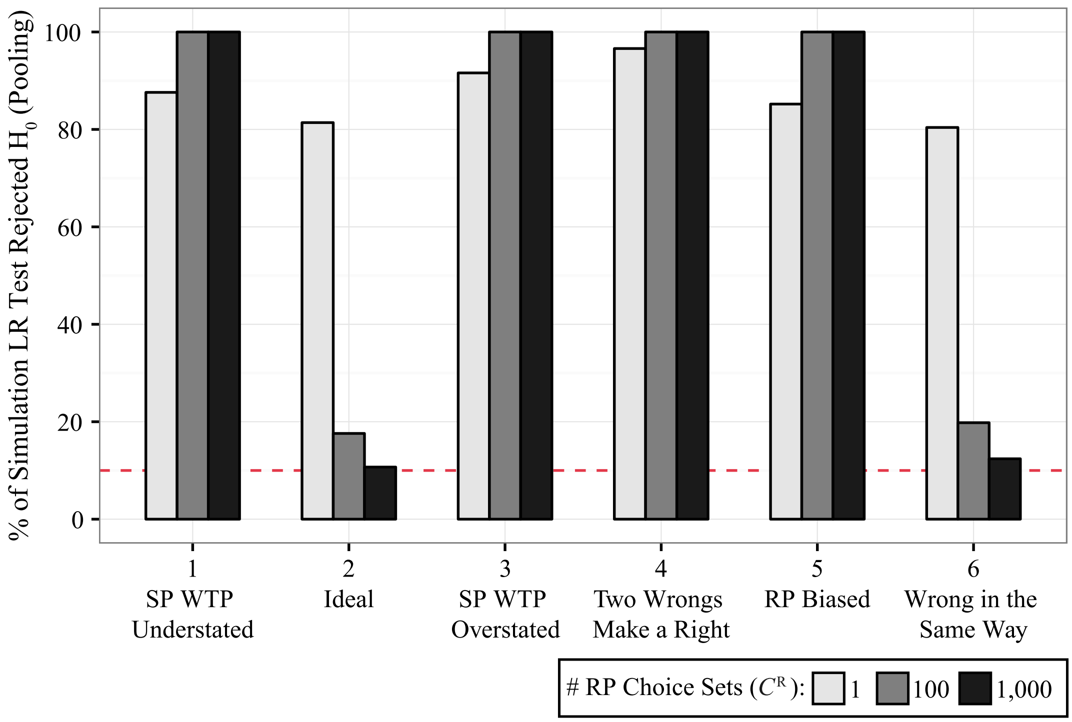


Figure 4: Likelihood ratio test rejection frequency for the null hypothesis $\beta^{\mathrm{R}} = \beta^{\mathrm{S}}$ as a function of ρ_{pz}, δ, and C^{R}. Each bar is computed from 1,000 simulations using the base case parameters in Table 2. The dashed horizontal red line indicates the nominal 10% confidence level rejection rate.

Figure 5 illustrates how the error term dependence produced by having a small number of choice sets affects the sampling distribution of the LR test statistic for test case 2 (when the true RP and SP parameters are the same). The "QQ" plot compares the empirical quantiles of the LR test statistics versus the theoretical asymptotic chi-squared quantiles. The plot shows that when there is only one choice set in the RP data (the light gray points) the sampling distribution of the LR test statistic is far from the theoretical chi-squared form from which we compute the threshold for the hypothesis test. As C^{R} increases and the number of observations per choice set approaches one ($C^{\mathrm{R}} \to N^{\mathrm{R}}$ and $n^{\mathrm{R}} \to 1$), the error term dependence decreases and the sampling distribution of the LR test statistic becomes closer to the asymptotic chi-squared distribution.[12] Importantly, if the sampling distribution of the LR test statistic is not chi-squared, then increasing the number of observations will not improve the LR test rejection

[12] Due to the omitted variable z_j, the error term in the RP data generating process is the sum of a Gumbel random variable and a normal random variable, which could also affect whether the LR test statistic is asymptotically chi-squared. However, as we see in Figure 5, this issue appears to be minor in comparison to the violation of the IID

rate. This suggests that researchers should use caution when applying the LR test to RP data with a small number of choice sets or with low variation in attributes across choice sets.

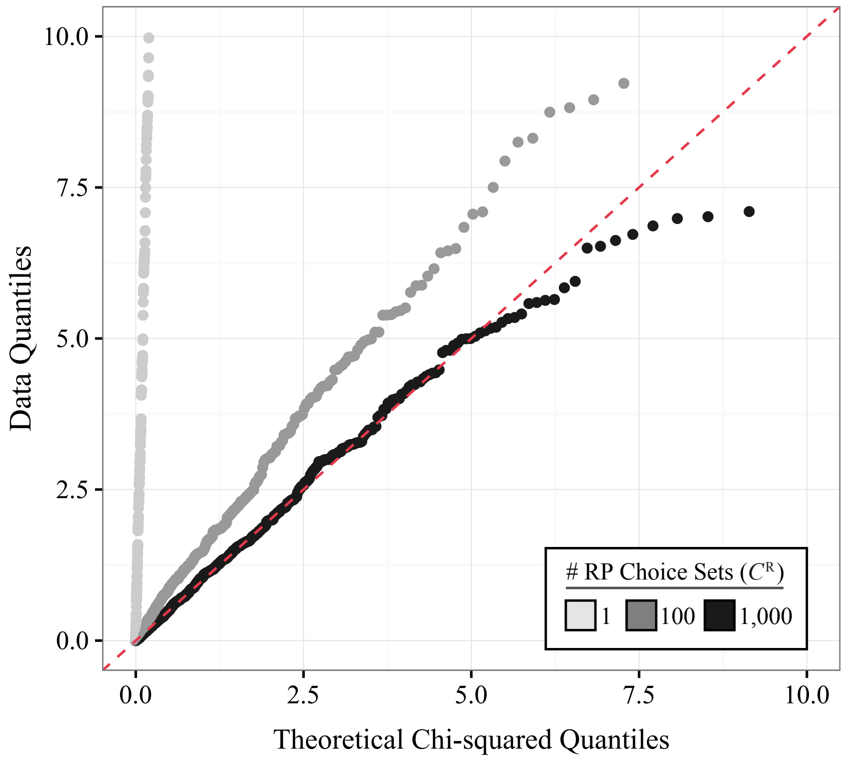


Figure 5: "QQ" plot of empirical quantiles of the LR test statistics versus the theoretical asymptotic Chi-squared quantiles for test case 2. Test statistics are computed from 1,000 simulations using the base case parameters in Table 2 and varying C^{R} while holding N^{R} fixed at 1,000.

5.2 Multiple Pooled Parameters

In our base case we have only one pooled attribute x for simplicity, but in typical pooled models there are multiple common parameters that could potentially be pooled. To examine this situation, we run a case with five common attributes in $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5]'$. In order to isolate the effect of a context difference, we model x_1 as the only attribute where true consumer WTP differs between RP and SP contexts (controlled by δ). The data generation process is the same as in our base case with all the same parameters, and the additional pooled parameters β_2^{R} through β_5^{R} are each set to 1.

Results are nearly identical with those in our base case except that the additional WTP coefficients for x_2 through x_5 in cases 4 through 6 are also biased upward due to the fact that price endogeneity affects all WTP coefficients (see Figure 9 in the supplemental information). We also ran a case with five common attributes, three RP-specific attributes, and three SP-specific attributes, and results were again similar.

5.3 Parametric Sensitivity Analysis

We conduct an extensive parametric study by running simulations across the full sensitivity range of each parameter in Table 2. For each parameter, we compare the WTP results from the RP, SP, and pooled models as in Figure 2. Figures for each sensitivity case as well as a

assumption.

table summarizing the qualitative effect of each sensitivity case are provided in the supplemental information. The parameters that control the number of choice observations and the sizes of the choice sets (N^{R}, N^{S}, A^{R}, A^{S}, and C^{R}) primarily affect the relative balance of information in the pooled model. Decreasing (increasing) the RP WTP parameter (β^{R}) or the scale parameters (λ^{R} and λ^{S}) increases (decreases) the sampling variance across simulations and also decreases (increases) the information in the respective data set.

In addition to these general observations, we observe that the sign and magnitude of the ζ coefficient (the WTP coefficient for the unobserved attribute z) influences our conclusions about the LR test in the presence of endogeneity in the RP data. As we noted in Section 4, our base case for test case 6 results in biased estimates of $\hat{\beta}$ that share the same direction and magnitude from the RP and SP data; as a result, the LR test largely fails to reject pooling. However, when we increase the size of ζ, the bias from the endogeneity in the RP data becomes larger in magnitude than that from the SP data, and the LR test largely rejects pooling. Likewise, when ζ is negative, cases 4 and 6 swap in their interpretation, with case 4 having biases in the same direction from each data source and case 6 having biases in opposite directions. The magnitude of the RP bias is also affected by the sign and magnitude of ρ_{pz} (i.e. a higher correlation between p and z increases the severity of the endogeneity).

6 Additional Considerations

6.1 Information Balance

The comparison between the sampling distributions for the RP estimates in Figure 2 and Figure 3 raises an important issue when pooling data: different data sources carry different levels of information about unknown model parameters. The amount of information can be influenced by a number of factors, including the number of observations, the levels of correlation and variation among observed attributes, the number of distinct choice sets, and the number of alternatives in the choice sets (Huber and Zwerina, 1996). For example, SP data are often highly informative even with few choice observations because the observations come from many different choice sets with alternatives and attributes that were designed specifically to create orthogonal variation in the attributes.[13] In contrast, as we can see in Figure 3, RP data often have large sample sizes but can be relatively uninformative if they have only one or a small number of choice sets and/or highly correlated attributes within choice sets (as is common in automotive choice data). Thus, the balance of information between the RP and SP data cannot be gauged simply by comparing sample sizes.

The pooled log-likelihood in equation (24) implicitly weights pooled parameter estimates by the respective amounts of information available in each data set. When the RP and SP parameters differ, the pooled estimate is effectively weighted closer to the parameter value of the more informative data source. One way to characterize this implicit weighting is to compare the elements of the respective Fisher information matrices, $\mathbf{I}^{\mathrm{S}}$ and $\mathbf{I}^{\mathrm{R}}$, which measure the amount of information a data set carries about the unknown model parameters. The information matrix for each data set can be computed as the negative of the Hessian of the log-likelihood function. Using the chain rule, we compute the RP and SP information matrices from the SP component and RP component of the log-likelihood objective function in equation (24):

$$\mathbf{I}^{\mathrm{S}} = -\sum_{c=1}^{C^{\mathrm{S}}} \sum_{i \in \mathcal{N}_c^{\mathrm{S}}} \sum_{j \in \mathcal{J}_c^{\mathrm{S}}} y_{ijc} \frac{P_{jc}^{\mathrm{S}} \nabla^2 P_{jc}^{\mathrm{S}} - \nabla P_{jc}^{\mathrm{S}\prime} \nabla P_{jc}^{\mathrm{S}}}{\left(P_{jc}^{\mathrm{S}}\right)^2} \tag{25}$$

$$\mathbf{I}^{\mathrm{R}} = -\sum_{c=1}^{C^{\mathrm{R}}} \sum_{i \in \mathcal{N}_c^{\mathrm{R}}} \sum_{j \in \mathcal{J}_c^{\mathrm{R}}} y_{ijc} \frac{P_{jc}^{\mathrm{R}} \nabla^2 P_{jc}^{\mathrm{R}} - \nabla P_{jc}^{\mathrm{R}\prime} \nabla P_{jc}^{\mathrm{R}}}{\left(P_{jc}^{\mathrm{R}}\right)^2} \tag{26}$$

[13] Note that conjoint designs typically optimize the determinant of the FI for the preference space model, with the exception of Toubia and Hauser (2007) who propose to optimize the determinant of the FI for the WTP parameters.

where the ∇^2 symbol is used to indicate the Hessian and the gradient and Hessian are taken with respect to the vector of estimated model parameters. For the case of a logit model, equations (25) and (26) reduce to:

$$\mathbf{I}^{\mathrm{S}} = -\sum_{c=1}^{C^{\mathrm{S}}} \sum_{i \in \mathcal{N}_c^{\mathrm{S}}} \sum_{j \in \mathcal{J}_c^{\mathrm{S}}} y_{ijc} \left[\nabla^2 v_j^{\mathrm{S}} - \sum_{k \in \mathcal{J}_c^{\mathrm{S}}} \left(\nabla v_k^{\mathrm{S}\prime} \nabla P_{kc}^{\mathrm{S}} + P_{kc}^{\mathrm{S}} \nabla^2 v_k^{\mathrm{S}} \right) \right] \tag{27}$$

$$\mathbf{I}^{\mathrm{R}} = -\sum_{c=1}^{C^{\mathrm{R}}} \sum_{i \in \mathcal{N}_c^{\mathrm{R}}} \sum_{j \in \mathcal{J}_c^{\mathrm{R}}} y_{ijc} \left[\nabla^2 v_j^{\mathrm{R}} - \sum_{k \in \mathcal{J}_c^{\mathrm{R}}} \left(\nabla v_k^{\mathrm{R}\prime} \nabla P_{kc}^{\mathrm{R}} + P_{kc}^{\mathrm{R}} \nabla^2 v_k^{\mathrm{R}} \right) \right] \tag{28}$$

From equations (27) and (28), the Fisher information depends not only on the sample size, but also on the variation in the attributes.

In our base case simulation experiment in the WTP space, the parameters are λ^{R}, λ^{S}, and β, and the observable component of utility is

$$v_j^{\mathrm{S}} = \lambda^{\mathrm{S}} \left(\beta x_j - p_j \right) \quad \therefore \nabla v_j^{\mathrm{S}} = \begin{bmatrix} \beta x_j - p_j \\ \lambda^{\mathrm{S}} x_j \end{bmatrix} \quad \therefore \nabla^2 v_j^{\mathrm{S}} = \begin{bmatrix} 0 & x_j \\ x_j & 0 \end{bmatrix} \tag{29}$$

$$v_j^{\mathrm{R}} = \lambda^{\mathrm{R}} \left(\beta x_j - p_j \right) \quad \therefore \nabla v_j^{\mathrm{R}} = \begin{bmatrix} \beta x_j - p_j \\ \lambda^{\mathrm{R}} x_j \end{bmatrix} \quad \therefore \nabla^2 v_j^{\mathrm{R}} = \begin{bmatrix} 0 & x_j \\ x_j & 0 \end{bmatrix} \tag{30}$$

For a pooled model, we compute the observed information from the SP and RP data sets at the pooled model estimates. For each element in the diagonal of these matrices, we compute a value, $0 \leq \omega_k \leq 1$, which measures the proportion of expected information from the SP data set:

$$\omega_k = \frac{d_k^{\mathrm{S}}}{d_k^{\mathrm{R}} + d_k^{\mathrm{S}}} \tag{31}$$

where d_k^{S} and d_k^{R} are the k^{th} elements in the *diagonals* of $\mathbf{I}^{\mathrm{S}}$ and $\mathbf{I}^{\mathrm{R}}$. By computing an estimate of ω_k using the observed information at the pooled model estimates, the modeler can gain an understanding of how much each data set is informing the pooled model estimates (e.g. individual attributes with limited variation in the RP data will be more strongly influenced by the SP data—for an example, see Feit et al., 2010).[14] Values of ω_k closer to 0 suggest that the RP data has a stronger influence on the parameter associated with attribute k while values closer to 1 suggest the SP data has a stronger influence.

Figure 6 below shows how different balances of information between SP and RP data sets can change the outcome of pooled model estimates for test case 4 ($\rho_{pz} = 0.5$, $\delta = 0.5$). The information balance is varied by increasing the SP sample size, N^{S}, and the mean ω is computed from the observed information at the estimated pooled parameters. As more SP data are collected, the variance of the SP results shrinks, and the pooled model estimate moves toward the SP estimate, reflected by ω.

[14]Our calculation of the Fisher Information for the multinomial logit model assumes that the errors are IID; as a result, the FI for the RP data may be optimistic when error terms are correlated, as they are in our simulated data.

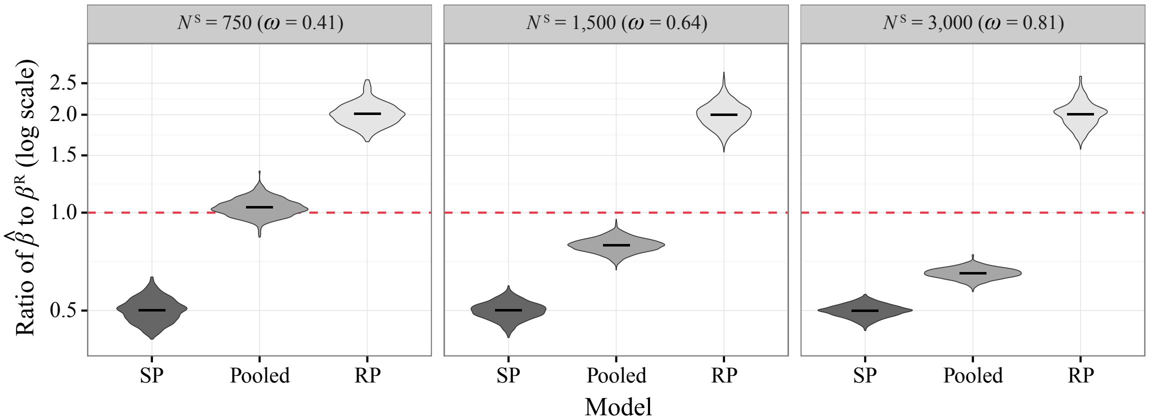


Figure 6: SP, RP, and pooled model results of $\hat{\beta}/\beta^{\mathrm{R}}$ for test case 4 ($\rho_{pz} = 0.5$, $\delta = 0.5$). Each box plot represents results from 1,000 simulated data sets using the base case parameters in Table 2 with different values of N^{S} and thus different information balances.

We chose our base case simulation parameters to produce data sets with relatively balanced levels of information about the WTP parameter. To confirm this balance, we compute the mean ω for $\hat{\beta}$ at the estimated pooled parameters across all 1,000 simulations for each test case. Table 4 shows that on average the base case SP data are slightly more informative than the RP data for all test cases, which is consistent with the pooled estimates of $\hat{\beta}$ in Figure 2 being slightly closer to the SP estimates than RP estimates. The endogeneity in the RP data in cases 4 through 6 pushes the balance even further towards the SP data.

Table 4: RP and SP Information Balance for Base Case Simulations

Case	mean(d_β^{R})	mean(d_β^{S})	mean(ω_β)
1: “SP WTP Understated”	768	909	0.54
2: “Ideal”	552	784	0.59
3: “SP WTP Overstated”	256	399	0.61
4: “Two Wrongs Make a Right”	423	766	0.64
5: “RP Price Endogenous”	317	586	0.65
6: “Wrong In The Same Way”	153	252	0.62

6.2 Endogeneity Corrections

Outside of the literature on pooled models, researchers who work regularly with RP data have developed several approaches to detect and correct endogeneity biases. We summarize those approaches briefly (see Guevara, 2015, for a review in the transportation context).

Since most endogeneity problems are motivated by an omitted variable, one way to prevent endogeneity is to make sure not to omit any variables that affect choice (Rossi, 2014) or to use some proxy for the omitted variable (Guevara, 2015). Unfortunately, this is nearly impossible for a complex product like an automobile. Most published automotive choice models omit important product characteristics, particularly difficult-to-quantify attributes like styling, interior quality, or sound quality.

A more common approach is to use a control function to correct the endogeneity (Villas-Boas and Winer, 1999; Petrin and Train, 2010). This involves finding an instrument for each potentially-endogenous observed variable, i.e. a variable that is correlated with the endogenous variable but uncorrelated with the outcome. Given a valid instrument, there are statistical tests for detecting endogeneity such as the Hausman test (Wooldridge, 2015). While many potential instruments have been proposed for endogenous prices in discrete choice models, including competitor prices or lagged prices, there remain serious concerns that these instruments are weak or invalid, resulting in parameter estimates that can be even more biased after a correction attempt (Rossi, 2014; Guevara, 2015; Haaf et al., 2016).

If products are observed repeatedly across multiple choice sets (e.g. vehicle models observed across multiple years in US automotive data), then another approach to account for omitted variables is to estimate an alternative specific constant (ASC) for each product (constant across choice sets). This fixed-effect acts as a proxy for the omitted variable and resolves the endogeneity. For example, if a vehicle design has particularly attractive styling and styling was omitted from the RP model, then the estimated ASC for the model will be relatively high, reflecting the greater appeal of the design due to the (omitted) styling. This eliminates the variation due to $\mathbf{z}$ from the error term, thereby correcting the endogeneity. Identification of this model requires variation in the observed correlated variable (e.g. *price*) without variation in the unobserved variable (e.g. *styling*).

Despite the limitations of these methods, one might ask the question, "Why not use one of these methods to correct the endogeneity in the RP data before pooling?" While this is theoretically feasible, there are two potential drawbacks. First, the endogeneity correction approaches require strong assumptions and/or data that is difficult or impossible to obtain. Second, these procedures change the information balance between the RP and SP data. For example, the instrumental variable approach effectively reduces the information in the RP data, essentially discarding variation in the RP data that could inform the parameters because it is potentially contaminated by endogeneity (Rossi, 2014). This may result in parameter estimates that become largely informed by the SP data, and while this is not necessarily an undesirable result, it does call into question the value of a pooled model over an SP model if they produce nearly the same result.

To illustrate how an endogeneity correction can be used in a pooled model, we conduct a simulation where product alternatives appear in multiple choice sets. Rather than independently generating each of C^{R} choice sets (each representing a particular market) as before in equation (17), we instead assume that each alternative appears in multiple choice sets with the observed attributes varying across choice sets but the unobserved attribute fixed across choice sets. This approach allows us to examine the performance of endogeneity correction with ASCs in a situation where we know the assumptions are correct. In practice, additional challenges emerge when the unobserved attributes vary across choice sets or when the variation in observed attributes across choice sets is insufficient.

Specifically, we generate alternatives j for RP choice sets such that each alternative appears an equal number of times across all choice sets. For each alternative j, we draw an ASC ξ_j (equivalent to drawing z_j for a fixed ζ in equation (17)) from the standard normal distribution, and we draw one set of observed attributes $[x_{jc}, p_{jc}]$ for each choice set $c \in \mathcal{C}_j^{\mathrm{R}}$ in which alternative j appears, conditional on ξ_j to maintain correlation ρ_{pz} in equation (18):

$$u_{jc}^{\mathrm{R}} = \lambda^{\mathrm{R}} \left(\beta^{\mathrm{R}} x_{jc} - p_{jc} + \xi_j\right) + \varepsilon_{jc}^{\mathrm{R}} \qquad \forall j \in \mathcal{J}_c^{\mathrm{R}}, \quad \forall c \in \mathcal{C}_j^{\mathrm{R}} \tag{32}$$

where $\mathcal{C}_j^{\mathrm{R}}$ is the set of choice sets that contain alternative j, ξ_j is the WTP for unobserved attributes of alternative j (the ASC), p_{jc} is the price for alternative j when it appears in choice set c, and x_{jc} is the other observed attribute for alternative j when it appears in choice set c. We simulate data sets using two scenarios: 3 RP choice sets of 2 alternatives each, and 10 RP choice sets of 5 alternatives each. We compare the results with and without ASCs and examine test cases 5 and 6. In both test cases price is endogenous in the RP data, but in test case 6 the SP WTP is over-stated whereas in test case 5 it is not. Figure 7 shows the results of each simulation.

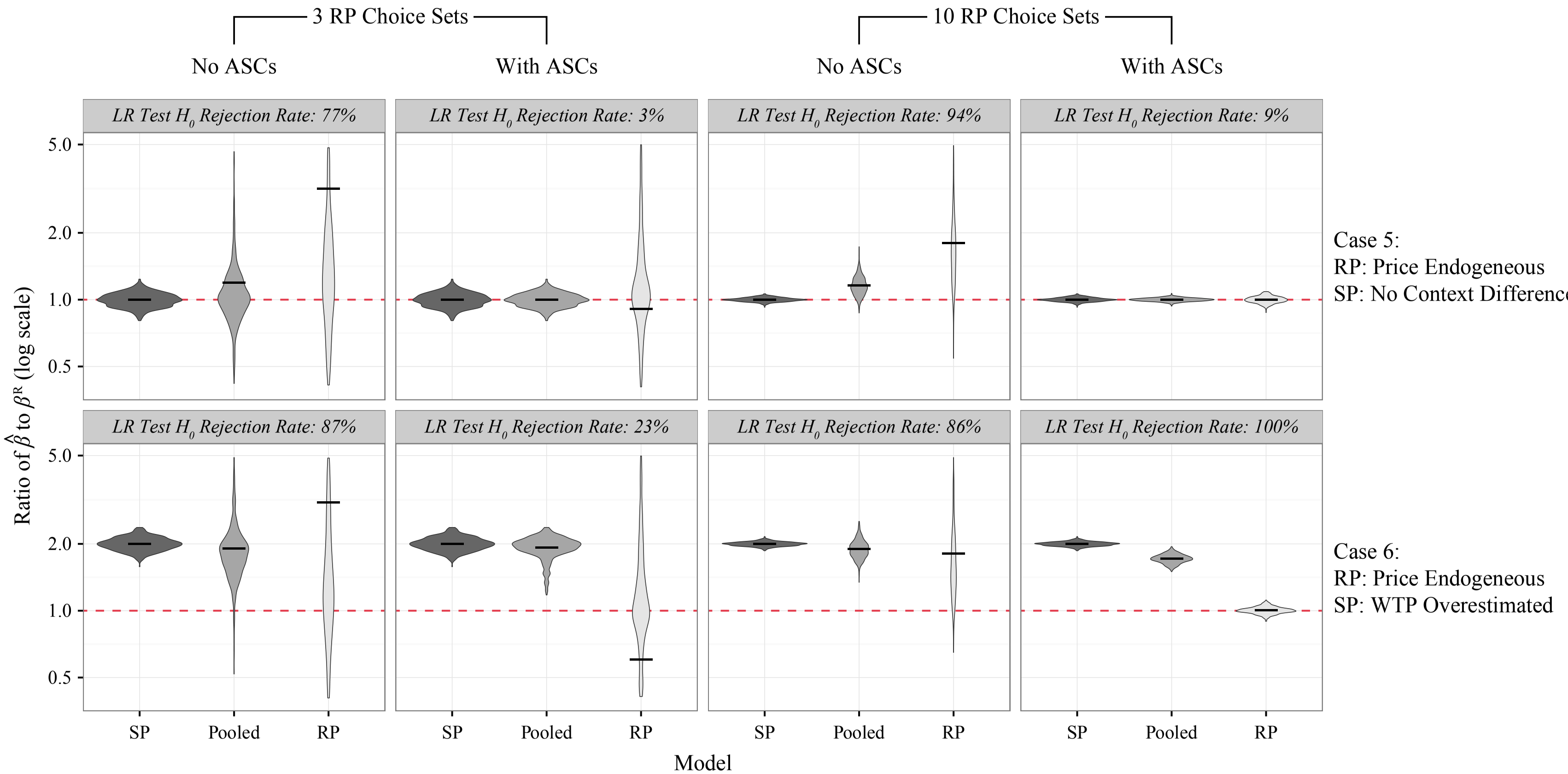


Figure 7: SP, RP, and pooled model results of the ratio of $\hat{\beta}$ to β^{R} for test cases 5 and 6. In each test case, we compare the impact of using ASCs to correct the RP price endogeneity as well as having fewer versus more choice sets. Each plot represents results from 1,000 simulated data sets with the sampling distribution visualized as a kernel density plot. For each data set, the LR test of pooling is conducted at 10% significance.

For test case 5 (the top row), including the ASCs corrects the endogeneity in the RP data and the LR test performs as expected, accurately rejecting pooling without the endogeneity correction and accepting pooling with it. For test case 6 (the bottom row), the effect of including the ASCs is sensitive to the number of RP choice sets. With 10 choice sets (the lower-right boxes), including the ASCs corrects the endogeneity bias and the LR test accurately rejects pooling. However, with only 3 RP choice sets (the lower-left boxes), the addition of the ASCs does correct the endogeneity bias, but due to the low number of choice sets the LR test largely fails to reject pooling, only rejecting in 23% of the simulations. In this particular situation, the modeler may falsely conclude that the RP and SP parameters are the same after correcting the endogeneity when they are in fact not, resulting in a biased pooled WTP estimate driven by the SP data.

This simulation result illustrates several drawbacks of using ASCs to correct the endogeneity bias in pooled models. First, this approach requires that the omitted variable be associated with an alternative that is observed repeatedly over multiple choice sets; otherwise it is not identified. Further, the prices and other observed attributes must vary across choice sets yet still be correlated with the unobserved variable, and the unobserved variable must be constant across choice sets. Even under these conditions, using data with fewer choice sets and less variation in pricing, we found the estimator can be unstable.[15] Second, if the WTP parameters are weakly identified by the RP data, the LR test may fail to reject pooling, providing false confidence that the RP and SP parameters are the same.

In summary, endogeneity corrections could work in a pooled model so long as the required assumptions of the endogeneity correction approach are satisfied and appropriate data are available. However, even if a data set has a structure that satisfies these strong assumptions, the endogeneity correction could tilt the information balance between the RP and SP data. This balance could be checked using the information balance statistics proposed in Section 6.1.

6.3 Limitations

We focus on (1) price endogeneity from omitted variables and (2) contextual differences as two specific issues that affect parameter estimates, but there are a number of other modeling concerns that we have not addressed, such as other forms of model misspecification and measurement error that can also lead to biased parameter estimates. We also do not address concerns with state-dependence effects (e.g. when the RP choices an individual makes influences his or her SP choices) or serial correlations across multiple responses in cases where the RP and SP respondents are the same (Bhat and Castelar, 2002; Morikawa, 1994).

Our simulation experiment makes several simplifying assumptions. For example, the homogeneous mulitnomial logit model used in this experiment has the Independence of Irrelevant Alternatives (IIA) property (Train, 2009). While it is unclear how pooled models will be affected by more flexible substitution patterns such as mixed logit models (McFadden and Train, 2000; Brownstone et al., 2000) or hierarchical models (Feit et al., 2010), we expect our general observations will hold, since misspecified models with endogenous parameters will produce biased parameter estimates regardless of the model structure if the endogeneity is not corrected.

7 Conclusions

Using a synthetic data experiment, we test the performance of pooled RP-SP models in recovering true market preference parameters when (1) there is potential for endogeneity problems in the RP data and (2) when consumer willingness to pay for attributes from the survey context, β^{S}, may differ from that of the market context, β^{R}.

[15] The maximum likelihood estimator for the multinomial logit model (or any limited dependent variable model) with fixed-effects is inconsistent if the number of choice sets is held fixed (Greene, 2004), i.e. the estimator does not converge to the true value as the number of observations increases and the number of choice sets is held fixed.

Our results suggest that modelers considering pooling SP and RP data for parameter estimation should first assess whether each data source can provide unbiased estimates of the true market preference parameters β^{R}. The SP data may provide good estimates of β^{S} but poor estimates of β^{R} if consumers respond differently in the survey context than in a market context ($\beta^{\mathrm{R}} \neq \beta^{\mathrm{S}}$), and the RP data may provide poor estimates of β^{R} if unobserved attributes correlated with observed attributes create endogeneity bias. If either data set alone produces biased estimates of β^{R}, pooling is unlikely to improve parameter estimates. Furthermore, the LR test may not offer clear guidance about whether or not to pool when there is a possibility that both sources of bias are in the same direction (Case 6 in our simulation experiment). We also introduced a new metric for assessing information balance that can help the modeler understand the extent to which parameter estimates are informed by each data source.

Addressing the source of bias (e.g. by providing appropriate respondent incentives in the SP context and by controlling for endogeneity when modeling the RP data) is a potential solution. However, even under ideal conditions, endogeneity correction may be an imperfect solution unless all of the data requirements and assumptions of the endogeneity correction are met. Endogeneity corrections may also alter the balance of information between the two data sets, which could be assessed using the proposed information balance metric.

Even when no endogeneity bias is present, the chi-squared LR test that has been widely used to justify pooling relies on the assumption that errors are IID. If the RP data contain repeated choice observations in only a small number of RP choice sets, as is common in practice, the errors may not be close to IID, and the LR test may produce misleading results, such as artificially inflating the rejection rate of the null pooling hypothesis. Alternatively, though it does not offer a statistical test, a scatterplot of the estimates of $\hat{\beta}^{\mathrm{S}}$ from the SP model and $\hat{\beta}^{\mathrm{R}}$ from the RP model can help the modeler assess whether pooling is supported and identify any attributes that should not be pooled, although this approach may still be unable to identify the case where parameter biases are in the same direction (Case 6 in our simulation experiment).

When the goal is to build a model that produces unbiased marketplace preference parameter estimates, modelers should carefully consider the potential for statistical bias in the RP data before making pooling choices, since pooling does not necessarily reduce these biases and in some cases can make them worse. In the absence of statistical biases in the RP data or context differences in the SP data, pooling data sources provides many useful purposes, such as adding additional information about the parameters, reducing multicollinearity, and allowing the incorporation of attributes that do not appear in the market.

Acknowledgments

The authors would like to thank the three reviewers who provided excellent feedback that improved the study. This work was supported in part by a grant from the Link Foundation, a grant from the National Science Foundation #1064241, and a grant from Ford Motor Company. The opinions expressed are those of the authors and not necessarily those of the sponsors.

References

W. Adamowicz, J. J. Louviere, and M. Williams. Combining revealed and stated preference methods for valuing environmental amenities. *J. Environ. Econ. Manage.*, 26:271–292, 1994.

W. Adamowicz, J. Swait, P. Boxall, J. Louviere, and M. Williams. Perceptions vs Objective Measures of Environmental Quality in Combined Revealed-Stated Preference Models of Environmental Valuation. *J. Environ. Econ. Manage.*, 32:65–84, 1997.

A. A. Ahern and N. Tapley. The use of stated preference techniques to model modal choices on interurban trips in Ireland. *Transp. Res. Part A Policy Pract.*, 42(1):15–27, 2008. ISSN 09658564. doi: 10.1016/j.tra.2007.06.005.

J. Axsen, D. C. Mountain, and M. Jaccard. Combining stated and revealed choice research to simulate the neighbor effect: The case of hybrid-electric vehicles. *Resour. Energy Econ.*, 31(3):221–238, aug 2009. ISSN 09287655. doi: 10.1016/j.reseneeco.2009.02.001. URL `http://linkinghub.elsevier.com/retrieve/pii/S0928765509000037`.

M. J. Beck, S. Fifer, and J. M. Rose. Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. *Transportation Research Part B: Methodological*, 89:149–167, 2016. ISSN 01912615. doi: 10.1016/j.trb.2016.04.004.

M. Ben-Akiva and T. Morikawa. Estimation of Switching Models from Revealed Preferences and Stated Intentions. *Transp. Res. Part A Policy Pract.*, 24A(6):485–495, 1990.

M. Ben-Akiva, M. Bradley, T. Morikawa, J. Benjamin, T. Novak, H. Oppewal, and V. Rao. Combining Revealed and Stated Preferences Data. *Mark. Lett.*, 5(4):335–350, 1994.

S. Berry, J. Levinsohn, and A. Pakes. Automobile Prices in Market Equilibrium. *Econometrica*, 63(4):841–890, 1995.

C. R. Bhat and S. Castelar. A unified mixed logit framework for modeling revealed and stated preferences: Formulation and application to congestion pricing analysis in the San Francisco Bay area. *Transp. Res. Part B Methodol.*, 36(7):593–616, 2002. ISSN 01912615. doi: 10.1016/S0191-2615(01)00020-0.

E. Birol, A. Kontoleon, and M. Smale. Combining revealed and stated preference methods to assess the private value of agrobiodiversity in Hungarian home gardens. Technical Report 156, International Food Policy Research Institute, 2006.

M. Börjesson. Joint RP-SP data in a mixed logit analysis of trip timing decisions. *Transp. Res. Part E Logist. Transp. Rev.*, 44(6):1025–1038, 2008. ISSN 13665545. doi: 10.1016/j.tre.2007.11.001.

K. Brooks and J. L. Lusk. Stated and revealed preferences for organic and cloned milk: Combining choice experiment and scanner data. *American Journal of Agricultural Economics*, 92 (4):1229–1241, 2010. ISSN 00029092. doi: 10.1093/ajae/aaq054.

D. Brownstone and K. a. Small. Valuing time and reliability: Assessing the evidence from road pricing demonstrations. *Transp. Res. Part A Policy Pract.*, 39(4 SPEC. ISSS.):279–293, 2005. ISSN 09658564. doi: 10.1016/j.tra.2004.11.001.

D. Brownstone and K. E. Train. Forecasting new product penetration with flexible substitution patterns. *J. Econom.*, 89, 1999.

D. Brownstone, D. S. Bunch, and K. E. Train. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. *Transp. Res. Part B*, 34(5): 315–338, jun 2000. ISSN 01912615. doi: 10.1016/S0191-2615(99)00031-4. URL `http://linkinghub.elsevier.com/retrieve/pii/S0191261599000314`.

R. T. Carson and T. Groves. Incentive and informational properties of preference questions. *Environ. Resour. Econ.*, 37(1):181–210, 2007.

M. Ding, R. Grewal, and J. Liechty. Incentive-aligned conjoint analysis. *J. Mark. Res.*, 42(1): 67–82, 2005.

D. Dissanayake and T. Morikawa. A Combined RP / SP Nested Logit Model of Vehicle Ownership, Mode Choice and Trip Chaining to Investigate Household Travel Behavior in Developing Countries. In *82nd Annu. Meet. Transp. Res. Board*. Transportation Research Board, 2003. ISBN 8152789356.

E. M. Feit, M. A. Beltramo, and F. M. Feinberg. Reality Check: Combining Choice Experiments with Market Data to Estimate the Importance of Product Attributes. *Manage. Sci.*, 56(5): 785–800, 2010. doi: 10.1287/mnsc.1090.1136.

W. Greene. The behaviour of the maximum likelihood estimator of limited dependent variable models in the presence of fixed effects. *The Econometrics Journal*, 7(1):98–119, 2004.

C. A. Guevara. Critical assessment of five methods to correct for endogeneity in discrete-choice models. *Transp. Res. Part A Policy Pract.*, 82: 240–254, 2015. ISSN 09658564. doi: 10.1016/j.tra.2015.10.005. URL `http://linkinghub.elsevier.com/retrieve/pii/S0965856415002554`.

C. G. Haaf, W. R. Morrow, I. M. Azevedo, E. M. Feit, and J. J. Michalek. Forecasting light-duty vehicle demand using alternative-specific constants for endogeneity correction versus calibration. *Transportation Research Part B: Methodological*, 84:182–210, 2016.

M. Haghani and M. Sarvi. Stated and revealed exit choices of pedestrian crowd evacuees. *Transp. Res. Part B*, 95:238–259, 2017. ISSN 0191-2615. doi: 10.1016/j.trb.2016.10.019. URL `http://dx.doi.org/10.1016/j.trb.2016.10.019`.

N. Hardt, Y. Kim, M. Joo, J. Kim, and G. M. Allenby. Reconciling Stated and Revealed Preferences. Working Paper, 2017.

D. A. Hensher. Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B: Methodological*, 44(6):735–752, 2010. ISSN 01912615. doi: 10.1016/j.trb.2009.12.012. URL `http://dx.doi.org/10.1016/j.trb.2009.12.012`.

D. A. Hensher and M. Bradley. Using stated response data to enrich revealed preference discrete choice models. *Mark. Lett.*, 4:2(1993):139–151, 1993.

D. A. Hensher, J. Louviere, and J. Swait. Combining sources of preference data. *J. Econom.*, 89:197–221, 1999.

D. A. Hensher, J. M. Rose, and W. H. Greene. Combining RP and SP data: biases in using the nested logit 'trick' - contrasts with flexible mixed logit incorporating panel and scale effects. *J. Transp. Geogr.*, 16(2):126–133, 2008. ISSN 09666923. doi: 10.1016/j.jtrangeo.2007.07.001.

J.-c. Huang, T. C. Haab, and J. C. Whitehead. Willingness to Pay for Quality Improvements: Should Revealed and Stated Preference Data Be Combined? *J. Environ. Econ. Manage.*, 34: 240–255, 1997.

J. Huber and K. Zwerina. The Importance of Utility Balance in Efficient Choice Designs. *J. Mark. Res.*, 33(3):307–317, 1996.

M. Lavasani, M. S. Hossan, H. Asgari, and X. Jin. Examining methodological issues on combined RP and SP data. *Transp. Res. Procedia*, 25:2335–2348, 2017. ISSN 23521465. doi: 10.1016/j.trpro.2017.05.218. URL `http://dx.doi.org/10.1016/j.trpro.2017.05.218`.

J. B. Loomis. What's To Know About Hypothetical Bias in Stated Preference Valuation Studies? *J. Econ. Surv.*, 25(2):363–370, 2011. ISSN 10685502. doi: 10.1111/j.1467-6419.2010.00675.x.

J. J. Louviere, D. A. Hensher, and J. Swait. *Stated Choice Methods: Analysis and Applications.* Cambridge University Press, Cambridge, Massachusetts, 2000.

T. Mark and J. Swait. Using stated preference and revealed preference data fusion modeling in health care. In *Using Discret. Choice Exp. to Value Heal. Heal. Care*, chapter 10, pages 217–234. 2008. ISBN 978-1-4020-4082-5. doi: 10.1007/978-1-4020-5753-3.

T. L. Mark and J. Swait. Using stated preference and revealed preference modeling to evaluate prescribing decisions. *Health Econ.*, 13(6):563–573, 2004. ISSN 1057-9230. doi: 10.1002/hec.845. URL http://doi.wiley.com/10.1002/hec.845.

D. McFadden and K. E. Train. Mixed MNL models for discrete response. *J. Appl. Econom.*, 15(5):447–470, sep 2000. ISSN 0883-7252. doi: 10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1. URL http://doi.wiley.com/10.1002/1099-1255(200009/10)15:5<447::AID-JAE570>3.0.CO;2-1.

T. Morikawa. *Incorporating stated preference data in travel demand analysis.* PhD thesis, Massachusetts Institute of Technology, 1989.

T. Morikawa. Correcting state dependence and serial correlation in the RP / SP combined estimation method. *Transportation (Amst).*, 21(2):153–165, 1994.

A. Petrin and K. Train. A control function approach to endogeneity in consumer choice models. *Journal of Marketing Research*, 47(1):3–13, 2010.

A. Polydoropoulou and M. Ben-Akiva. Combined revealed and stated preference nested logit access and mode choice model for multiple mass transit technologies. *Transp. Res. Rec. J. Transp. Res. Board*, 1771:38–45, 2001.

C. Román, R. Espino, and J. C. Martín. Competition of high-speed train with air transport: The case of Madrid–Barcelona. *J. Air Transp. Manag.*, 13(5): 277–284, 2007. ISSN 09696997. doi: 10.1016/j.jairtraman.2007.04.009. URL http://linkinghub.elsevier.com/retrieve/pii/S0969699707000452.

P. E. Rossi. Even the rich can make themselves poor: A critical examination of iv methods in marketing applications. *Marketing Science*, 33(5):655–672, 2014.

K. A. Small, C. Winston, and J. Yan. Uncovering the distribution of motorists' preferences for travel time and reliability. *Econometrica*, 73(4):1367–1382, 2005. ISSN 00129682. doi: 10.1111/j.1468-0262.2005.00619.x.

G. Sonnier, A. Ainslie, and T. Otter. Heterogeneity distributions of willingness-to-pay in choice models. *Quant. Mark. Econ.*, 5(3):313–331, aug 2007. ISSN 1570-7156. doi: 10.1007/s11129-007-9024-6. URL http://link.springer.com/10.1007/s11129-007-9024-6.

J. Swait and R. Andrews. Enriching scanner panel models with choice experiments. *Mark. Sci.*, 22(3):442–460, 2003. ISSN 0732-2399. doi: 10.1287/mksc.22.4.442.24910. URL http://pubsonline.informs.org/doi/abs/10.1287/mksc.22.4.442.24910.

J. Swait and J. Louviere. The Role of the Scale Parameter in the Estimation and Comparison of Multinomial Logit Models. *J. Mark. Res.*, 30(3):305–314, 1993.

J. Swait, J. J. Louviere, and M. Williams. A sequential approach to exploiting the combined strengths of SP and RP data: Application to freight shipper choice. *Tranportation*, 21:135–152, 1994.

O. Toubia and J. R. Hauser. Research Note On Managerially Efficient Experimental Designs. *Marketing Science*, 26(6):851–858, 2007. ISSN 0732-2399. doi: 10.1287/mksc.1060.0244.

K. E. Train. *Discrete Choice Methods with Simulation.* Cambridge University Press, 2nd edition, 2009.

K. E. Train and M. Weeks. Discrete Choice Models in Preference and Willingness-to-Pay Space. In *Appl. Simul. Methods Environ. Resour. Econ.*, chapter 1, pages 1–16. 2005. URL `http://link.springer.com/chapter/10.1007/1-4020-3684-1_1`.

T. Truong, W. Adamowicz, and P. C. Boxall. Modelling the Effect of Chronic Wasting Disease on Recreational Hunting Site Choice Preferences and Choice Set Formation over Time. *Environ. Resour. Econ.*, pages 1–25, 2017. ISSN 15731502. doi: 10.1007/s10640-017-0120-0.

J. M. Villas-Boas and R. S. Winer. Endogeneity in brand choice models. *Management Science*, 45(10):1324–1338, 1999.

R. H. von Haefen and D. J. Phaneuf. Identifying demand parameters in the presence of unobservables: A combined revealed and stated preference approach. *Journal of Environmental Economics and Management*, 56(1):19–32, 2008. ISSN 00950696. doi: 10.1016/j.jeem.2008.01.002.

P. Wicker, J. C. Whitehead, B. K. Johnson, and D. S. Mason. The effect of sporting success and management failure on attendance demand in the Bundesliga: a revealed and stated preference travel cost approach. *Appl. Econ.*, 49 (52):5287–5295, 2017. ISSN 14664283. doi: 10.1080/00036846.2017.1305090. URL `https://doi.org/10.1080/00036846.2017.1305090`.

J. M. Wooldridge. *Econometric analysis of cross section and panel data.* MIT press, 2010.

J. M. Wooldridge. *Introductory econometrics: A modern approach.* Nelson Education, 2015.

8 Supplemental Information

8.1 Table of Previous Literature on Pooled RP-SP Models

Table 5: Summary of previous literature on pooled RP-SP models

Study	Pooling Motivation	Conclusions
Adamowicz et al. (1994)	Improve RP collinearity; include non-existing attributes	SP improves “quality” of RP estimates; RP collinearity reduced; attribute ranges not presently available analyzed
Adamowicz et al. (1997)	Improve RP collinearity; include non-existing attributes; Compare perceived vs. objective attributes	Respondent decisions based on attribute perceptions rather than objective value
Ahern and Tapley (2008)	RP data for actual travel data, SP data for non-existing alternatives.	Differences between RP and SP parameters were not sufficient to reject the hypothesis of parameter equality, and the pooling...hsould allow the advantages of each to be maximised and the disadvantages to be minimised.
Axsen et al. (2009)	Improve RP collinearity; include non-existing attributes; RP adds realism	Greater RP influence, better on statistical measures; greater SP influence, more realistic WTP
Ben-Akiva and Morikawa (1990)	Include non-existing attributes; RP corrects SP biases	“...combined estimation...can be used to exploit their advantages. In particular, the combined estimation explicitly identifies the differences between the RP and SP data generating processes.”
Bhat and Castelar (2002)	Add flexible substitution patterns (MXL model) and state dependency in joint models	Suggest “using SP experiments as the main data source for analysis and supplementing with small samples of RP data for anchoring with actual market activity.”
Birol et al. (2006)	Improve RP collinearity; include non-existing attributes	“...combined estimation enables more robust and efficient identification”
Börjesson (2008)	Improve RP collinearity; SP data might be less trustworthy for trip timing	Different scheduling disutility across RP and SP choices imply temporal differences in RP and SP choice situations
Brooks and Lusk (2010)	Improve RP collinearity; include non-existing attributes	“RP data...critical for obtaining realistic body-type choice and scaling information...SP data are critical for obtaining information about attributes not available in the marketplace, but pure SP models...give implausible forecasts.”
Brownstone et al. (2000)	Improve RP collinearity; include non-existing attributes	Pooled model performs better in out-of-sample prediction. Pooled coefficients used for consumer valuation.

Dissanayake and Morikawa (2003)	Include non-existing attributes; "to improve the accuracy of parameter estimates while exploiting the advantages of both RP and SP [data]"	"It has been clearly observed that the combined estimation of RP and SP data in travel demand modeling is an effective technique for expressing complex travel behavior and forecasting the travel demand for new transport services."
Feit et al. (2010)	Adjusting conjoint parameters to be more consistent with observed market choices	Joint model benefits from well-conditioned conjoint data and predicts market data much better than conjoint model
Haghani and Sarvi (2017)	Most work on crowd evacuee exit choices based on SP data, need to verify consistency with RP context.	Despite some similarities and differences between RP and SP parameters, model predictions were similar.
Hensher and Bradley (1993)	Introduces the use of a nested model to estimate scale differences (the FIML procedure)	"...additional information from SP data gives the RP model increased richness and sensitivity for prediction. The use of SP data to estimate alternative-specific constants for new products is a major contribution to enriching an RP application in the presence of a new alternative."
Hensher et al. (1999)	Surveys past studies on joint RP-SP approaches, in particular different error structures.	"...we are able to utilise the well-behaved SP design matrix to correct sign and collinearity problems in the RP data...we are also able to obtain more robust parameter estimates...The availability of the RP data...contributes a 'real-world flavour' to the joint model by establishing alternative-specific constants that reflect population characteristics."
Hensher et al. (2008)	"This paper promotes the replacement of the NL 'trick' method with an error components model that can accommodate correlated observations as well as reveal the relevant scale parameter for subsets of alternatives."	"The nested logit approach is...not capable of accounting for the potential correlation induced through repeated observations on one or more pooled data sets."
Huang et al. (1997)	"The purpose of this paper is to demonstrate the conditions for consistently combining revealed (trip demands) and stated (contingent valuation) data for an improvement in environmental quality."	"Our results show that revealed and stated data should not be combined under the same assumed preference structure unless the two decisions imply the same change in behavior induced by the quality change."
Lavasani et al. (2017)	Consider scale parameters, error components, and state dependency factors in pooled RP-SP models.	"model results show significant SP to RP scale parameter indicating higher variances in the SP data."
Mark and Swait (2004)	Improve RP collinearity; include non-existing attributes; Improve the ability to evaluate choice of health care products	"This paper illustrates how SP data (hypothetical prescription choices)...and RP data...(perceived medication attributes and reported medication usage) can be employed to understand the factors influencing physician prescribing decisions."

Mark and Swait (2008)	Improve RP collinearity; include non-existing attributes; "our interest will be in improved prediction of some form of market behaviour."	"Data enrichment allows one to capitalise on the realism of actual health care choices with the favourable statistical characteristics of hypothetical choices"
Morikawa (1994)	Parameter efficiency, "bias correction," and identification of new attributes	Develops a method for correcting state dependence and serial correlation in the RP / SP combined estimation method.
Poly-doropoulou and Ben-Akiva (2001)	"[pooling]...provides more reliable estimation results because the RP data...counteract the SP-related biases, and it provides the capability to estimate the demand for new mass transit technologies."	"The results demonstrate the...advantages of simultaneously estimating models with different data sets and sharing common coefficients."
Román et al. (2007)	RP data for actual travel data, SP data for non-existing alternatives	"results...cast doubts on the competition that HSTs can exert in markets characterized by high-frequency air services."
Small et al. (2005)	Improve RP collinearity; include non-existing attributes; improve statistically precision	"...we are able to measure properties of travel preferences that have eluded other studies. We find that travel time and its predictability are highly valued by motorists and that there is significant heterogeneity in these values."
Swait et al. (1994)	New modeling approach (sequential) to "exploit the strengths and avoiding the weaknesses of each data source."	RP data ground model in reality (with ASCs), SP data help reduce statistical problems in RP data; "...choice forecasts of the Sequential model are practically indistinguishable from those of the RP-only model, and in quite a few cases actually improved over the performance of the latter."
Swait and Andrews (2003)	"The fact that different choice data sources have diverse strengths and weaknesses suggests it might be possible to pool multiple sources to achieve improved models, due to offsetting advantages and disadvantages."	Pooled model performed better on holdout predictions even though the LR test rejected parameter homogeneity.
Truong et al. (2017)	Increase sample size, incorporate choice set formation.	"Including choice set formation improves the statistical properties of choice models and generates welfare measures that differ from choice models that exclude choice set generation."
von Haefen and Phaneuf (2008)	Improve RP collinearity; include non-existing attributes; improve statistically precision; correct for endogenous RP parameters	"...our combined RP/SP approach to identifying preference parameters in the presence of unobserved determinants of choice represents a feasible and in many ways attractive alternative to RP approaches."
Wicker et al. (2017)	Increase sample size; include non-existing attributes.	The likelihood ratio test statistic...indicates that the restriction is statistically appropriate.

8.2 Supplemental Figures

Base Case: All Pooled Model Parameters

Figure 8 shows the ratio of estimated to true parameters for all of the pooled model parameters. Depending on the case, the modeler could make false conclusions about both consumer WTP for attributes ($\hat{\beta}$) as well as how consistently consumers make choices in the RP versus SP contexts ($\hat{\lambda}^{\mathrm{R}}$ and $\hat{\lambda}^{\mathrm{S}}$). For example, in case 2 the estimated RP scale parameter is less than it's true parameter $\lambda^{\mathrm{R}} = 1$ because the omitted unobserved variable increases the variance of the error term, thus decreasing scale. The presence of endogeneity exacerbates the effect; in particular, in test case 6 when the LR test would largely accept pooling, the modeler may falsely conclude that respondents make much more consistent choices in the RP context than SP context.

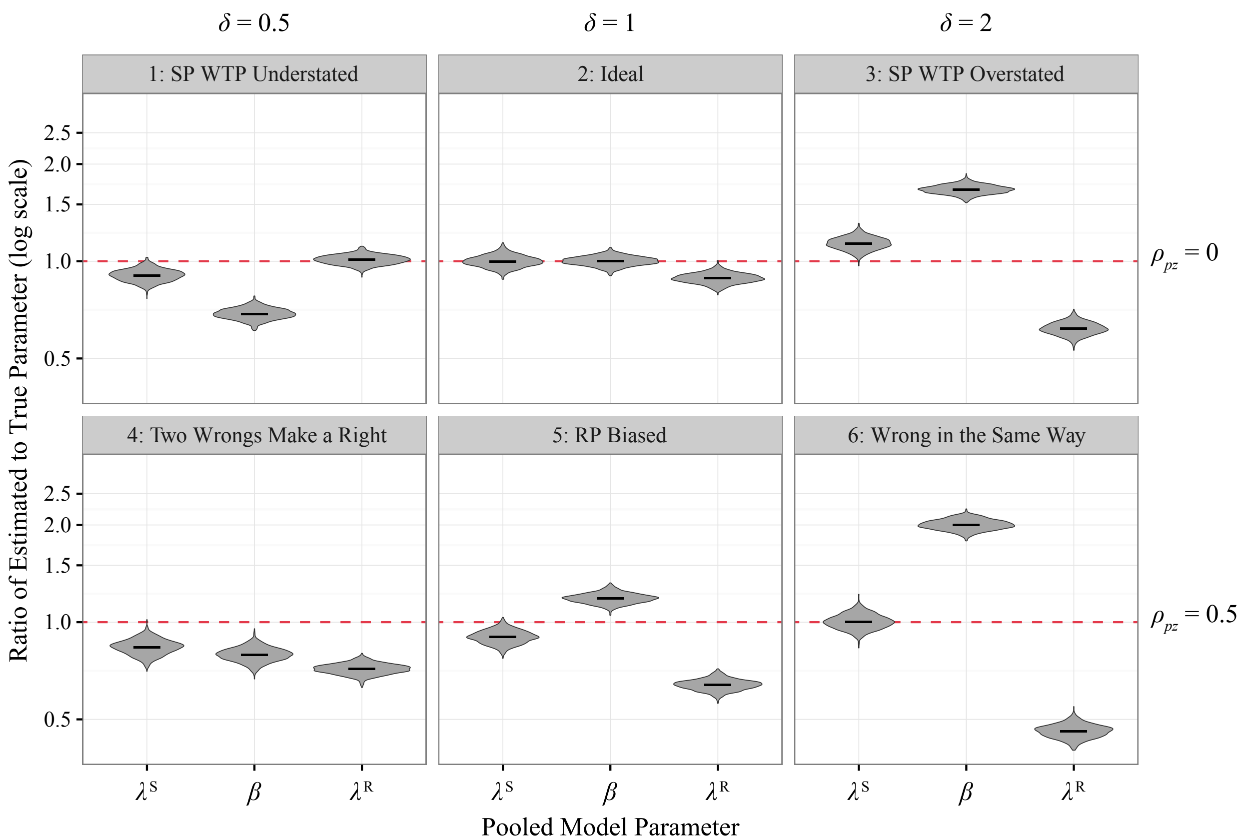


Figure 8: Ratio of estimated to true parameters for all pooled model parameters in each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. Sampling variation is visualized with a kernel density plot. The red dashed line shows the true parameter values of the data-generating function.

Sensitivity Case: Multiple Pooled Parameters

Figure 9 shows the WTP coefficients from 1,000 simulations of the model with five pooled attributes. Results are nearly identical to those in our base case except that the additional WTP coefficients for x_2 through x_5 in cases 4 through 6 are also biased upward due to the fact that price endogeneity affects all WTP coefficients. We also ran a case with five common attributes, three RP-specific attributes, and three SP-specific attributes, and results were again are similar to those in Figure 9.

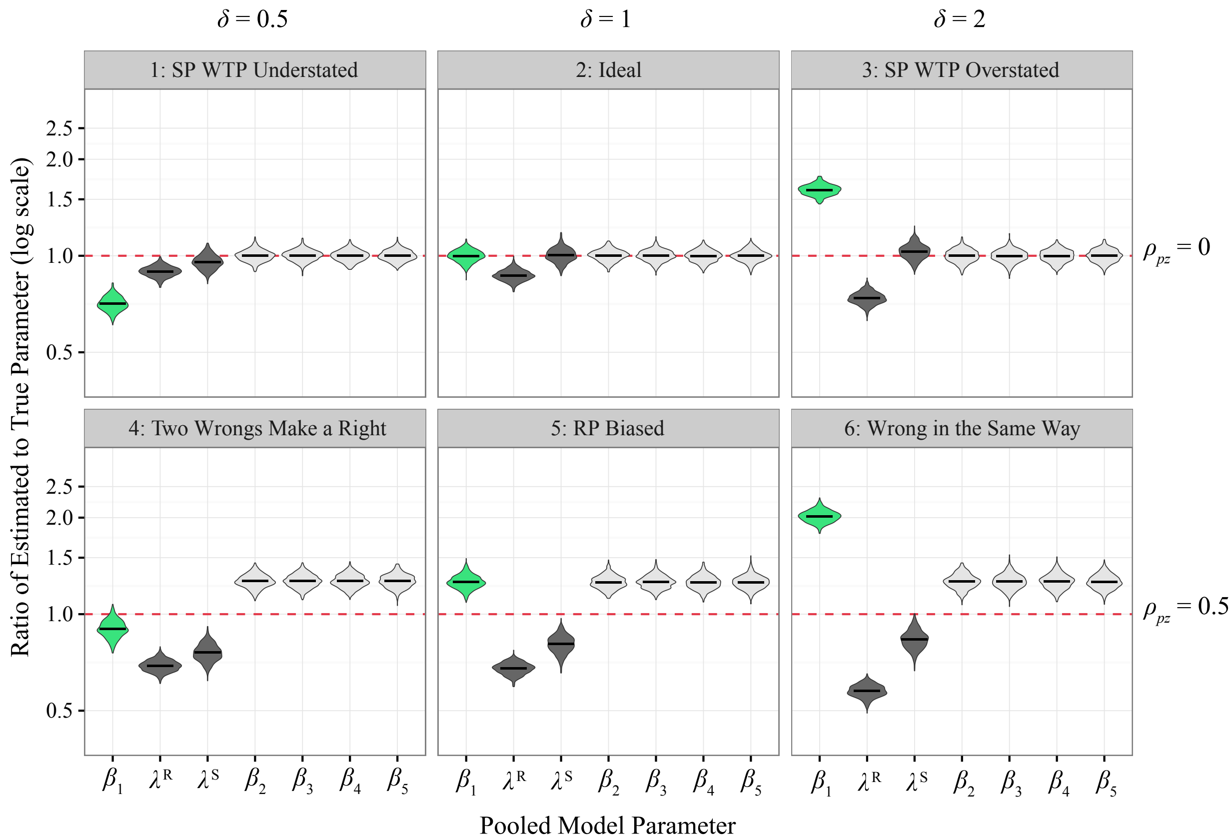


Figure 9: Ratio of estimated to true parameters for pooled model with 5 pooled attributes. The context effect controlled by δ only affects β_1. Each box plot represents 1,000 simulations using the parameters in Table 2. Sampling variation is visualized with a kernel density plot. The red dashed line shows the true parameter values of the data-generating function.

8.3 Base Case Results in the Preference Space

The Pooled Model in the Preference Space

For readers more familiar with the preference space utility model, we conduct our base case simulation in the preference space for comparison with the results in the WTP space in Section 4. We begin by specifying the utility for SP and RP data in the preference space, which follows the form of equation (3):

$$u_j^{\mathrm{S}} = \boldsymbol{\beta}^{\mathrm{S}'}\mathbf{x}_j - \alpha^{\mathrm{S}} p_j + \boldsymbol{\gamma}^{\mathrm{S}'}\mathbf{y}_j^{\mathrm{S}} + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{33}$$

$$u_j^{\mathrm{R}} = \boldsymbol{\beta}^{\mathrm{R}'}\mathbf{x}_j - \alpha^{\mathrm{R}} p_j + \boldsymbol{\gamma}^{\mathrm{R}'}\mathbf{y}_j^{\mathrm{R}} + \boldsymbol{\zeta}'\mathbf{z}_j + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{34}$$

As in Section 2.2, we separate out the attributes into two vectors: $\mathbf{x}_j$ represents attributes that are common between the RP and SP data sets, and $\mathbf{y}_j$ represents those attributes observed in only one of the contexts and not the other (price, p_j, is noted separately from $\mathbf{x}_j$ and exists in both contexts). Finally, $\mathbf{z}_j$ represents attributes unobserved by the modeler and is only present in the RP utility expression.

Recall that the parameters α and $\boldsymbol{\beta}$ in the utility expressions in equations (33) and (34) are scaled by the scale of the error term ($\lambda = 1/\sigma$) to achieve an identifiable utility model with a standardized error term $\varepsilon_j = (\varepsilon_j^*/\sigma)$, where ε_j^* is the unscaled error term with variance $\sigma^2(\pi^2/6)$. To account for potential scale differences between the RP and SP data, the pooled model restricts (33) and (34) such that $\lambda^{\mathrm{S}}\alpha^{\mathrm{S}} = \lambda^{\mathrm{R}}\alpha^{\mathrm{R}}$ and $\lambda^{\mathrm{S}}\boldsymbol{\beta}^{\mathrm{S}} = \lambda^{\mathrm{R}}\boldsymbol{\beta}^{\mathrm{R}}$. Since both λ^{S} and λ^{R} terms are not separately identifiable, a "scale ratio" term is defined as $\lambda = \lambda^{\mathrm{S}}/\lambda^{\mathrm{R}}$. The resulting pooled utility specification is given by:

$$u_j^{\mathrm{S}} = \lambda\left(\boldsymbol{\beta}'\mathbf{x}_j - \alpha p_j + \boldsymbol{\gamma}^{\mathrm{S}'}\mathbf{y}_j^{\mathrm{S}}\right) + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{35}$$

$$u_j^{\mathrm{R}} = \boldsymbol{\beta}'\mathbf{x}_j - \alpha p_j + \boldsymbol{\gamma}^{\mathrm{R}'}\mathbf{y}_j^{\mathrm{R}} + \boldsymbol{\zeta}'\mathbf{z}_j + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{36}$$

where α and $\boldsymbol{\beta}$ are now modeled as parameters common to the two utility models and λ accounts for potential differences in error scaling.

Base Case Simulation in the Preference Space

To conduct our base case simulation experiment in the preference space, we mirror the same data generation process as described in Section 3.1 except we use different equations for the SP and RP utility functions. Specifically, for the SP and RP data, we substitute the following preference space utility models for equations (14) and (17):

$$u_j^{\mathrm{S}} = \beta^{\mathrm{S}}x_j - \alpha^{\mathrm{S}}p_j + \varepsilon_j^{\mathrm{S}}, \qquad \varepsilon_j^{\mathrm{S}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{37}$$

$$u_j^{\mathrm{R}} = \beta^{\mathrm{R}}x_j - \alpha^{\mathrm{R}}p_j + \zeta z_j + \varepsilon_j^{\mathrm{R}}, \qquad \varepsilon_j^{\mathrm{R}} \sim \text{Gumbel}\left(0, \frac{\pi^2}{6}\right) \tag{38}$$

SP context differences are defined in the same way as the WTP space model where SP WTP is equal to the RP WTP parameter scaled by δ:

$$\left(\beta^{\mathrm{S}}/\alpha^{\mathrm{S}}\right) = \delta\left(\beta^{\mathrm{R}}/\alpha^{\mathrm{R}}\right) \tag{39}$$

We then estimate a pooled model where the z term is omitted from the RP utility specification as an unobserved variable:

$$u_j^{\mathrm{S}} = \hat{\lambda}\left(\hat{\beta}x_j - \hat{\alpha}p_j\right) + \varepsilon_j^{\mathrm{S}} \tag{40}$$

$$u_j^{\mathrm{R}} = \hat{\beta}x_j - \hat{\alpha}p_j + \varepsilon_j^{\mathrm{R}} \tag{41}$$

where $\hat{\alpha}$ and $\hat{\beta}$ are common between the two models, and the hats on the parameters indicate that they are estimated.

Base Case Results in the Preference Space

We simulate 1,000 sets of RP and SP data for each of the six test cases in Table 3 and then use them to estimate RP, SP, and pooled models. Figures 10 and 11 show the ratio between the estimated and true model coefficients for α^{R} and β^{R}, respectively, for each test case using the base case parameters in Table 2 (we use 1 for α^{S} and α^{R}). Figure 12 shows the results of all pooled model parameters (λ, α, and β). Finally, Figure 13 shows the ratio of the computed WTP from the estimates of the preference space models ($\hat{\beta}/\hat{\alpha}$) to the true computed WTP in the RP context ($\beta^{\mathrm{R}}/\alpha^{\mathrm{R}}$). The plots show the distribution of the ratios across 1,000 simulations using a logarithmic y-axis for comparing ratios.

These results can initially be challenging to interpret. Figure 10 shows that for the price parameter, $\hat{\alpha}$, the estimates are biased as would be expected, with the pooled model estimate generally falling between the SP-only and RP-only estimates. However, Figure 11 shows that the pooled model estimate for β is sometimes *lower* than either the SP-only or RP-only estimate.

To interpret these trends, we must simultaneously consider the role of the estimated scale ratio parameter in these models. Figure 12 shows the estimates for $\hat{\lambda}$, $\hat{\alpha}$ and $\hat{\beta}$ for the pooled model. The estimates for case 1, where the SP WTP is understated, show that the scale ratio ($\lambda = \lambda^{\mathrm{S}}/\lambda^{\mathrm{R}}$) is estimated to be about 1.75. When the WTP is altered by δ, it also affects the relative error scale between the two data sources. When we multiply the estimated $\hat{\beta}$ by the estimated error scale $\hat{\lambda}$, we approximately recover the true β^{R}. Comparing the pooled model results in the preference space in Figure 12 to those in the WTP space in Figure 8, one can see that the WTP space pooled model also has biased scale parameters due to the SP context difference, but this only affects the estimates of the scale terms $\hat{\lambda}^{\mathrm{S}}$ and $\hat{\lambda}^{\mathrm{R}}$ and does not affect other model parameters besides $\hat{\beta}$.

To compare the preference-space results more directly to the WTP-space model, we compute WTP *post-hoc* from the parameter estimates of the preference space model (i.e. $\hat{\beta}/\hat{\alpha}$). Figure 13 summarizes these results, which closely match those from our base case simulation in the WTP space (Figure 2). Thus the results from a preference space parameterization of our simulation are equivalent to those from a WTP space parameterization. We find interpretation cleaner in the WTP-space, which parametrically isolates WTP (ratio of utility of non-price attributes to utility of price) from the scale terms (ratio of utility of attributes to the scale of the error term), rather than jointly interpreting the estimates of $\hat{\alpha}$, $\hat{\beta}$, and $\hat{\lambda}$ in Figures 10, 11, and 12 in the preference space. The WTP space facilitates comparison of model parameter estimates across different models without the need for extra computations after model estimation, and we focus on WTP-space models exclusively in the main text.

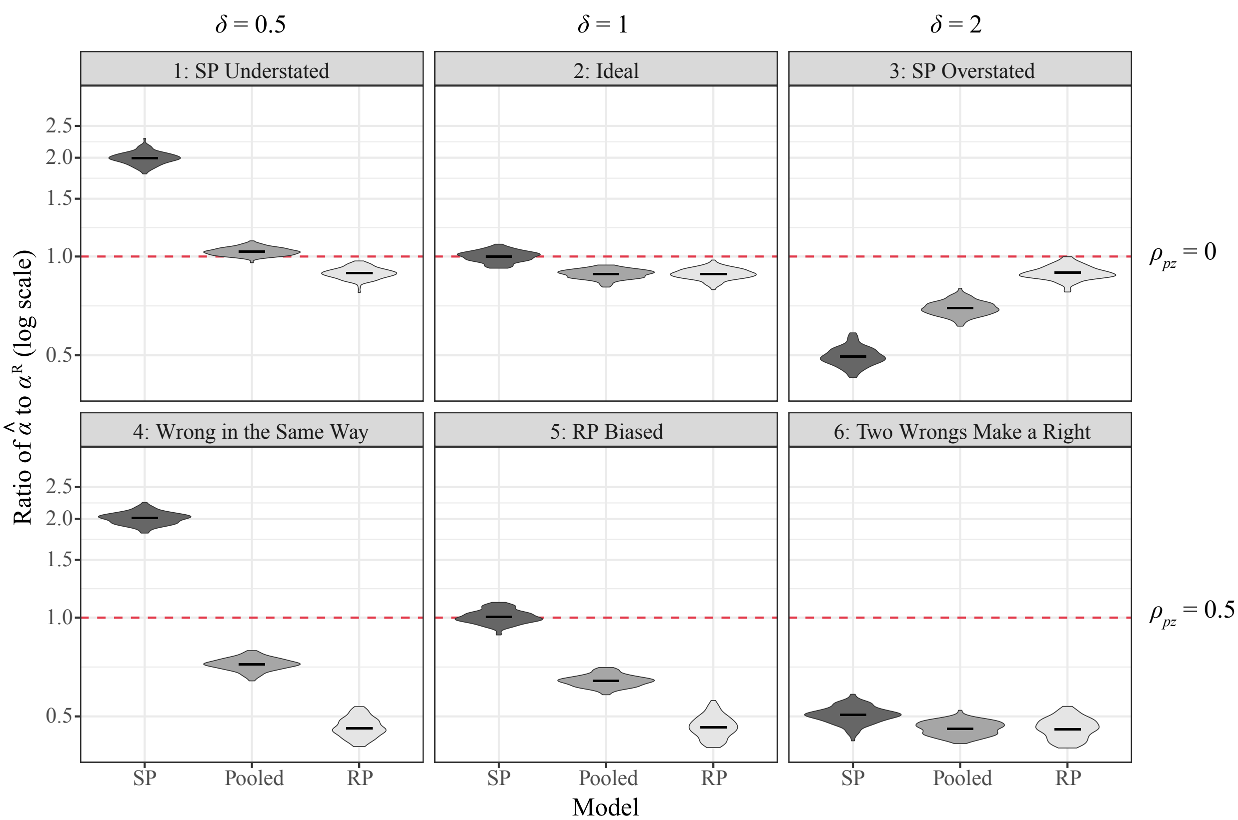


Figure 10: SP, RP, and pooled model results of the ratio of $\hat{\alpha}$ to α^{R} for each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. Sampling variation is visualized with a kernel density plot.

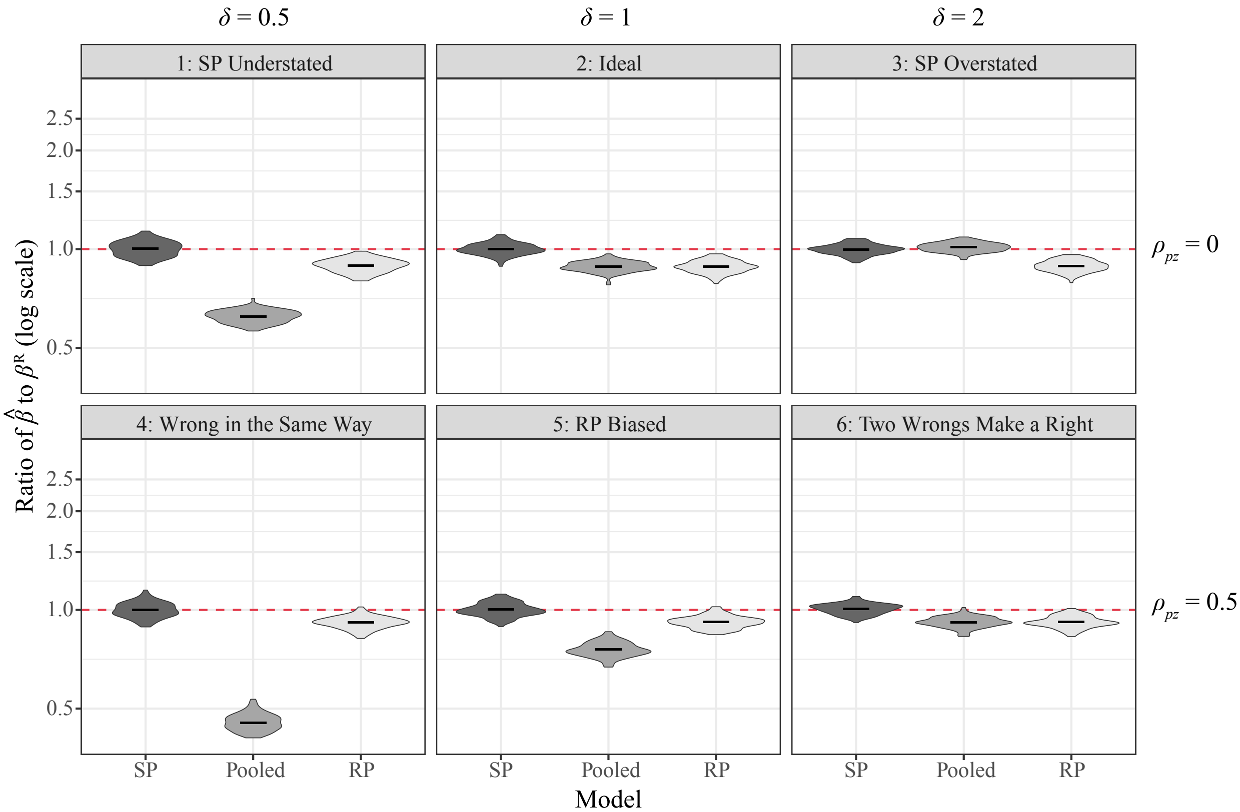


Figure 11: SP, RP, and pooled model results of the ratio of $\hat{\beta}$ to β^{R} for each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. Sampling variation is visualized with a kernel density plot.

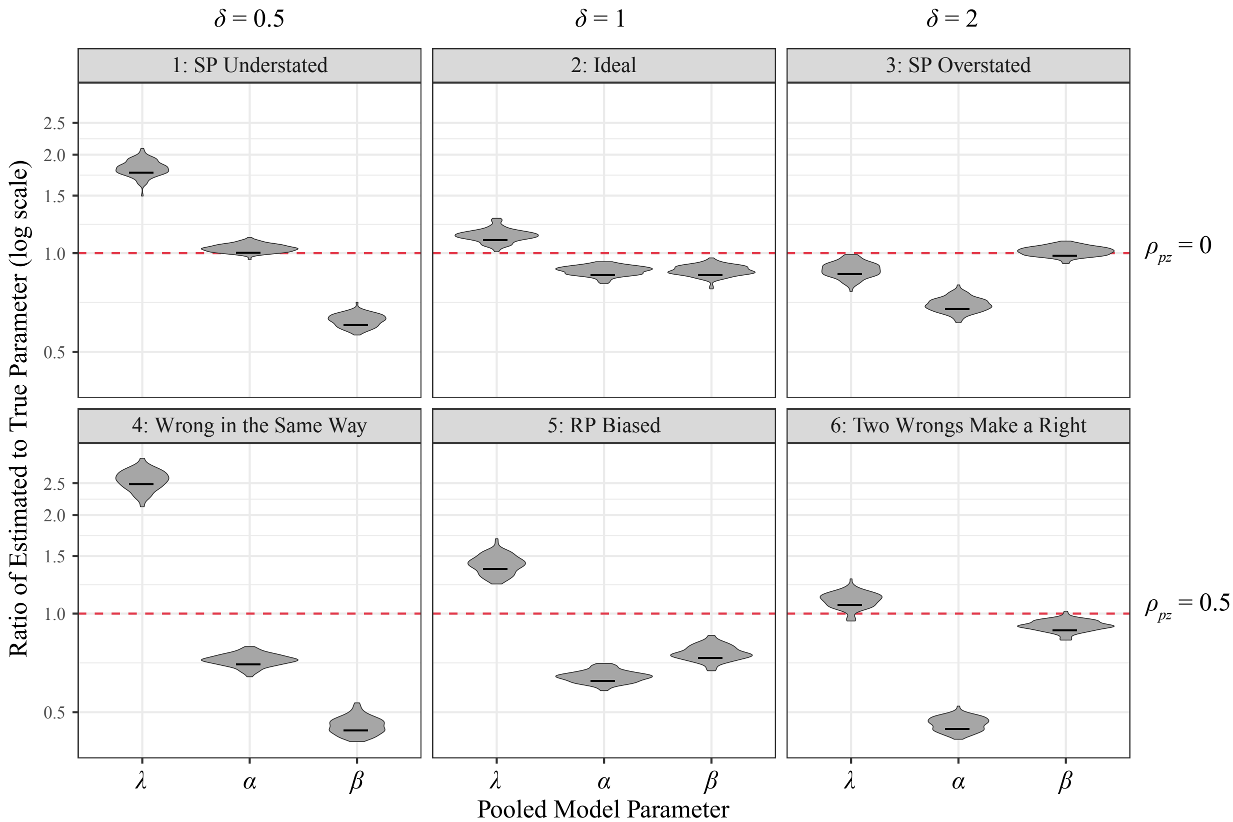


Figure 12: Ratio of estimated to true parameters for all pooled model parameters in each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. Sampling variation is visualized with a kernel density plot.

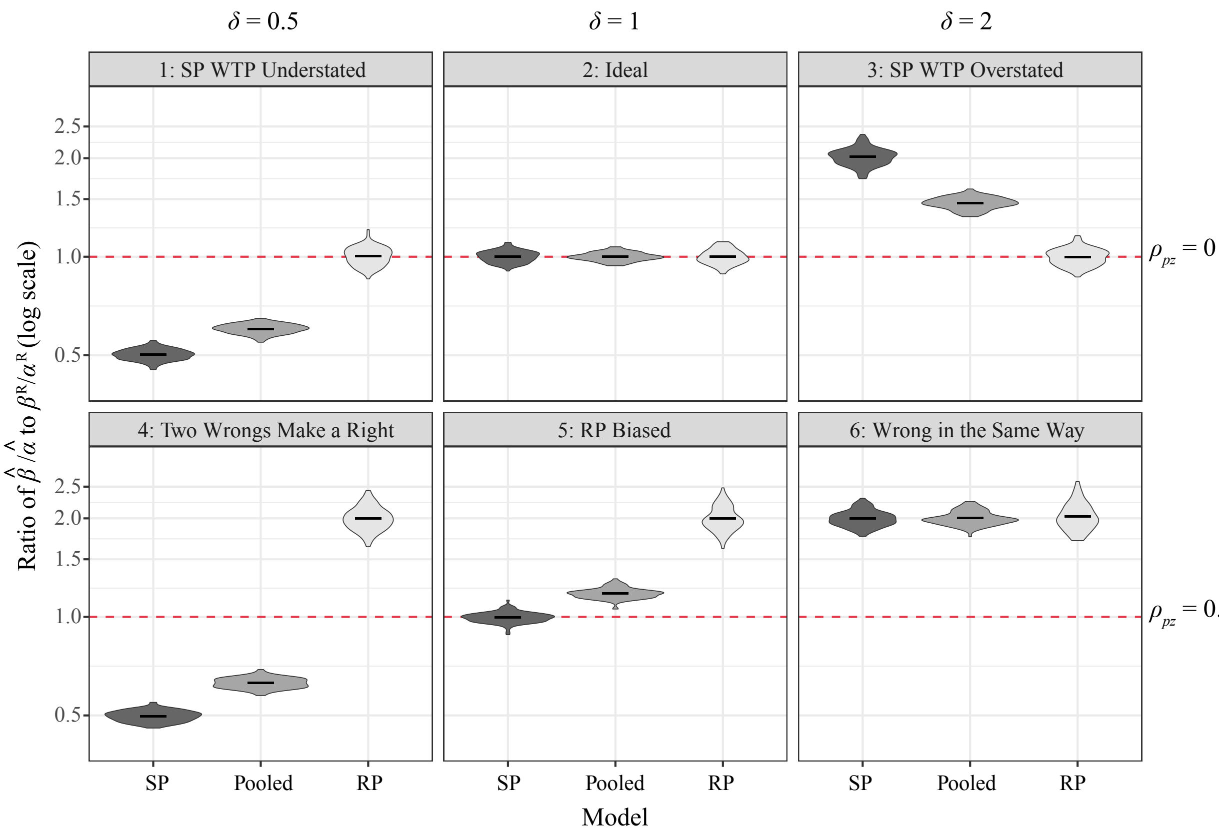
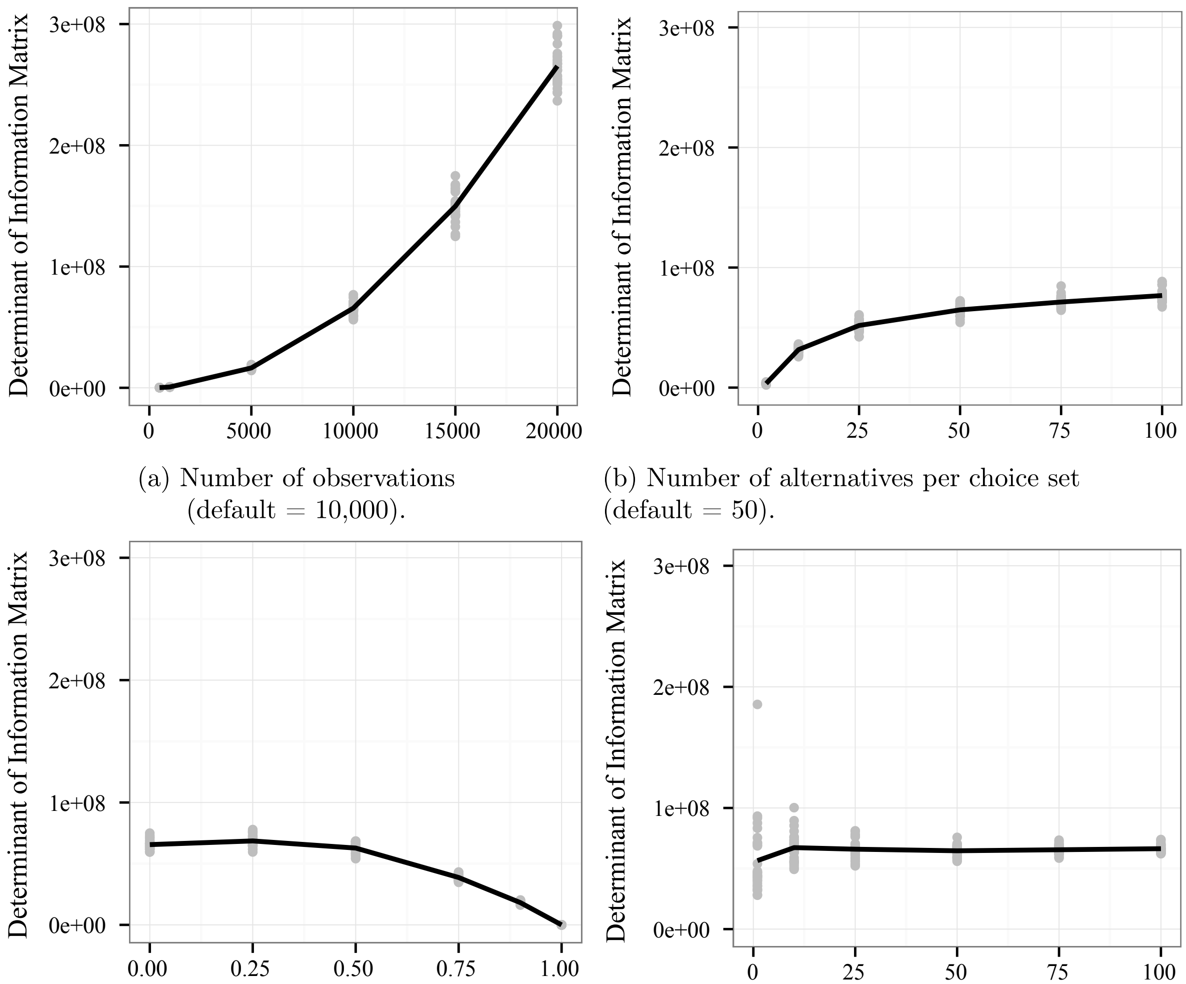


Figure 13: SP, RP, and pooled model results of the ratio of the computed WTP from the estimates of the preference space model ($\hat{\beta}/\hat{\alpha}$) to the computed WTP from the true RP context preference space parameters ($\beta^{\mathrm{R}}/\alpha^{\mathrm{R}}$) for each test case in Table 3. Each plot represents results from 1,000 simulated data sets using the base case parameters in Table 2. Sampling variation is visualized with a kernel density plot.

8.4 Data Set Characteristics and Fisher Information

A number of characteristics influence the amount of information a data set carries about the unknown model parameters. In particular, we examine the relationships between the Fisher Information and the number of observations, the correlation among observed attributes, the number of different choice sets, and the number of alternatives in the choice sets. To illustrate these relationships, we simulate sets of choice data for a simple two product case and then compute the determinant of the information matrix at the true parameters. By varying one characteristic while holding all others constant, we can visualize the relationships between these attributes and data set information, as shown in Figure 14. We use the determinant of the information matrix as an approximation for the overall total amount of information in a data set.

The amount of information is quadratically related to the number of choice observations (Figure 14a), making sample size a large determinant of the overall amount of information. Increasing the number of alternatives in a choice set (Figure 14b) has diminishing returns on information and follows a logarithmic relationship. As Figure 14c illustrates, the correlation between attributes in the data set is critical. While low correlations have a limited impact on information, highly correlated attributes can dramatically reduce the level of information. Finally, increasing the number of choice sets in a data set, which is the same as adding more variation among the attributes, does not necessarily affect the overall *amount* of information but rather the *variation* in the amount of information across different data sets. As Figure 14d shows, a low number of choice sets results in high variation in the amount of information



(a) Number of observations (default = 10,000).

(b) Number of alternatives per choice set (default = 50).

(c) Correlation between attributes (default = 0).

(d) Number of choice sets (default = 50).

Figure 14: Relationships between data set characteristics and Fisher Information.

(depending on the random draw of data), but as the number of choice sets increases the amount of variation among the attributes also increases which decreases the variation in the information. Taking all of these factors together illustrates how some data sets (such as aggregate market data with highly correlated attributes) can be relatively uninformative about attributes even with large sample sizes.

8.5 Sensitivity Analysis Figures

This section provides plots of the WTP parameter estimates for each sensitivity case in Table 2. In each plot, the light colors represent the base case and the dark colors represent the sensitivity case. Table 6 below shows the sensitivity ranges examined for each variable as well as the figures associated with each sensitivity case.

Table 6: Summary of sensitivity cases

Parameter	Base Case	Sensitivity Case	Sampling Variance	RP Data Information	SP Data Information	Information Balance	Figure Number
λ^{R}	1	0.1	Increases	Decreases	–	SP greater	15
		5	Decreases	Increases	–	RP greater	16
λ^{S}	1	0.1	Increases	–	Decreases	RP greater	17
		5	Decreases	–	Increases	SP greater	18
β^{R}	1	0.5	Increases	Decreases	–	SP greater	19
		2	Decreases	Increases	–	RP greater	20
ζ	1	-1	–	Decreases	–	SP greater	21
		1.5	–	Decreases	–	SP greater	22
ρ_{px}	0	0.5	–	Decreases	–	SP greater	23
N^{R}	1000	500	Increases	Decreases	–	SP greater	24
		5000	Decreases	Increases	–	RP greater	25
N^{S}	1500	500	Increases	–	Decreases	RP greater	26
		5000	Decreases	–	Increases	SP greater	27
A^{R}	15	3	Increases	Decreases	–	SP greater	28
		100	Decreases	Increases	–	RP greater	29
A^{S}	3	2	Increases	–	Decreases	RP greater	30
		10	Decreases	–	Increases	SP greater	31
C^{R}	1000	1	Increases	Decreases	–	SP greater	32
		1000	Decreases	Increases	–	RP greater	33

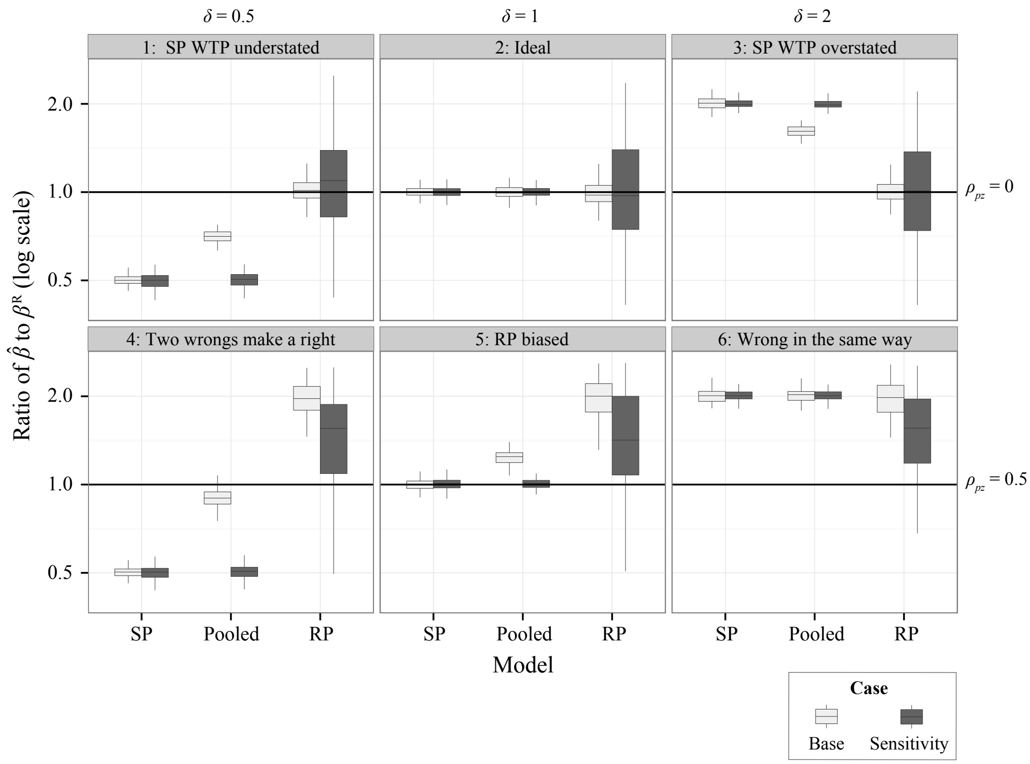


Figure 15: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\lambda^{\mathrm{R}} = 0.1$. Each box plot contains 100 simulated data sets.

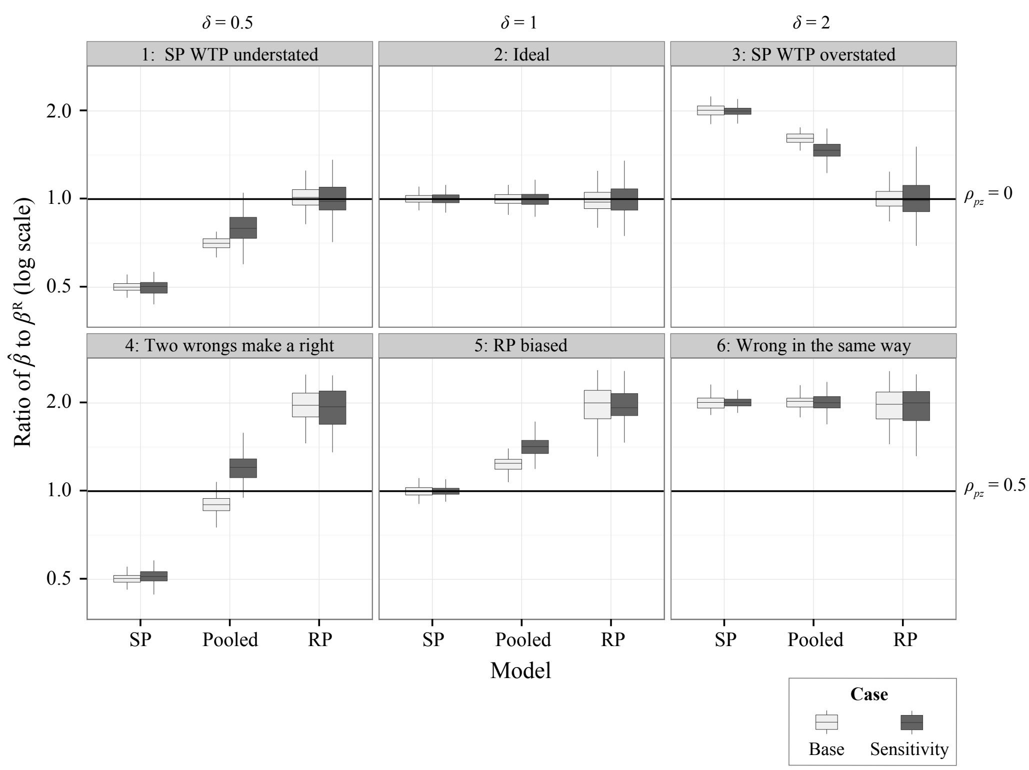


Figure 16: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\lambda^{\mathrm{R}} = 5$. Each box plot contains 100 simulated data sets.

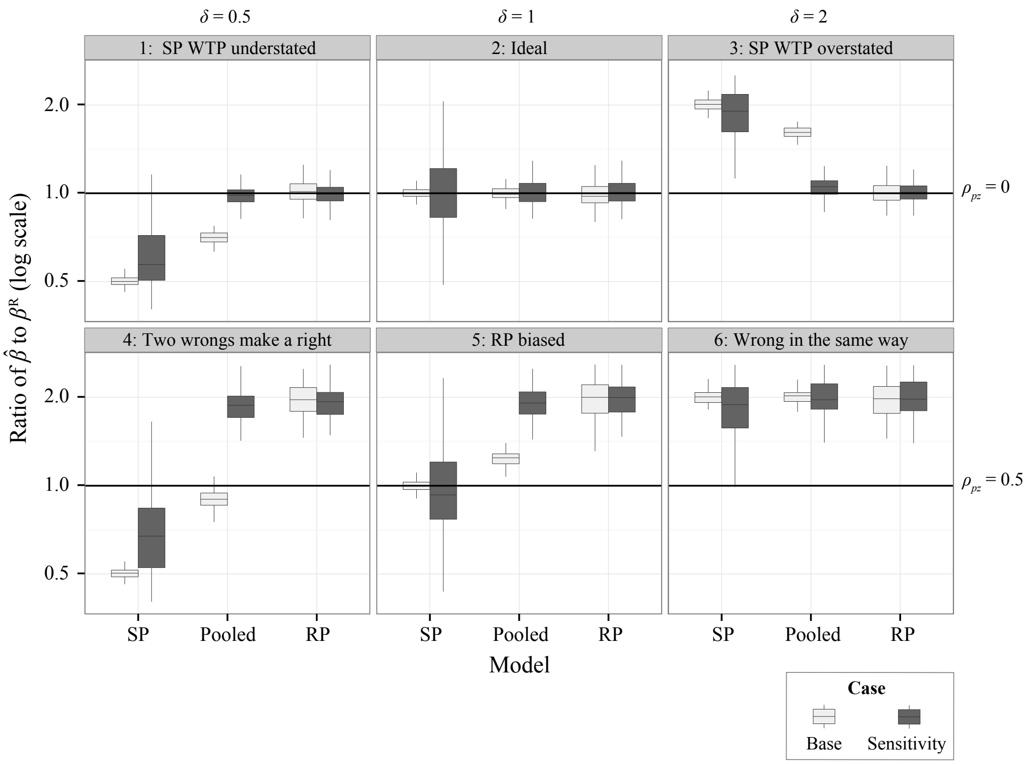


Figure 17: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\lambda^{\mathrm{S}} = 0.1$. Each box plot contains 100 simulated data sets.

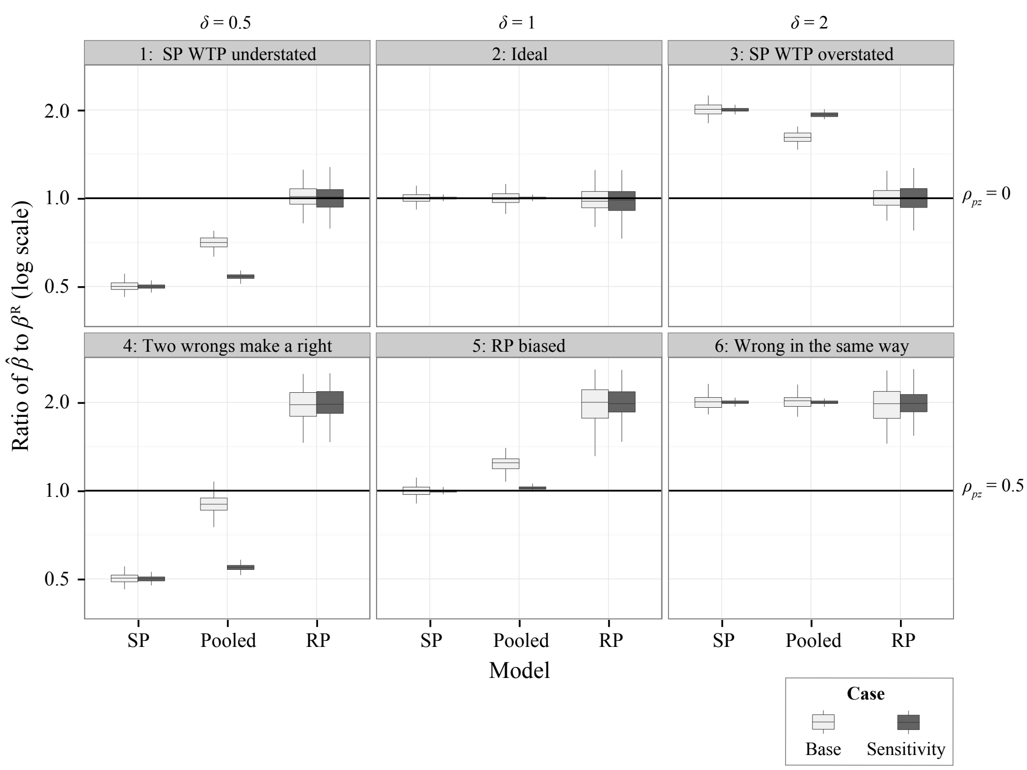


Figure 18: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\lambda^{\mathrm{S}} = 5$. Each box plot contains 100 simulated data sets.

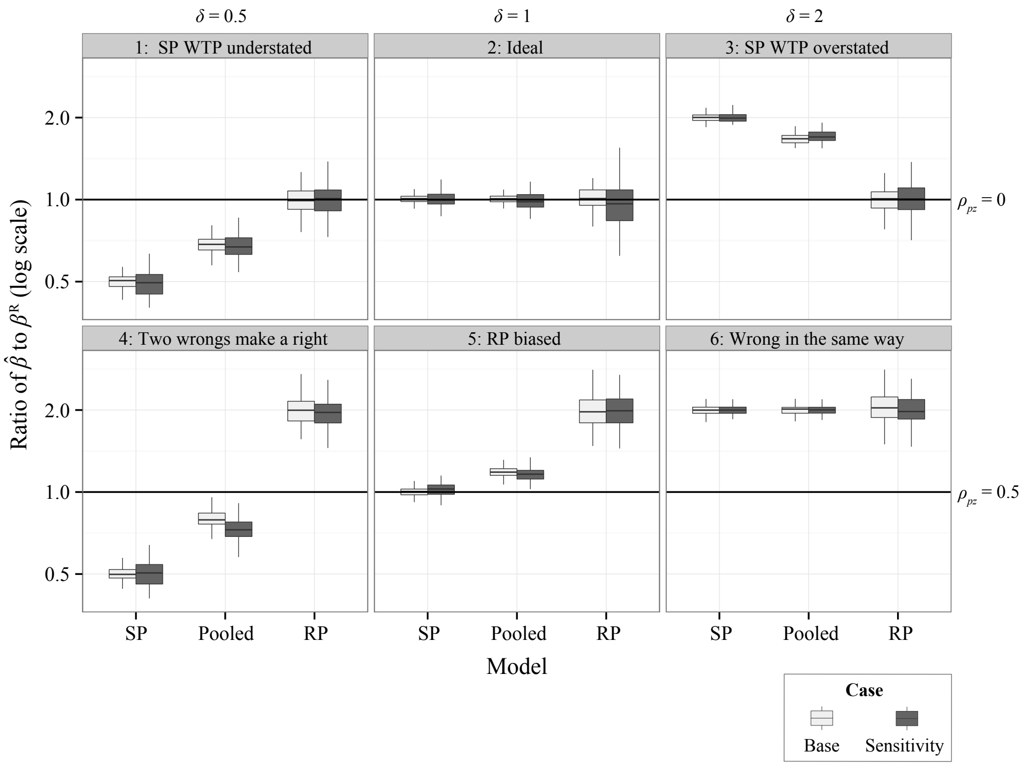


Figure 19: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\beta^{\mathrm{R}} = 0.5$. Each box plot contains 100 simulated data sets.

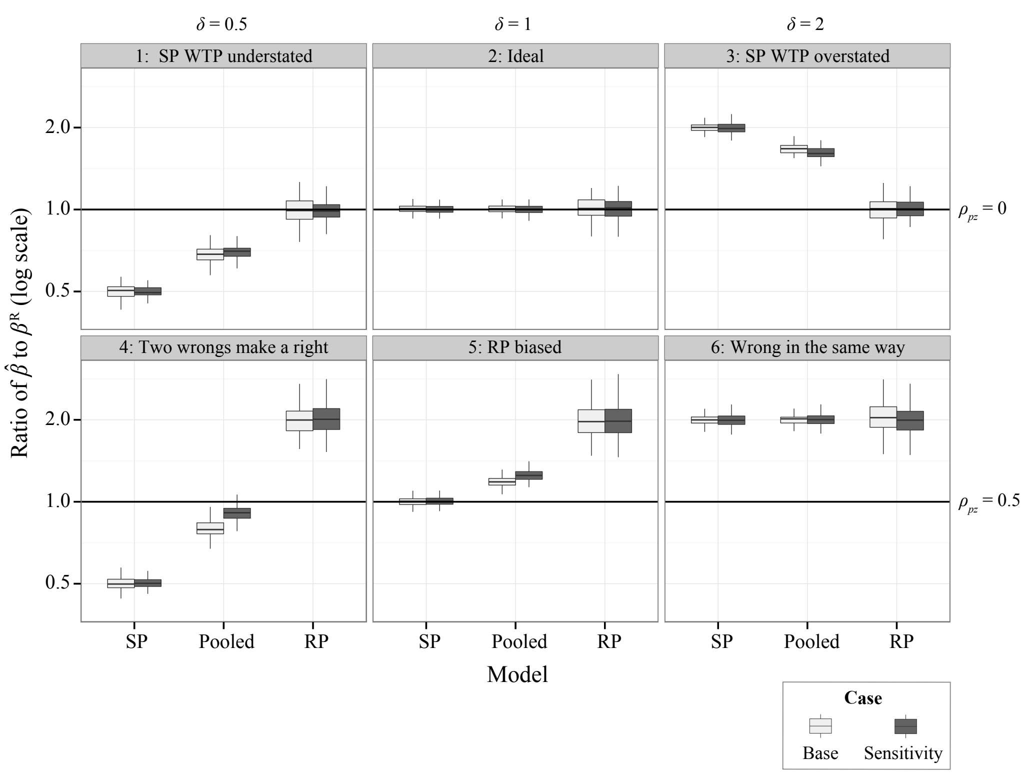


Figure 20: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\beta^{\mathrm{R}} = 2$. Each box plot contains 100 simulated data sets.

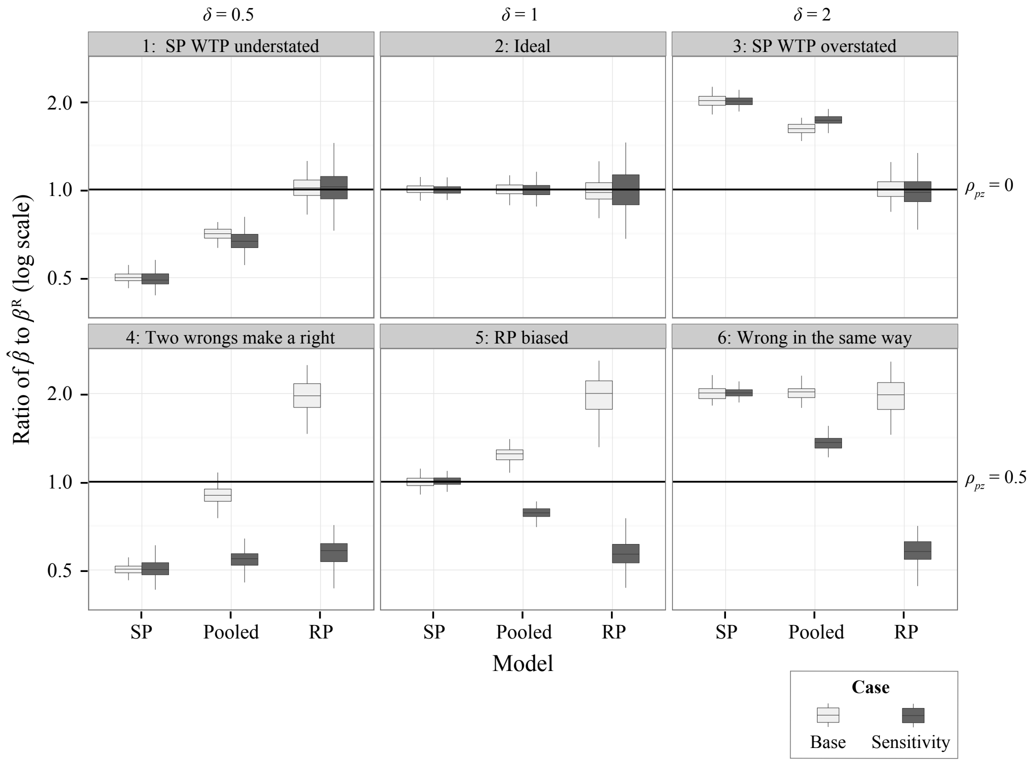


Figure 21: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\zeta = -1.5$. Each box plot contains 100 simulated data sets.

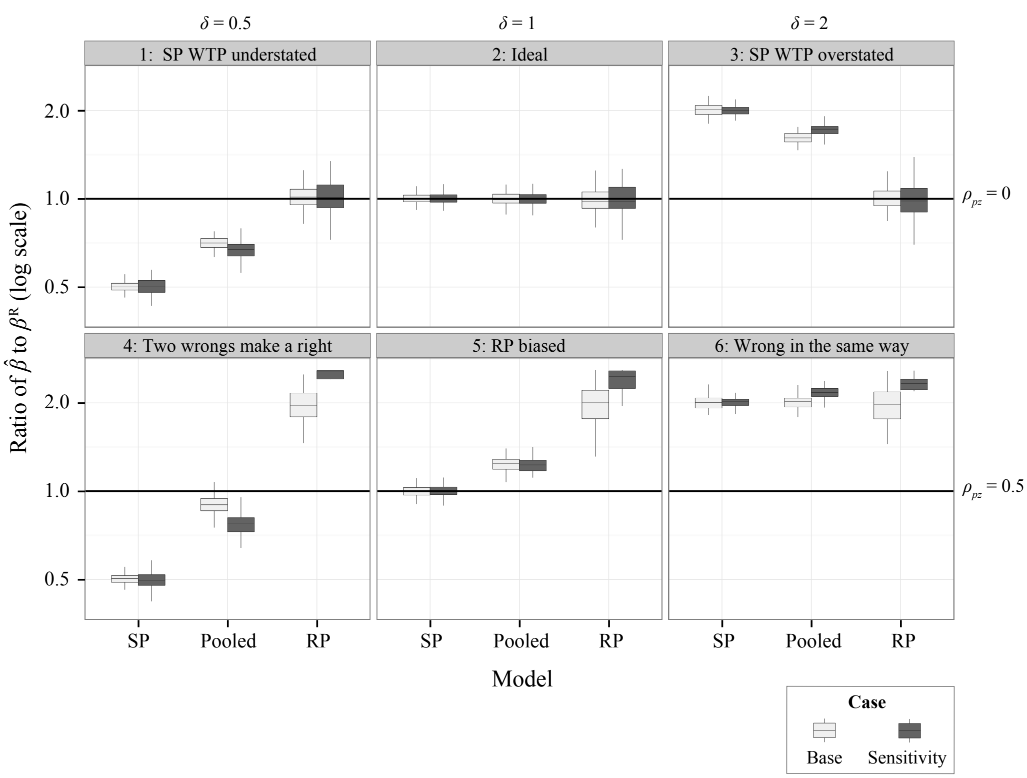


Figure 22: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\zeta = 1.5$. Each box plot contains 100 simulated data sets.

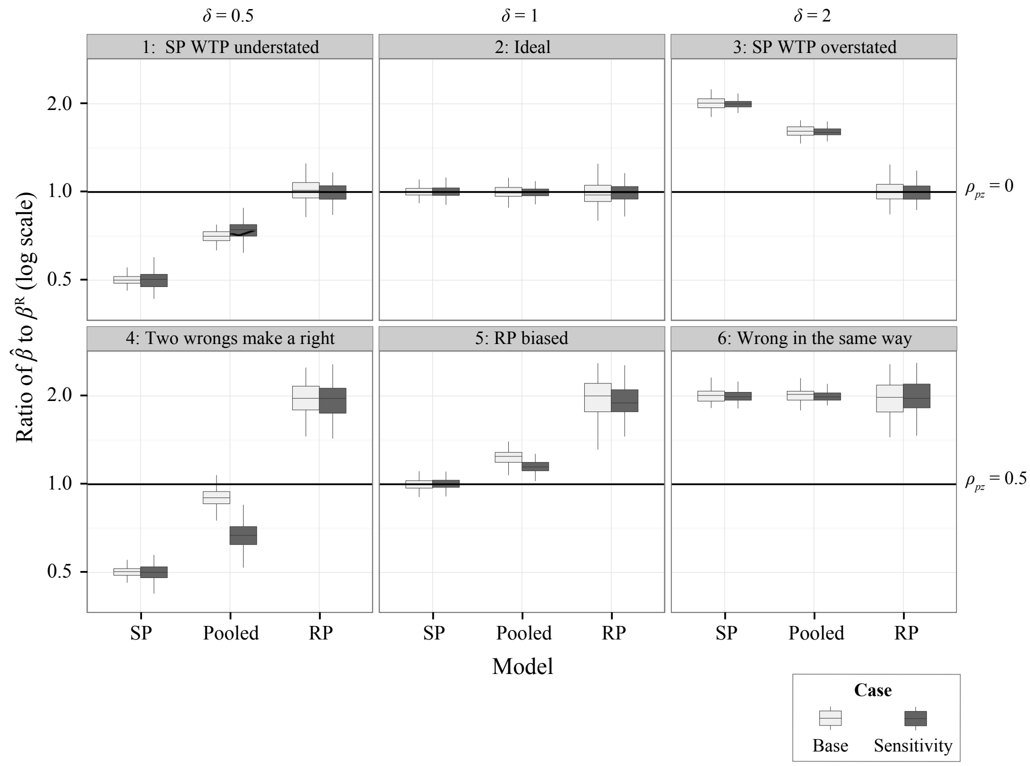


Figure 23: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $\rho_{px} = 0.5$. Each box plot contains 100 simulated data sets.

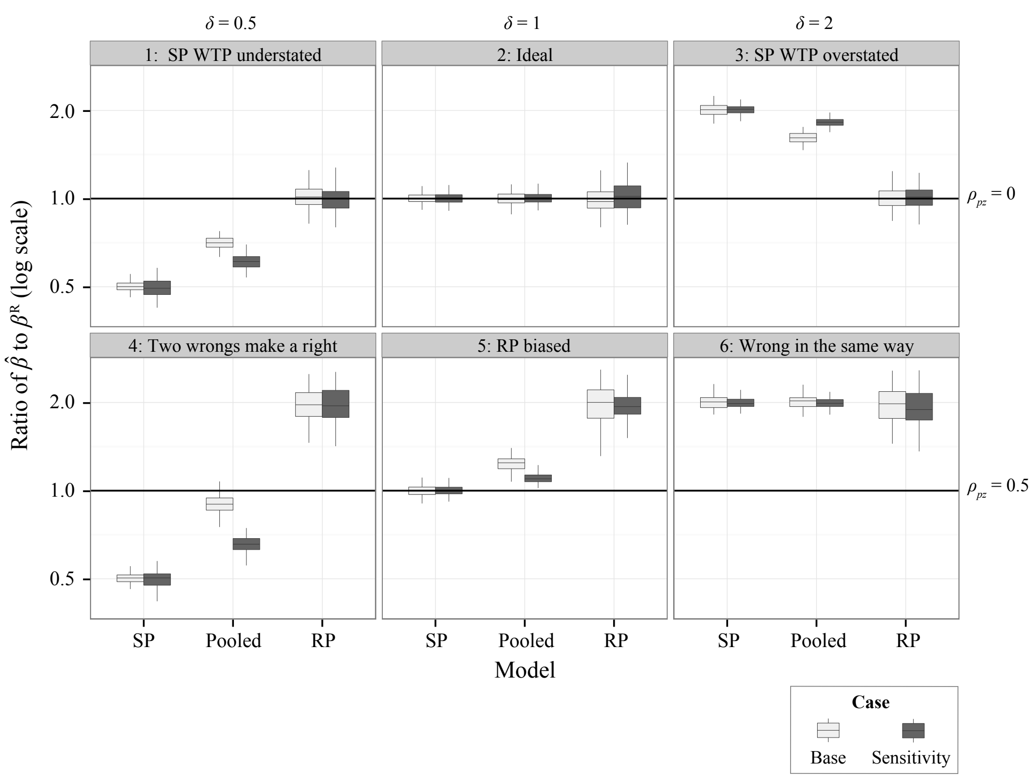


Figure 24: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $N^{\mathrm{R}} = 500$. Each box plot contains 100 simulated data sets.

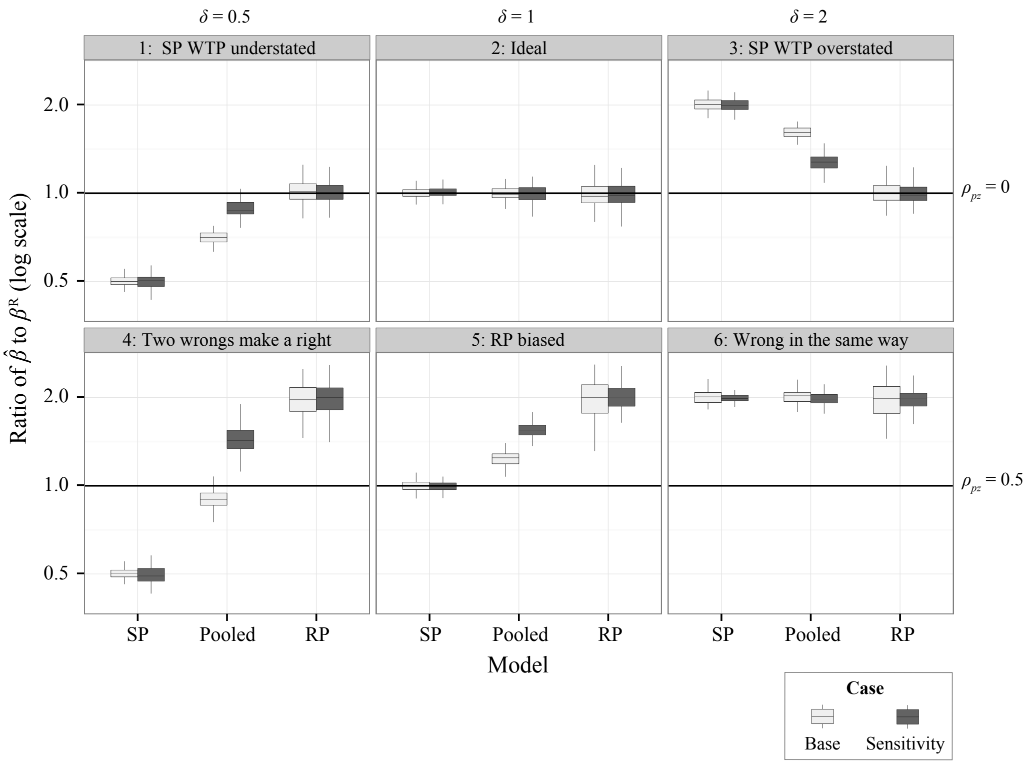


Figure 25: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $N^{\mathrm{R}} = 5000$. Each box plot contains 100 simulated data sets.

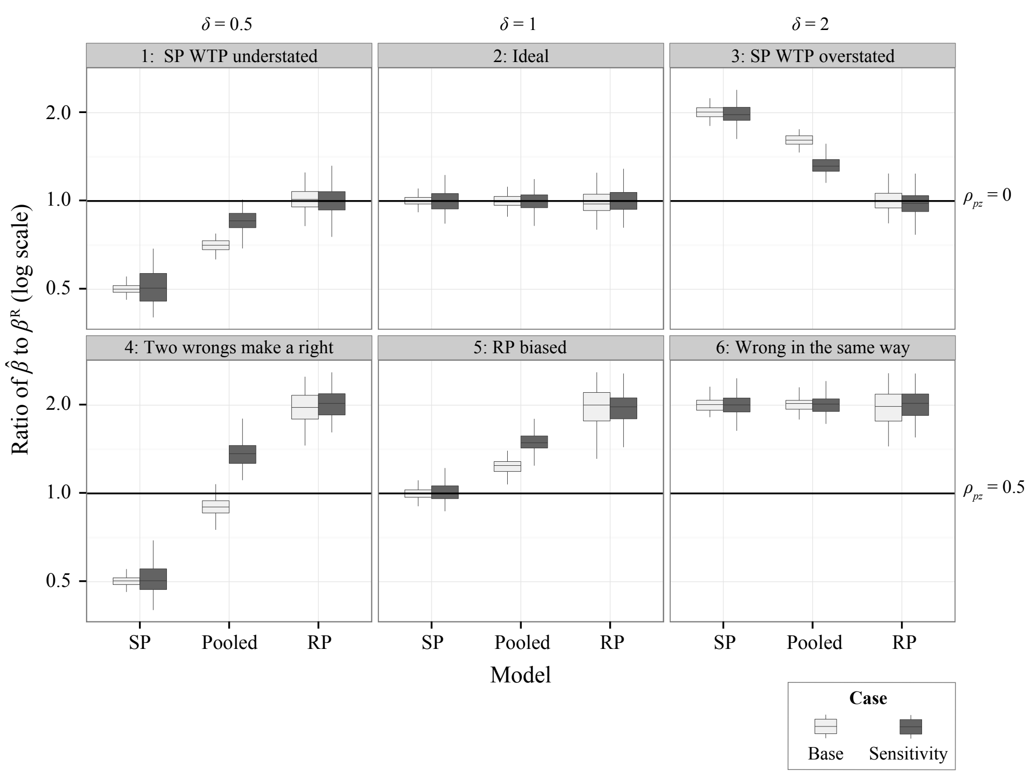


Figure 26: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $N^{\mathrm{S}} = 500$. Each box plot contains 100 simulated data sets.

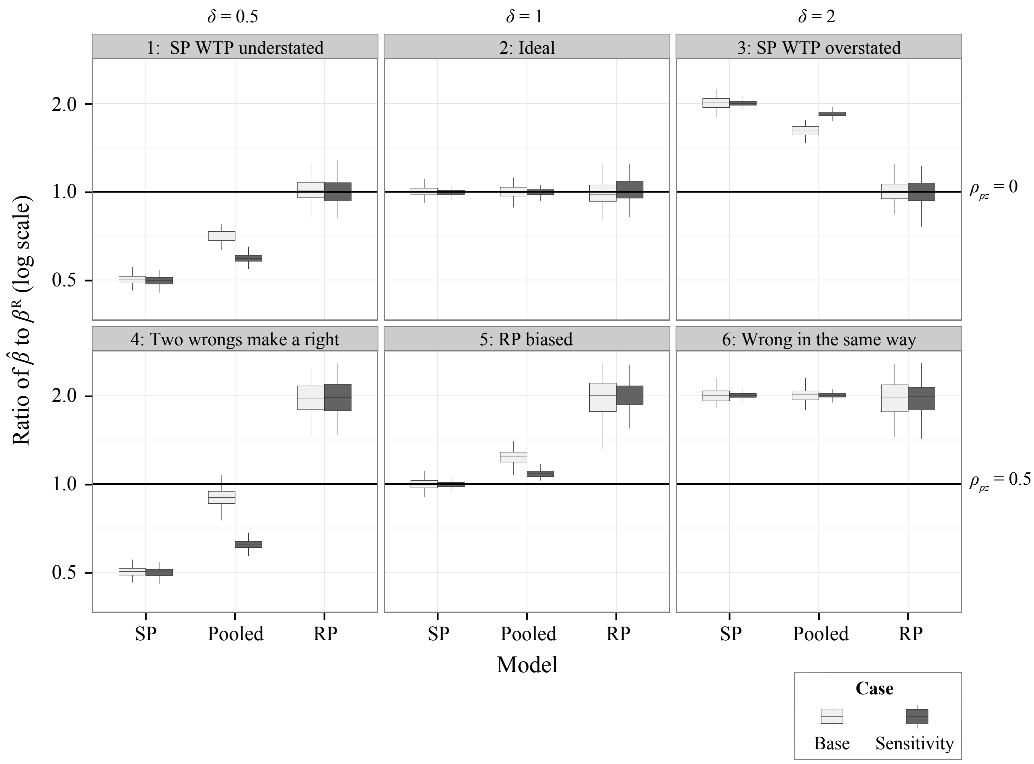


Figure 27: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $N^{\mathrm{S}} = 5000$. Each box plot contains 100 simulated data sets.

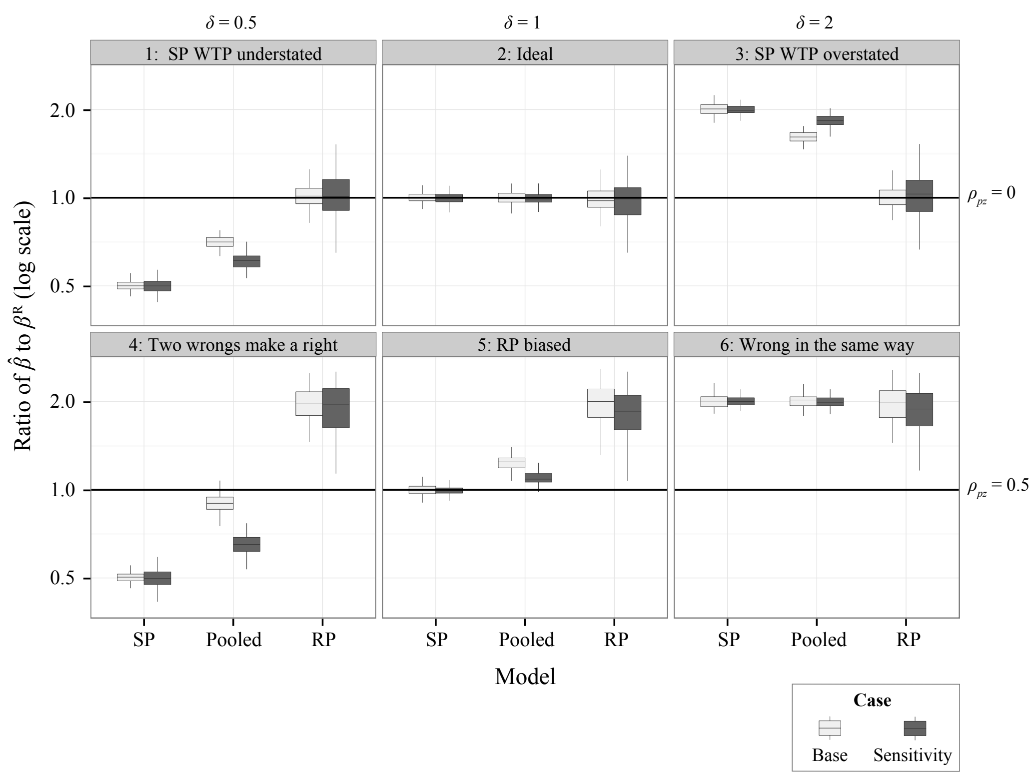


Figure 28: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $A^{\mathrm{R}} = 3$. Each box plot contains 100 simulated data sets.

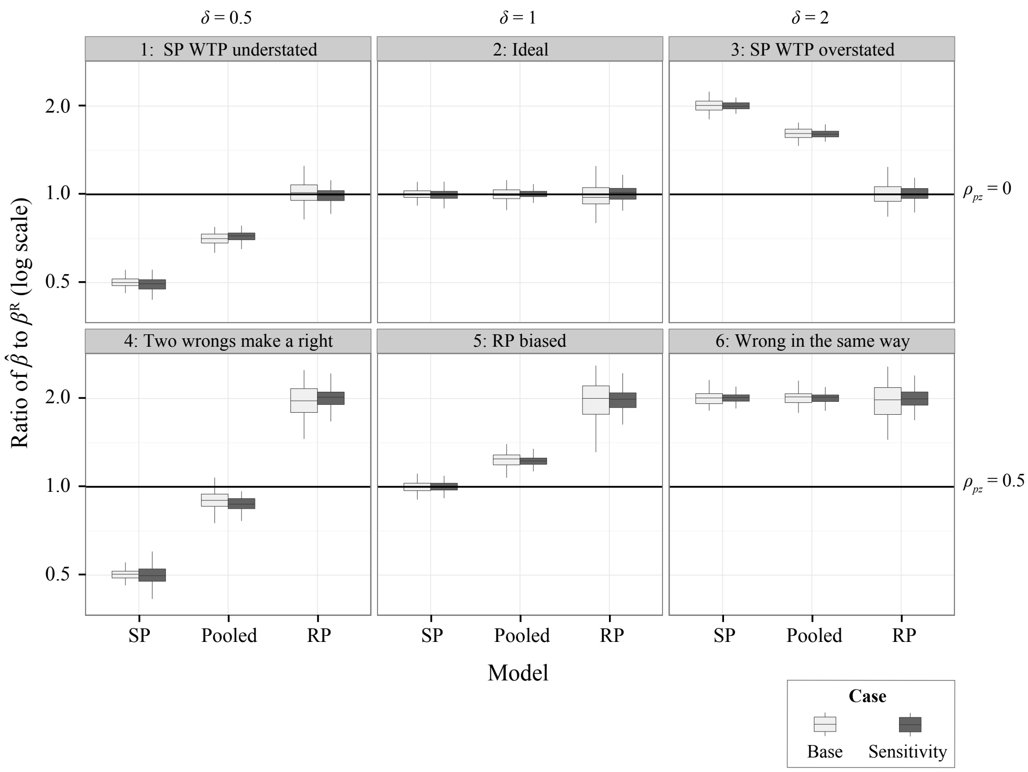


Figure 29: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $A^{\mathrm{R}} = 100$. Each box plot contains 100 simulated data sets.

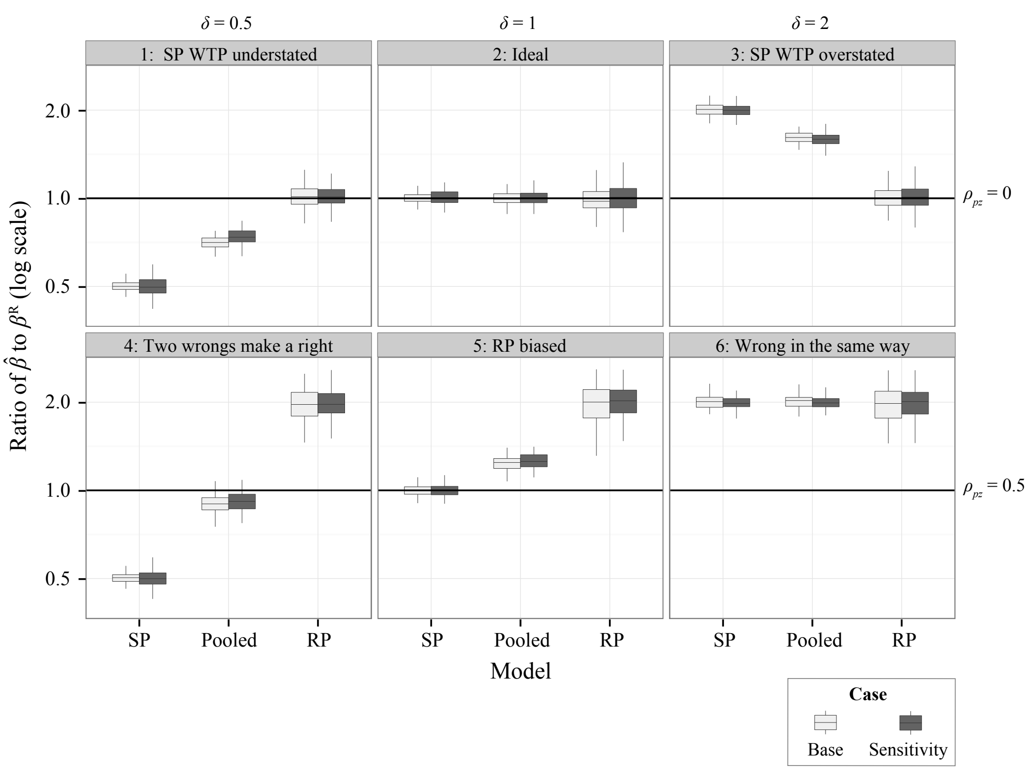


Figure 30: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $A^{\mathrm{S}} = 2$. Each box plot contains 100 simulated data sets.

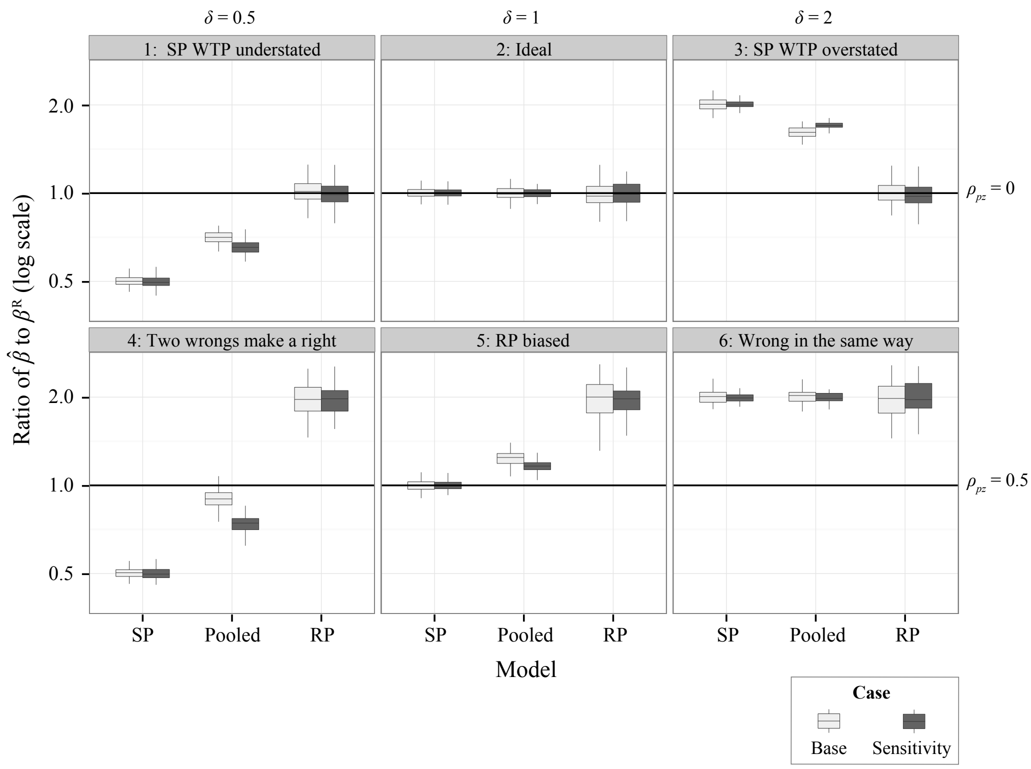


Figure 31: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $A^{\mathrm{S}} = 10$. Each box plot contains 100 simulated data sets.

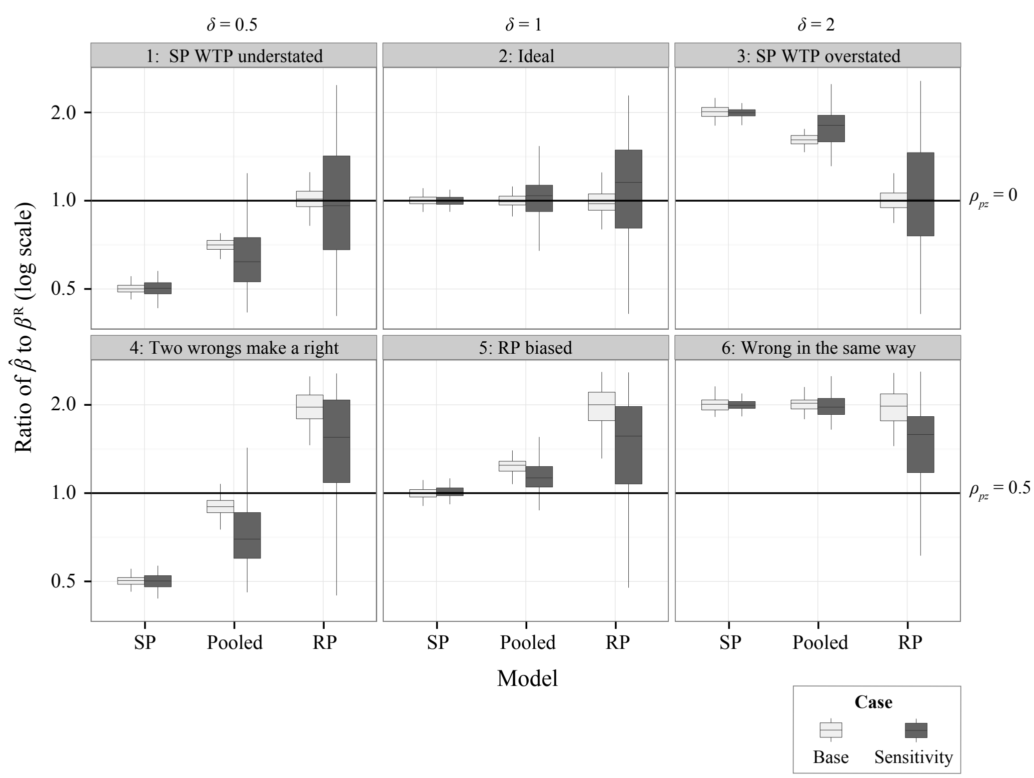


Figure 32: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $T^{\mathrm{R}} = 1$. Each box plot contains 100 simulated data sets.

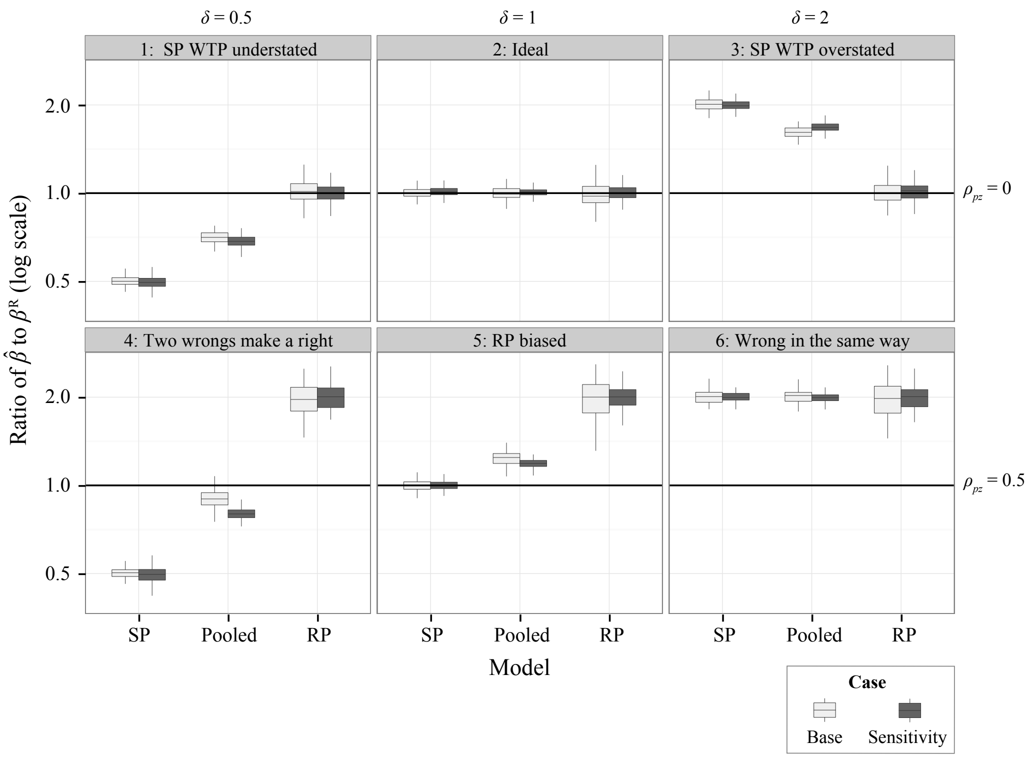


Figure 33: Ratio of $\hat{\beta}$ to β for the base case (light color) and sensitivity case (dark color) where $C^{\mathrm{R}} = 200$. Each box plot contains 100 simulated data sets.