![White Rose university consortium logo - Universities of Leeds, Sheffield & York]

eprints@whiterose.ac.uk
https://eprints.whiterose.ac.uk/

Stubbing out hypothetical bias: improving tobacco market predictions by combining stated and revealed preference data

John Buckell, Health Policy and Management, School of Public Health, Yale University

Stephane Hess, Choice Modelling Centre & Institute for Transport Studies, University of Leeds

Corresponding author: John Buckell, Yale University, 135 College Street, New Haven, CT, USA. john.buckell@yale.edu

Abstract

In health, stated preference data from discrete choice experiments (DCEs) are commonly used to estimate discrete choice models that are then used for forecasting behavioral change, often with the goal of informing policy decisions. Data from DCEs are potentially subject to hypothetical bias. In turn, forecasts may be biased, yielding substandard evidence for policymakers. Bias can enter both through the elasticities as well as through the model constants. Simple correction approaches exist (using revealed preference data) but are seemingly not widely used in health economics. We use DCE data from an experiment on smokers in the US. Real-world data are used to calibrate scale of utility (in two ways) and the alternative-specific constants (ASCs); several innovations for calibration are proposed. We find that embedding revealed preference data in the model makes a substantial difference to the forecasts; and that how models are calibrated also makes a substantial difference.

Highlights:

- We combine SP with multiple sources of RP data in choice models
- We study the impact of a range of calibrations on predictions
- Model calibration itself makes a substantial impact on predictions
- How model calibration is conducted makes a substantial impact on predictions

Key words: Stated preference; revealed preference; hypothetical bias; tobacco; discrete choice experiment; policy predictions

JEL codes: C35; I12; I18

1. Introduction

Discrete choice models are used extensively in health economics, with a major focus on the use of data from experiments in surveys, variably referred to as stated preference (SP), stated choice (SC) or discrete choice experiments (DCE), in contrast to the Revealed Preference (RP) data sources for which choice modelling was initially developed (see Louviere and Lancsar, 2009; de Bekker-Grob et al., 2012; Clark et al., 2014; Soekhai et al., 2018, for backgrounds in the use of choice modelling in health). The models estimated on such data provide insights into the relative importance of different product or service characteristics in determining the choice of an individual decision-maker. After estimation, it is straightforward to use these choice models in forecasting, i.e. predicting the demand for services and/or the changes in demand as a result of changes in the population of decision-makers and/or the characteristics of the products/services.

Forecasts of this type can be extremely valuable to policymakers. While generating forecasts from models estimated on SP data is a straightforward process, the reliability of SP-based forecasts is not guaranteed. The main objection to SP data is hypothetical bias (Hausman, 2012). That is, what people say they will do and what people actually do can be very different. This can, in some cases, lead to specious forecasts lacking external validity. Policymaking based on these estimates is therefore at risk to this extent. As such, this is one of the most important, if not the most important, issues concerning forecasts derived from choice models estimated on hypothetical data.

The impact of hypothetical bias in SP data has been studied widely in contingent valuation (CV) studies, where it is thought that individuals typically overstate willingness to pay (List and Gallett, 2001; Little and Berrens, 2004; Murphy et al., 2005; Whynes et al., 2005; Donofouet et al., 2013). In the DCE setting, there are fewer examples, owing to the relative difficulty and cost of conducting such studies (Fifer et al., 2014). In terms of attributes, the extant evidence suggests that hypothetical bias plays a role, but its effect on the direction of preference estimates varies (Fifer et al., 2014; Beck et al., 2016; Rakotonarivo et al., 2016). In the health setting, one study found that WTP from DCEs was higher than that of CV (Ryan and Watson, 2009). Elsewhere, Ozdemir et al. (2009) find that the use of "cheap talk" (i.e. instructing respondents of the importance of their responses in an attempt to elicit truthful responses) reduces estimates of WTP (versus no use of cheap talk); interpreting this as having reduced hypothetical bias. This is broadly

consistent with findings in the wider DCE literature (Beck et al., 2016). However, it is not clear that cheap talk can eradicate hypothetical bias entirely; indeed, methods of this ilk are not generalizable, proven remedies for hypothetical bias (Harrison, 2014, Wuepper et al., 2018).

The type of bias discussed above related to the possibility that the way in which respondents react differently to attributes in a hypothetical setting compared to a real-world setting varies across attributes, leading to differences in WTP. In addition, there is the possibility of respondents reacting more or less strongly overall to the stimuli they are faced with in a hypothetical setting. This is in line with earlier work showing that, in terms of product preferences, respondents in DCEs are thought to overstate their propensity to purchase products, so market shares may be biased (Train, 2009). A literature review of predictions versus choices in health DCEs suggests that there is reasonable concordance between predicted and actual choices (Quaife et al., 2018). The remaining error, however, suggests that hypothetical bias is at play. Thus, the evidence suggests that hypothetical bias in SP-based DCEs impacts on estimates of both the attribute and product preferences.

Choice modelling is multidisciplinary, and it is useful to reflect on the views held in other disciplines. One of the more established fields of choice modelling research is transportation, where the general perception is that data from hypothetical surveys are useful for understanding relative sensitivities of decision makers, e.g. how important is price as a characteristic compared to waiting time, but is potentially subject to severe bias in terms of the absolute sensitivities (Train, 2009; Hess and Daly, 2014). While commonly done in consulting work, very few academic studies in transport would use hypothetical data for forecasting, especially without additional corrections such as discussed in the present paper. In particular, the general perception is that, when answering hypothetical choice tasks (as in SP data), decision makers face these choices in a highly isolated setting (in contrast with their real-world decision-making, as in RP data, where numerous outside factors are at play) and may thus overstate their reaction to changes in those variables included in a survey (Harrison, 2014; Hensher et al., 2015). If this is the case, then parameters, and derived elasticities from those parameters, in models estimated on hypothetical data are likely to be overstated, leading to biased forecasts, and in particular an overstated impact of interventions. It is the scale of utility[1] which reflects the degree of randomness in behavior in the choice

---

[1] the scale is inversely proportional to the variance of the error (i.e. the unobserved factors in utility; see Train, 2009, chapter 2)

model – if the elasticities in SP data are biased upwards, then the scale in SP data will be higher than the RP scale.

Another possible source of bias arises when the market shares for the different options in a dataset used for modelling are different from overall real-world shares. This issue can arise with RP data, as a result of the sample of decision makers not being representative of the overall population, but is especially prevalent in hypothetical data. Indeed, not only does the same possibility of a non-representative sample arise (which may be intentional if targeting specific population subsets), but the way in which the alternatives are described in the survey may influence the shares obtained in the survey. It is the product constants in the choice model (commonly termed alternative-specific constants (ASCs)) through which this form of bias arises, driven directly by the market shares in the estimation data.

Both types of bias can potentially be addressed with a mix of different RP-based methods, but these have received relatively little exposure in health economics. This paper gives an overview of these corrections, scale calibration and ASC calibration, and illustrates their impacts in the case of a typical stated choice survey in health. We provide several innovations that are useful to researchers in the field. First, we combine both scale and ASC calibration, and compare results to uncalibrated models and models that calibrate either feature. Second, we compare two approaches to calibrating the scale of utility – one is using joint SP-RP modeling, and another is scale calibration based on elasticity measures in the literature. Third, we propose a novel, simple technique for calibrating model constants, that we term *partial calibration*, which allows for ambiguity in the interpretation of the outside good[2] (also referred to as the "opt-out" option). We show not only that these methods can have a profound effect on forecasts, but also that how these methods are applied can have a significant impact, too. This is important for those who are making predictions to inform policymaking.

We use a choice experiment on US adult smokers and recent quitters (Buckell et al., 2018). The application to tobacco enables us to use several sources of RP data: product use and purchasing data from individuals that took the experiment, RP data from the literature (Pesko et al., 2018) and Population Assessment of Tobacco and Health (PATH) – a large, nationally-collected tobacco use data set (Hyland et al., 2017). Use of these data allows for the alternative approaches to calibration, insights into the mechanisms of calibration, and an ensuing discussion as to which is more appropriate for the task at hand. Further,

---

[2] The partial calibration is new to health economics. To the best of our knowledge, it is new to the field of choice modelling, too.

tobacco regulation is an area in which policy predictions are of significant value and have been made in several recent studies (Kenkel et al., 2017; Buckell et al., 2018; Marti et al., 2019). Therefore, understanding the precision of forecasts is critical for generating high quality evidence for policymaking. In addition, this example highlights the importance of calibration, as follows. In principle, hypothetical bias should not pervade tobacco DCEs in the same way as other health applications. That is, in this setting, the options are realistic, the products are familiar to respondents and the products are well-described by their attributes; all of which are conducive to the DCE functioning as intended (McFadden, 2014). Thus, the impact of RP calibration here should be smaller than for other applications. We show that, even in this setting, model calibration makes a significant impact on forecasts.

2. Methods

Correcting the scale

Revealed preference (RP) data do not suffer from hypothetical bias. Thus, if available, incorporating RP data in choice models can abate hypothetical bias in model estimates and the derived metrics such as forecasts (Lancsar and Louviere, 2008; Lancsar and Burge, 2014). Such joint SP-RP estimation allows the researcher to match the elasticities in the choice model directly to real-world behaviors and still use the information from the variation in the experimental data. This draws simultaneously on the respective strengths of the two data sources; RP provides the scale while SP provides the relative sensitivities to different attributes. See Hensher et al. (2015), chapter 19, for an overview of this topic.

If compatible data are available from both an RP and an SP source (sharing at least one attribute), then joint estimation on the two data sources can be performed relatively easily by specifying the utilities in the joint model as follows:

$$U_{njt} = \left(x_{RP,nt} + \mu_{SP}x_{SP,nt}\right)\left(\delta_j + \beta'x_{njt}\right) + \varepsilon_{njt} \tag{1}$$

where U_{njt} is the utility for decision-maker *n* for alternative *j* in choice situation *t*. With the above notation, $x_{RP,nt}$ = 1 if choice situation *t* for decision-maker *n* is an RP observation (and zero otherwise) while $x_{SP,nt} = 1$ if choice situation *t* is an SP observation (and zero otherwise). δ_j is an alternative specific constant (ASC) for alternative *j*. x_{njt} are products' attributes, and the vector β contains the estimated

marginal utility parameters (with possible socio-demographic effects). ε_{njt} is the typical extreme value error term. The crucial term in Equation (1) is μ_{SP}, which is a scale parameter for the SP data. If the elasticities in the SP data are higher than in the RP data, μ_{SP} will be larger than 1, otherwise it will be smaller. As the scale for RP data is fixed at 1, the estimated model parameters (ASCs and marginal utility parameters) are of the RP scale, meaning that they reflect real-world (rather than hypothetical data) elasticities. The separate identification of μ_{SP} is made possible by the use of joint parameters in the utility functions for the two data sources (Ben-Akiva and Morikawa, 1990; Hensher et al., 1998; Bradley and Daly, 1994). In the joint model, SP and RP data are used together, and μ_{SP} is estimated with β and δj; see Train (2009) for details. An alternative approach is to use the so-called nested logit trick (Hensher and Bradley, 1993). However, this is not actually needed in software that allows for non-linear utility functions, where we can then simply use $\mu_{SP} \cdot \beta \cdot x$ in the SP utility functions. Other parameters are estimated from their own respective data sources: RP parameters are estimated from variation in RP data, and SP parameters are estimates from variation in SP data. Of course, if only either RP or SP data is available, then Equation 1 simplifies accordingly, and we do not estimate the additional μ_{SP} term[3].

We note that the estimation of differences in scale has been used for other purposes in choice models beyond SP-RP calibration. Louviere and Swait (1993) estimate scale factors across SP data sets; Bradley and Daly (1994) account for respondent fatigue through scale heterogeneity; and Hess et al., (2017) use scaling to align responses across different arms of their experimental design. It is also possible to capture deterministic heterogeneity in the scale of utility according to observed individual characteristics; see Vass et al. (2017) and Wright et al. (2018) for health-based applications.

In many cases, compatible RP data will not be available to analysts to enable joint SP-RP estimation. However, even in those cases, analysts will often have access to an elasticity from past work for at least one of the attributes from the model. Let us assume an analyst has a target elasticity of e_k^* for attribute k retrieved from prior literature. Using the estimated model parameters, the analyst is able to calculate an elasticity from the model (either analytically or through sample enumeration; see Hensher et al., 2015), say e_k^0. If $\left|e_k^0\right| \geq \left|e_k^*\right|$, then the scale in the model needs to be reduced (with the opposite applying if $\left|e_k^0\right| \leq \left|e_k^*\right|$). This can be done relatively easily after estimation. With the utility functions from the estimated models given by

---

[3] Then the model simplifies to the classic utility function used widely in health economics and beyond:
$U_{njt} = \delta_j + \beta'x_{njt} + \varepsilon_{njt}$

$$U_{njt} = \delta_j + \beta'x_{njt} + \varepsilon_{njt} \tag{2}$$

we can equivalently say:

$$U_{njt} = \mu_0(\delta_j + \beta'x_{njt}) + \varepsilon_{njt} \tag{3}$$

Where $\mu_0 = 1$, i.e. ensuring that U_{njt} is at the scale estimated from the data, and gives us the elasticity e_k^0. We can then set $\mu_1 = \mu_0 \frac{e_k^*}{e_k^0}$, and calculate the elasticities from the rescaled model, i.e. one in which μ_0 in Equation (3) is replaced by μ_1. This yields e_k^1 which can then again be compared to e_k^*. If $\left|e_k^1\right|$ is close enough to $\left|e_k^*\right|$, where the level of precision depends on the analyst's criteria, we use μ_1 in model application. In practice, several iterations of the correction approach may be required. This would imply calculating $\mu_s = \mu_{s-1} \frac{e_k^*}{e_k^{s-1}}$ in iteration s of the algorithm. In each iteration, we thus use the scale and elasticity from the previous iteration (just like we used the base scale and elasticity in the first iteration) to calculate the new scale parameter.

In health, calibration to RP scale has been limited, in large part due to the lack of RP data available in health markets (Lancsar and Swait, 2014; Lancsar and Burge, 2014). Some studies have used joint SP-RP models for calibration (Mark and Swait, 2004; Kestyernich et al., 2013; Kenkel et al., 2017), but we are not aware of published studies using the elasticity approach, in either health or elsewhere, though we have seen it used in unpublished consulting work.

Correcting the market shares

In a random utility model, alternative specific constants (ASC) are used to capture the mean effect of any factors not explained through the specification of the utility function (i.e. the impact of observed attributes of the alternatives and the decision makers). These ASCs ensure that the model perfectly recovers the market shares observed in the estimation data, at the sample level.

If the market shares in the data are not representative of real-world shares or shares in the specific application setting that an analyst is interested in, then the ASCs can be recalibrated to match the target market shares in model application, rather than the market shares in the data, as shown in Train (2009). Using the cigarette alternative as our example, we would use the estimated ASC for cigarette (δ_{cig}^0) and

all other model parameters to calculate the market share for cigarette that the uncalibrated model predicts on the estimate data. This market share, $\widehat{CS}^0_{cig}$, is then compared to the target (typically RP) market share of cigarettes, say MS^{RP}_{cig}. If the model underpredicts the real-world market share for an alternative, the constant for this alternative needs to be increased, with the opposite applying if it overpredicts the market share. Specifically, in the first iteration of this calibration, we would use:

$$\delta^1_{cig} = \hat{\delta}^0_{cig} + \ln\left(\frac{MS^{RP}_{cig}}{\widehat{CS}^0_{cig}}\right) \tag{4}$$

Where δ^1_{cig} is the recalibrated ASC for cigarette. As can be seen from the above, if model market shares predicted by the model match the target market shares, $\ln\left(\frac{MS^{RP}_{cig}}{\widehat{CS}^0_{cig}}\right)$ will be zero and no adjustment is made. If the target market share for a given alternative is higher than that predicted by the uncalibrated model, a positive shift is added to the ASC (thus increasing the market share) with the reverse happening if the model overpredicts the market share.

We would then use this new ASC to calculate the new market share, i.e. $\widehat{CS}^1_{cig}$, which would again be compared to the target market share MS^{RP}_{cig}. If $\widehat{CS}^1_{cig}$ differs from MS^{RP}_{cig}, then $\hat{\delta}^1_{cig}$ needs updating again. Specifically, in iteration *s* of the calibration, we would update the calibrated constant and market share from iteration *s-1* to calculate $\delta^s_{cig} = \hat{\delta}^{s-1}_{cig} + ln\left(\frac{MS^{RP}_{cig}}{\widehat{CS}^{s-1}_{cig}}\right)$. The predicted choice share of all alternatives will match the target market share after a small number of iterations of the above procedure. Of course, if the target market shares are substantially different from those in the estimation data, the calibration required will be so excessive as to undermine the role of the explanatory variables in the model.

ASC calibration has been used relatively rarely in health economics (Fiebig et al., 2011; Sivey et al., 2012; Ghijben et al., 2014; Ride and Lancsar, 2016; Ramos et al., 2018).

Correcting both scale and market shares

In practice, both the scale and market shares will require correction. Here, it is important to recognize that these two processes interact with each other. A correction of the scale will also affect the market shares, as the explanatory variables now matter more, while a correction of the constants will also affect

the model scale. When calibrating both scale and constants, it is then good practice to iterate a few times between the two calibrations to ensure that both the scale and market shares are sufficiently close to the targets set by the analyst. Of course, the degree of precision to be used deserves some thought and is likely application-specific.

SP data: Sample and Choice Experiment

An online discrete choice experiment (DCE) on 2,031 US adult smokers (1531 current smokers; and 500 self-reported recent quitters[4]) was conducted (Buckell et al., 2018). We sampled from the population according to quotas derived from the Behavioral Risk Factor Surveillance System (BRFSS) data in 2013/14 based on gender, age, education and region to make the sample representative. The sample size is well in excess of minimum sample size calculations (de Bekker-Grob et al., 2015). A series of exercises were conducted to promote the quality of the data (e.g. attention checks in the survey, minimum time threshold, etc.; see Buckell et al. (2018) for details). In each scenario, individuals chose between cigarettes, e-cigarettes and an opt-out. The opt-out was labelled as "none of these". Attributes and levels are shown in Table 1. Some levels are omitted to make choices realistic (e.g. fruit/sweet cigarettes are not on the market in the US). This design is based on a review of the literature and a pilot study.

The principle of Bayesian D-optimality was used to generate the experimental design (Hensher et al., 2015). Priors were obtained from analysis of pilot study data on 87 respondents. 36 total choice sets were divided into 3 blocks of 12, and individuals were randomized to each block[5]. Each individual thus answered 12 choice sets which balances concerns of learning and respondent fatigue (Hess et al., 2012). A practice choice scenario was given to all respondents to ensure that they understood how the choice scenarios worked.

	E-cigarette	Cigarette
Flavor	Plain tobacco	Plain tobacco
	Menthol	Menthol
	Fruit	
	Sweet	

Life years lost by average user	10 5 2 Unknown	10
Level of nicotine	High Medium Low None	High Medium Low
Price	$4.99 $7.99 $10.99 $13.99	$4.99 $7.99 $10.99 $13.99

Table 1: Experimental design: Products, attributes and levels

RP data sources

Information on individuals' smoking-related and purchasing behavior was collected alongside the experiment, which has been treated as RP data elsewhere (Kenkel et al., 2017). Specifically, data on individuals' product use, prices and e-cigarette flavors were collected. Individuals whose reported prices were implausible, under $2 and above $20[6], were removed; 1,903 individuals remained. Based on respondents' use, they were categorized as either a smoker (only uses cigarettes, 51%), dual user (uses cigarettes and e-cigarettes, 31%), vaper (uses only e-cigarettes, 7%) or a recent quitter (uses neither cigarettes nor e-cigarettes, 11%).

For scale calibration based on prior literature, we used the results from Pesko et al. (2018) on high school students' tobacco use behavior which gives an own price elasticity of participation for e-cigarettes of -0.54. It is an average of elasticity of participation(s) for two types of e-cigarette (disposable and reusable), as per Jawad et al. (2018). This elasticity is based on reported use of e-cigarettes and store prices from

---

[6] In preliminary modelling, using different cut-offs yielded similar results to those reported.

scanner data. The figure is also broadly in line with participation elasticities for cigarettes in several literature reviews[7] (Chaloupka and Warner, 2000; Gallet and List, 2003; Rice et al., 2010).

ASC calibration is based on the PATH data (Hyland et al., 2017). This is a large, national data set of tobacco use behaviors in the US population. Use of a range of tobacco products is reported. Kasza et al. (2017) analyze this data and provide the % of adults that use each type of tobacco products. Using current use, market shares are derived for cigarettes (66%), e-cigarettes (20%) and all other remaining tobacco products, i.e. cigars, cigarillos, non-combustible tobacco, etc. (14%).

Implementation of correction approaches

In our work, we estimated models on the SP data alone as well as models estimated jointly on the SP and RP data, leading to a direct correction of the scale made possible by the fact that both models use a price coefficient. For models estimated on the SP data alone, we also test the calibration of the scale without joint SP-RP estimation, by adjusting the scale such that the own-price elasticity of choice for e-cigarettes estimated in the model matches the own-price elasticity of participation for e-cigarettes from the literature, as per Equation (3).

Two ASC calibrations are applied. A *full calibration* computes market shares from the PATH data and applies them directly. Here, the outside good market share is the remainder of all tobacco product use in PATH after cigarettes and e-cigarettes (14.0%). Thus, a rather strong assumption is imposed on the outside good: that it represents the choice of these products. In reality, choosing the outside good in the experiment may confer a range of behaviors, namely other tobacco products, the respondent's own brand of cigarette/e-cigarette, cessation behavior, or simply not to purchase the product. To allow for this, we apply a *partial calibration*, wherein the choice share of the outside good is that which occurs in estimation. In other words, the opt-out is uncalibrated. Here, importantly, no assumption is made as to its meaning and so interpretation can generalize to the aforementioned range of possible behaviors. Then, the remaining choice share is divided proportionally between cigarettes and e-cigarettes. The proportion is determined by the ratio of their market shares in the PATH data, 3.3:1[8].

---

[7] This is different to Callison and Kaestner (2014), but we note they use daily use of cigarettes for elasticities; we use current use. Also they only estimate elasticities for cigarette consumption.

[8] We also conducted an approach that used the RP survey data. The outside good choice share is set as the proportion of behaviors that are neither smoking nor vaping (6.8% of the sample) and the same procedure was applied. There are three issues with this approach. One is that this measure ignores the possibility that the outside good was understood to be the respondent's own cigarette/e-cigarette, which is a plausible

A series of forecasts are made using sample enumeration, i.e. applying estimated models to a given choice scenario for each individual in our sample (Hensher et al., 2015). The starting point is the uncalibrated model from the SP data. Here, forecasts are made using the estimated utility weights and state-of-the-world configuration of attributes: the average of respondents' reported prices (since some observations were dropped), tobacco flavor for cigarettes (as it is the most common), fruit flavor for e-cigarettes (as it is the most common), medium level of nicotine. The health harm for cigarettes is set at 10 life years lost (as per the experiment) and for e-cigarettes, it is set at 2 life years lost, as this is closest to estimates of the harms of e-cigarettes (Goneiwitz et al., 2014; Shahab et al., 2017). This is termed the *base* scenario. From this, forecasts are made to predict the impact of e-cigarettes being more harmful than they are currently considered to be. This follows recent research showing e-cigarette-specific harms that are not present for cigarettes (Reidel et al., 2018; Scott et al., 2018; Erythropel et al., 2018). Here, the e-cigarette health harm is set to 5 life years lost; all else is held constant. This is the *e-cigarettes are more harmful* scenario.

The forecast from the *base* scenario to the *e-cigarettes are more harmful* scenario is made with 9 calibrations, (1) to (9). The first is the uncalibrated SP model described above (1). From this, either the scale, ASCs, or both are calibrated. The next two forecasts calibrate the scale only, either by using elasticity from the literature (2) (applying Equation (3)), or SP-RP scale calibration (3) through joint estimation. Next, three forecasts use full ASC calibration: full ASC calibration and no scale calibration (4); full ASC calibration and calibration of scale based on the elasticity from the literature (5); and full ASC calibration with SP-RP calibration (6)[9]. Finally, three forecasts use partial ASC calibration: (7) is partial ASC calibration alone; (8) is partial ASC calibration and scale calibration based on the elasticity from the literature; and (9) is partial ASC calibration and SP-RP scale calibration .

Estimation

---

behavior. Two is that survey data is required for this approach; this data is not always available (i.e. if it is not collected). The third is that it is lower than the outside good choice share of the full calibration, yet is meant to represent a wider range of behaviors than the full calibration opt-out (so should be at least as large, if not larger). For these reasons, this approach was discarded.

[9] NB – the scale is adjusted post-ASC calibration, using the elasticity from its ASC calibration-free counterpart, model (3), and applying the approach in Equation (1). Alternatively, we could have left the scale unadjusted. In preliminary analyses we did this, too, and the results were similar.

We estimate, given the assumption on the error term, a multinomial logit (MNL) model. The model is estimated using R software and user-written code (CMC, 2017).

Limitations

The study has some important limitations. First, despite several measures to promote data quality, we identified several poor-quality responses for respondents' reported own cigarette prices (some being implausibly high or low). These observations were dropped for the RP models. Different thresholds were used to discard responses and models re-estimated; model parameters were stable. More broadly, it is not always the case that RP will yield the true elasticities. As noted by Train (2009), real prices may not vary much and even if consumers are highly price elastic, this may not be reflected in the data given the small cost changes. Consumers' decision may be dominated by other attributes, and/or there may be unobservable correlation between price and other attributes (e.g. quality)[10]. We believe that in the cigarette context, this may be less of an issue as real-world prices do vary over time. An additional issue is that, the survey data, whilst considered RP, is ultimately self-reported data. Therefore, is it subject to bias from misreporting, stigma, etc. (Cawley and Ruhm, 2011). Note that by survey data we also refer to PATH which is also self-reported data. Second, whilst we have developed a solution to the ambiguity in interpretation of the opt-out in our modelling, it would be useful to understand how individuals consider the opt-out option. To this end, several opt-outs, with different labels (e.g. "I would choose my own cigarettes" or "I would rather quit" instead of just "none of these" as is the case in our experiment) could be used; or some follow-up questions in the survey post-experiment. Finally, we note that the own-price elasticity calibration was conducted using own-price elasticity of participation; whereas in the experiment, the margin is the elasticity of choice, which is a similar, but different concept (Hensher et al., 2015). Own-price elasticity of participation measures the response of participation (i.e. whether or not the individual uses a particular product) to variation in that product's price. Own-price elasticity of choice measures the response of choice (i.e. the probability of choosing a product from a set of available options) to variation in that product's price. We further note that prices derive from retail scanner data only; of course, much of the e-cigarette sales occur online. One estimate suggests that online sales account for around 25% of e-cigarette sales in the US (Herzog and Kanada, 2018). To the best of our knowledge, no RP estimates of the choice elasticity of cigarettes or e-cigarettes are available in the literature.

---

[10] We thank the reviewer for noting the latter point.

3. Results

Choice models

Table 2 reports the model estimates for the three choice models: RP, SP and joint SP-RP. Similarities in SP and RP preferences can be seen between the two models: all else being equal, cigarettes are preferred to e-cigarettes (comparing the ASCs in both models); respondents prefer lower prices; and fruit flavor is preferred to menthol (though it is not a perfect comparison because in the SP data, menthol flavors apply for both cigarettes and e-cigarettes; in the RP data, menthol flavors (the omitted category) are for e-cigarettes only). In the SP-RP model, the scale of the SP data is brought in line with the RP data via the scale parameter (note the similarity in the RP and SP-RP preference parameters). The μ_{SP} parameter is larger than 1, implying higher scale for the SP data, in line with expectations, albeit that it is only significantly different from 1 at low levels of significance. This suggests that responses in the SP setting exaggerate the effects of attributes on choices. This can be readily observed via μ_{SP} (=1.43). With only the price coefficient being shared between the two models, the estimation of the μ parameter in the SP-RP model means that this model does not impose any restrictions compared to the two separate models and the log-likelihood of the SP-RP model is thus the sum of the SP and RP models. In other words, there are the same number of parameters in the joint model and the sum of the separate models; and the fit of the model is unchanged (though the form of the model and its interpretation change).

	Model (i): RP MNL				Model (ii): SP MNL				Model (iii): Joint SPRP MNL			
	Estimate	Rob s.e.	t-ratio (0)	t-ratio (1)	Estimate	Rob s.e.	t-ratio (0)	t-ratio (1)	Estimate	Rob s.e.	t-ratio (0)	t-ratio (1)
δ (ASC: cigarette)					2.45	0.06	42.62	25.21	1.72	0.46	3.77	1.57
δ (ASC: e-cigarette)					1.55	0.07	23.06	8.19	1.09	0.29	3.75	0.30
β(price)	-0.07	0.02	-3.87	-60.41	-0.10	0.00	-30.03	-337.33	-0.07	0.02	-3.79	-58.99
β(no nicotine)					-0.05	0.03	-1.57	-31.43	-0.04	0.03	-1.45	-40.94
β(low nicotine)					-0.05	0.02	-1.84	-42.01	-0.03	0.02	-1.65	-52.92
β(high nicotine)					0.00	0.02	0.01	-49.81	0.00	0.01	0.01	-70.78
β(menthol flavor)					-0.34	0.03	-9.98	-39.41	-0.24	0.07	-3.53	-18.39
β(fruit flavor)					-0.20	0.04	-5.19	-30.94	-0.14	0.05	-3.04	-24.50
β(sweet flavor)					-0.15	0.03	-4.24	-32.89	-0.10	0.04	-2.77	-29.39
β(unknown years life lost)					0.48	0.05	10.09	-11.07	0.33	0.10	3.51	-6.97
β(2 years life lost)					0.67	0.05	13.27	-6.41	0.47	0.13	3.60	-4.01
β(5 years life lost)					0.17	0.04	3.94	-18.64	0.12	0.05	2.65	-19.04
δ (ASC: dual user)	-1.05	0.14	-7.60	-14.82					-1.05	0.14	-7.45	-14.53
δ (ASC: vaper)	-3.54	0.20	-17.37	-22.27					-3.54	0.21	-17.19	-22.05
δ (ASC: recent quitter)	-2.30	0.16	-14.79	-21.22					-2.30	0.16	-14.55	-20.88
β(fruit flavor)	1.52	0.13	11.46	3.90					1.52	0.13	11.46	3.90
β(other flavor)	0.35	0.46	0.77	-1.42					0.35	0.46	0.77	-1.42
β(no flavor)	1.80	0.30	6.09	2.71					1.80	0.30	6.08	2.71
μ_{SP}									1.43	0.38	3.77	1.12
Number of Parameters	7				12				19			
LL(b)	-1991.63				-39407.82				-41399.5			
LL(0)	-2638.12				-43668.76				-46306.88			
Observations	1903				24372				26275			
Individuals	1903				2031				2031			

Table 2: Multinomial logit (MNL) choice model results. RP – revealed preference; SP – stated preference. For the SP model, the omitted: choice option is the outside good; level of nicotine is "medium"; flavor is tobacco; and health harm is "10 years life lost". For the RP model, the omitted: status is "smoker"; and flavor is "menthol". In both models, price is treated continuously. Estimate – estimated parameter(s); Rob s.e. – robust standard errors (clustered by individual); t-ratio (0) – t-ratio of estimated parameter = 0; t-ratio (1) – t-ratio of estimated parameter = 1; LL(b) – log-likelihood of the fitted model; LL(0) – log-likelihood of the null model.

Forecast	Model	ASC calibration	Scale calibration	Elasticity	Base scenario			E-cigarettes are more harmful scenario			Relative change in choice share		
					Cigarette	E-cigarette	Opt-out	Cigarette	E-cigarette	Opt-out	Cigarette	E-cigarette	Opt-out
1	(ii)	None	None	-0.46	54.9	35.8	9.3	63.9	25.3	10.8	16.4%	-29.4%	16.5%
2	(ii)	None	Elasticity, literature	-0.54	57.5	35.1	7.4	68	23.3	8.7	18.2%	-33.6%	18.1%
3	(iii)	None	Elasticity, SPRP	-0.32	49.3	36.6	14.2	55.2	28.9	15.9	12.1%	-21.1%	12.1%
4	(ii)	Full	None	-0.57	66	20	14	71.6	13.2	15.2	8.5%	-34.0%	8.6%
5	(ii)	Full	Elasticity, literature	-0.54	66	20	14	71.4	13.4	15.2	8.2%	-32.7%	8.2%
6	(iii)	Full	Elasticity, SPRP	-0.32	65.9	20.2	13.9	69.4	16	14.6	5.3%	-20.8%	5.0%
7	(ii)	Partial	None	-0.57	65.8	20	14.2	71.4	13.2	15.4	8.5%	-34.1%	8.6%
8	(ii)	Partial	Elasticity, literature	-0.54	65.9	19.9	14.2	71.3	13.4	15.3	8.1%	-32.7%	8.1%
9	(iii)	Partial	Elasticity, SPRP	-0.32	65.9	20	14.1	69.4	15.8	14.8	5.2%	-20.8%	5.2%

Table 3: Forecasting with and without calibrations. Forecast – the number of the forecast; Model – the choice model (from table 2) used in the forecast; ASC calibration – the type of ASC calibration applied; Scale calibration – the type of scale calibration applied; Elasticity – price elasticity of choice for e-cigarettes; Base scenario – the forecasted choice shares per product from the state-of-the-world configuration; E-cigarettes are more harmful scenario - the forecasted choice shares per product when e-cigarette health harm is increased; Relative change in choice share – is the proportional change in the forecasted choice share per product from the base scenario to the e-cigarettes are more harmful scenario.

Forecasts

Table 4 presents 9 forecasts. For each forecast, the models used, calibration configurations and the choice elasticities are shown. Two choice shares per product are shown: the *base* scenario and the *e-cigarettes are more harmful* scenario. The difference between these is the prediction of the impact of e-cigarettes being more harmful, which is presented (in % terms) in the furthest right columns. Forecast (1), in which both scale and ASCs are uncalibrated, is the raw SP prediction. Different forms of calibration are applied from (2) to (9).

Overall, the calibration impacts the relative changes in choice shares considerably. That is, for the same prediction (e-cigarettes are more harmful than currently thought), the change in the predicted choice shares of the products varies according to the calibration applied. For cigarettes, the range is 5.2% to 18.2%; for e-cigarettes the range is 20.8% to 34.1%; and for the outside good the range is 5% to 18.1%.

Forecasts in which the ASCs are uncalibrated, (1) to (3), under-predict the choice share of cigarettes and over-predict the choice share of e-cigarettes. Cigarette base scenario choice shares of ASC-calibrated forecasts are around 66%; cigarette choice shares of ASC-uncalibrated forecasts are not close, ranging from 49% to 58%. E-cigarette choice shares of ASC-calibrated forecasts are around 20%; cigarette choice shares of ASC-uncalibrated forecasts are not close, ranging from 35% to 37%. This shows the importance of ASC calibration.

Forecasts (1) to (3) also highlight two important points about scale calibration. The first is that predictions from ASC-uncalibrated forecasts, in terms of relative choice share, are very different from their ASC-calibrated counterparts for cigarettes and the opt-out (though broadly comparable for e-cigarette choice shares). Second, as noted in the methods, the impact of scale calibration depends on the relative values of the RP and SP elasticities. In the case that $\left|e_k^0\right| \leq \left|e_k^*\right|$, as per forecast (2), then the overall scale of utility is increased with the calibration, the impact of attributes on utility is greater, and the forecasts are more sensitive. Thus the relative changes in forecast (2) are larger than in forecast (1). The reverse is true when $\left|e_k^0\right| \geq \left|e_k^*\right|$, as per forecasts (3); here, the relative changes in forecasts (3) are all smaller than in forecasts (1).

The full (forecasts (4) to (6)) and partial (forecasts (7) to (9)) ASC calibrations, are very similar. For the full calibration, by construction, the base scenario choice shares for all products are in line with the RP data. For the partial calibration, the choice share of the outside good is that which occurs without the ASC calibration applied[11]. Coincidentally, this happens to be similar, though larger, than the choice share of the outside good in the full calibration. Thus, the choice shares and relative changes of both approaches are similar. Since the outside good in the partial calibration represents a wider range of behaviors than the outside good in the full calibration, it is encouraging that its choice share is larger (though only slightly). As such, this is our preferred approach.

The ASC-calibrated forecasts, in terms of relative change, are substantially different from the ASC-uncalibrated forecasts, particularly for the double-calibrations (i.e. calibrating both scale and ASCs). This suggests that the joint calibration of both scale and ASCs is critical. For cigarettes and the outside good, the relative change is considerably reduced; for e-cigarettes the relative change is comparable. As before, the impact on the predictions of the scale calibration depends on the ratio of the calibrated elasticity to the uncalibrated elasticity. For example, comparing forecast (7) to forecasts (8) or (9), we see that these ratios are < 1. In turn, the forecasts are reduced. That is, the scale calibration reduces the impact of the change in the attribute on the choice shares. The literature RP scale calibration's forecasts are similar to those without scale calibration; the SP-RP scale calibration's forecasts are much lower, resulting in the most modest relative change(s) in choice share(s) of all models. This shows that the choice of the elasticity used to calibrate the scale of utility impacts on the predicted forecasts. Since the SP-RP approach is better suited to our data, this is our preferred approach.

4. Summary and Discussion

In this paper we consider methods for correcting forecasts from choice models estimated on discrete choice experiments in health, with a view to accounting for hypothetical bias. The main aim of this study is to show a range of techniques for calibrating choice models and the impact on model forecasts of doing so. We use a SP experiment of smokers' tobacco product choices and RP data from several sources. We estimate SP and RP models based on these data; and a joint SP-RP model on both. A set of forecasts based on a range of calibrations is conducted. Three features of calibration are assessed: calibrating both the

---

[11] This is similar to the same pre-calibration choice share of models (7) and (3).

scale of utility and alternative-specific constants; using different approaches to calibrating the scale of utility; and using different approaches to calibrating the alternative-specific constants.

The results indicate that forecasts in this setting are sensitive to both forms of calibration. Uncalibrated models under-predict cigarette choice and over-predict e-cigarette choices. When looking at predictions for future scenarios, uncalibrated forecasts over-estimate changes in choice shares, reflecting the common concern that hypothetical data alone tends to lead to inflated elasticities (though this is only a single example). Jointly calibrating both scale and model constants has a considerable impact on forecasts.

The impact of scale calibration depends on the elasticity used for calibration. Target elasticities that are larger (demand is more elastic) than the SP elasticity increase the scale of utility; and target elasticities that are smaller (demand is more inelastic) than the SP elasticity decrease the scale of utility. As shown, these cases have opposing effects on forecasts. Thus, the choice of elasticity for scaling can have a significant bearing on the forecasts of the impact of changes in attributes on choices.

In this study, we had two candidate elasticities: one based on findings from the literature, and one based on our RP data. We prefer the RP data-based elasticity for a number of reasons. First, the elasticity from the literature is based on youths, not adults, as in the experimental data. Second, whilst the figure is close to adult participation elasticities found elsewhere in the literature, we note that these other elasticities are rather dated; at this time, e-cigarettes were not available on the market. Further note that other elasticities are for cigarettes; we used e-cigarette participation elasticities here. It is not clear, therefore, that these are appropriate measures for the task at hand. Third, the e-cigarette participation elasticities are imprecisely estimated in the source paper. In fact, they are not significantly different from zero at usual levels of confidence. Conversely, our data the RP elasticity is for the correct population, the correct product, and the coefficient is precisely estimated (cf. table 2). Moreover, when we applied this calibration, the scale of utility was reduced, which was the expected impact based on the prior literature. Thus, our preferred models are those based on SP-RP elasticity calibration for the scale.

Partial calibration allows for a general interpretation of the choice of the outside good that does not require assumptions about the implied behavior. In the full calibration, the assumption on the opt-out is that it represents the choice of tobacco products in the PATH data other than cigarettes or e-cigarettes (e.g. smokeless tobacco, cigars, etc.). In reality, the choice of the opt-out could represent a wider set of

behaviors, e.g. not to purchase any of the options, cessation behavior, inter alia. In our results, the outside good choice share in the partial calibration was larger (although only slightly), which should be the case since it represents a wider set of possible behaviors than those imposed by using the PATH data alone. With this, the restrictive assumption of the full calibration is relaxed and the predictions are arguably more realistic. This is our preferred approach. Of course, the partial calibration may not be necessary if the opt-outs are defined clearly enough. Reasoning should be applied to determine the optimal approach. Indeed, an alternative approach (using RP data from our own survey) yielded a much smaller choice share for the outside good and thus, we argue, less realistic forecasts; we duly rejected this approach.

The preceding elasticity and outside good discussions highlight an important further point. Just because one can calibrate choice models, doesn't mean one should. It is tempting to assume that any use of RP data will necessarily improve forecasts. In fact, however, there may be imperfections with the RP data itself. The data itself may lack variation, e.g. if prices in a market do not vary, attributes are correlated, consumers are price inelastic (Train, 2009). In the real world, certain attributes may dominate consumers' decisions, or consumers may avoid information (Golman et al., 2017). Further, there may be practical difficulties, such as very large choice sets in the real world that are difficult to reduce for modelling (Brownstone et al,. 2000). Further, as is the case here, the application of the elasticity may be questionable if the nature of the elasticity differs between the literature and the application, the populations are different, the measure of elasticity is not statistically significant, or that part of the market is not covered (i.e. the elasticity data we used includes store prices but not online prices). Not only that, what constitutes RP data should be considered; what we term RP here, both our survey data and PATH, is ultimately self-reported (i.e. SP) data. Researchers should be aware of these issues when selecting – or not – RP data for calibration. In passing, we further note that in many cases, particularly in health, RP in its traditional form is unavailable.

Overall, the results show that, even in a setting such as tobacco where hypothetical bias should be relatively limited, substantial differences to forecasts are found in calibration of all forms. Large differences are observed comparing the uncalibrated SP forecasts in (1) to the doubly-calibrated forecasts in (9), and the forecasts in between. Therefore, two key points arise from this study. First, that calibration of both scale and constants can make a substantial difference to the forecasts from DCE studies. Where possible (RP data is not always available in health; nor is a suitable elasticity in the literature), and, where appropriate (e.g. the RP and SP data are well-suited, which was not the case in this study, but may be

elsewhere), calibration of both kinds should be conducted. Applications in health hitherto have used one form of calibration only; and numerous applications (including several tobacco forecasts) make forecasts without any calibration. Second, the means by which calibration is conducted can have a substantial impact on forecasts. Accordingly, reasoning should be applied when selecting sources of data for calibrating choice models. Calibrating choice modes can help to abate hypothetical bias in SP data and provide better quality empirical evidence for policymakers.

References

Ben-Akiva M, Morikawa T. Estimation of switching models from revealed preferences and stated intentions. *Transportation Research Part A: General.* 1990;24(6):485-495.

Beck MJ, Fifer S, Rose JM. Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. *Transportation Research Part B: Methodological.* 2016;89:149-167.

Bradley M, Daly A. Use of the logit scaling approach to test for rank-order and fatigue effects in stated preference data. Transportation. 1994;21(2):167-84.

Brownstone D, Bunch DS, Train K. Joint mixed logit models of stated and revealed preferences for alternative-fuel vehicles. Transportation Research Part B: Methodological. 2000;34(5):315-38.

Buckell J, Marti J, Sindelar JL. Should flavours be banned in cigarettes and e-cigarettes? Evidence on adult smokers and recent quitters from a discrete choice experiment. Tobacco Control. 2018:tobaccocontrol-2017-054165.

Callison K, Kaestner R. DO HIGHER TOBACCO TAXES REDUCE ADULT SMOKING? NEW EVIDENCE OF THE EFFECT OF RECENT CIGARETTE TAX INCREASES ON ADULT SMOKING. Economic Inquiry. 2013;52(1):155-72.

Cawley J, Ruhm CJ. Chapter Three - The Economics of Risky Health Behaviors1. In: Mark V. Pauly TGM, Pedro PB, editors. Handbook of Health Economics. Volume 2: Elsevier; 2011. p. 95-199.

Chaloupka FJ, Warner KE. Chapter 29 The economics of smoking. Handbook of Health Economics. Volume 1, Part B: Elsevier; 2000. p. 1539-627.

Clark M, Determann D, Petrou S, Moro D, de Bekker-Grob E. Discrete Choice Experiments in Health Economics: A Review of the Literature. PharmacoEconomics. 2014;32(9):883-902.

CMC (2017), CMC choice modelling code for R, Choice Modelling Centre, University of Leeds, www.cmc.leeds.ac.uk, accessed 9/7/2018

de Bekker-Grob EW, Ryan M, Gerard K. Discrete choice experiments in health economics: a review of the literature. Health Economics. 2012;21(2):145-72.

de Bekker-Grob E, Donkers B, Jonker M, Stolk E. Sample Size Requirements for Discrete-Choice Experiments in Healthcare: a Practical Guide. The Patient - Patient-Centered Outcomes Research. 2015;8(5):373-84.

Donfouet HP, Mahieu PA, Malin E. Using respondents' uncertainty scores to mitigate hypothetical bias in community-based health insurance studies. *The European journal of health economics.* 2013;14(2):277-285.

Erythropel HC, Jabba SV, DeWinter TM, et al. Formation of flavorant–propylene Glycol Adducts With Novel Toxicological Properties in Chemically Unstable E-Cigarette Liquids. *Nicotine & Tobacco Research.* 2018:nty192-nty192.

Fiebig. FD, Stephanie K, Rosalie V, Marion H, J. SD. Preferences for new and existing contraceptive products. Health Economics. 2011;20(1):35-52.

Fifer S, Rose J, Greaves S. Hypothetical bias in Stated Choice Experiments: Is it a problem? And if so, how do we deal with it? *Transportation Research Part A: Policy and Practice.* 2014;61:164-177.

Gallet CA, List JA. Cigarette demand: a meta-analysis of elasticities. Health Econ. 2003;12.

Ghijben P, Lancsar E, Zavarsek S. Preferences for Oral Anticoagulants in Atrial Fibrillation: a Best–Best Discrete Choice Experiment. PharmacoEconomics. 2014;32(11):1115-27.

Golman R, Hagmann D, Loewenstein G. Information Avoidance. *Journal of Economic Literature.* 2017;55(1):96-135.

Goniewicz ML, Hajek P, McRobbie H. Nicotine content of electronic cigarettes, its release in vapour and its consistency across batches: regulatory implications. Addiction. 2014;109(3):500-7.

Harrison G. Real choices and hypothetical choices. In: Hess S, Daly A, editors. Handbook of Choice modelling. Cheltenham: Edwards Elgar Publishing; 2014.

Hausman J. Contingent Valuation: From Dubious to Hopeless. Journal of Economic Perspectives. 2012;26(4):43-56.

Hensher DA, Bradley M. Using stated response choice data to enrich revealed preference discrete choice models. Marketing Letters. 1993;4(2):139-51.

Hensher D, Louviere J, Swait J. Combining sources of preference data. Journal of Econometrics. 1998;89(1–2):197-221.

Hensher D, Rose JM, Greene W. Applied Choice Analysis. Second ed. Cambridge: Cambridge University Press; 2015.

Herzog, B & Kanada, P. Nielsen: Tobacco 'All Channel' Data 1/27. Wells-Fargo. 2018-12-05. URL: https://1lbxcx1bcuig1rfxaq3rd6w9-wpengine.netdna-ssl.com/wp-content/uploads/2018/02/Nielsen-Tobacco-All-Channel-Report-Period-Ending-1.27.18.pdf. Accessed: 2018-12-05. (Archived by WebCite® at http://www.webcitation.org/74RUD2GEs)

Hess S, Daly A. *Handbook of Choice Modelling.* Cheltenham: Edward Elgar; 2014.

Hess S, Daly A, Dekker T, Cabral MO, Batley R. A framework for capturing heterogeneity, heteroskedasticity, non-linearity, reference dependence and design artefacts in value of time research. *Transportation Research Part B: Methodological.* 2017;96:126-149.

Hess S, Hensher DA, Daly A. Not bored yet – Revisiting respondent fatigue in stated choice experiments. Transportation Research Part A: Policy and Practice. 2012;46(3):626-44.

Hyland A, Ambrose BK, Conway KP, Borek N, Lambert E, Carusi C, et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. Tobacco Control. 2017;26(4):371-8.

Jawad M, Lee JT, Glantz S, Millett C. Price elasticity of demand of non-cigarette tobacco products: a systematic review and meta-analysis. Tobacco Control. 2018;27(6):689.

Kasza KA, Ambrose BK, Conway KP, Borek N, Taylor K, Goniewicz ML, et al. Tobacco-Product Use by Adults and Youths in the United States in 2013 and 2014. New England Journal of Medicine. 2017;376(4):342-53.

Kenkel D, Peng S, Pesko M, Wang H. Mostly Harmless Regulation? Electronic Cigarettes, Public Policy And Consumer Welfare. National Bureau of Economic Research Working Paper Series, No 23710; 2017.

Kesternich I, Heiss F, McFadden D, Winter J. Suit the action to the word, the word to the action: Hypothetical choices and real decisions in Medicare Part D. Journal of Health Economics. 2013;32(6):1313-24.

Lancsar E, Burge P. Choice modelling research in health economics. In: Hess S, Daly A, editors. Handbook of Choice Modelling. Cheltenham: Edward Elagar Publishing; 2014.

Lancsar E, Louviere J. Conducting Discrete Choice Experiments to Inform Healthcare Decision Making. PharmacoEconomics. 2008;26(8):661-77.

Lancsar E, Swait J. Reconceptualising the External Validity of Discrete Choice Experiments. PharmacoEconomics. 2014;32(10):951-65.

List JA, Gallet CA. What Experimental Protocol Influence Disparities Between Actual and Hypothetical Stated Values? *Environmental and Resource Economics.* 2001;20(3):241-254.

Little J, Berrens R. Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. *Econ Bull.* 2004;3(6):1-13.

Louviere JJ, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Economics, Policy and Law. 2009;4(04):527-46.

Mark TL, Swait J. Using stated preference and revealed preference modeling to evaluate prescribing decisions. Health Economics. 2004;13(6):563-73.

Marti J, Buckell J, Maclean JC, Sindelar J. TO "VAPE" OR SMOKE? EXPERIMENTAL EVIDENCE ON ADULT SMOKERS. Economic Inquiry. 2019;57(1):705-25.

McFadden D. The New Science of Pleasure: Consumer Choice Behavior and the Measurement of Well-Being. In: Hess S, Daly A, editors. Handbook of Choice Modelling. Cheltenham: Edward Elgar; 2014.

Murphy JJ, Allen PG, Stevens TH, Weatherhead D. A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. *Environmental and Resource Economics.* 2005;30(3):313-325.

Özdemir S, Johnson FR, Hauber AB. Hypothetical bias, cheap talk, and stated willingness to pay for health care. *Journal of Health Economics.* 2009;28(4):894-901.

Pesko M F, Huang J, Johnston LD, Chaloupka FJ. E-cigarette price sensitivity among middle- and high-school students: evidence from monitoring the future. Addiction. 2018;113(5):896-906.

Quaife M, Terris-Prestholt F, Di Tanna GL, Vickerman P. How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity. *The European Journal of Health Economics.* 2018;19(8):1053-1066.

Rakotonarivo OS, Schaafsma M, Hockley N. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. *Journal of Environmental Management.* 2016;183:98-109.

Ramos P, Alves H, Guimarães P, Ferreira MA. Junior doctors' medical specialty and practice location choice: simulating policies to overcome regional inequalities. The European Journal of Health Economics. 2017;18(8):1013-30.

Rice N, Godfrey C, Slack R, Sowden A, Worthy G. A systematic review of the effects of price on the smoking behaviour of young people. UK: PHRC Project Outputs; A2-06; 2010.

Ride J, Lancsar E. Women's Preferences for Treatment of Perinatal Depression and Anxiety: A Discrete Choice Experiment. PLOS ONE. 2016;11(6):e0156629.

Reidel B, Radicioni G, Clapp PW, Ford AA, Abdelwahab S, Rebuli ME, et al. E-Cigarette Use Causes a Unique Innate Immune Response in the Lung, Involving Increased Neutrophilic Activation and Altered Mucin Secretion. Am J Respir Crit Care Med. 2018;197(4):492-501.

Ryan M, Watson V. Comparing welfare estimates from payment card contingent valuation and discrete choice experiments. *Health Economics.* 2008;18(4):389-401.

Scott A, Lugg ST, Aldridge K, et al. Pro-inflammatory effects of e-cigarette vapour condensate on human alveolar macrophages. *Thorax.* 2018;73(12):1161.

Shahab L, Goniewicz ML, Blount BC, et al. Nicotine, carcinogen, and toxin exposure in long-term e-cigarette and nicotine replacement therapy users: A cross-sectional study. Annals of Internal Medicine. 2017;166(6):390-400.

Sivey P, Scott A, Witt J, Joyce C, Humphreys J. Junior doctors' preferences for specialty choice. Journal of Health Economics. 2012;31(6):813-23.

Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete Choice Experiments in Health Economics: Past, Present and Future. *PharmacoEconomics.* 2018.

Swait J, Louviere J. The role of the scale parameter in the estimation and comparison of multinational logit models. *Journal of Marketing Research.* 1993 08;30(3):305.

Train K. Discrete choice methods with simulation. Cambridge: Cambridge University Press; 2009.

Vass CM, Wright S, Burton M, Payne K. Scale Heterogeneity in Healthcare Discrete Choice Experiments: A Primer. *The Patient - Patient-Centered Outcomes Research.* 2018;11(2):167-173.

Whynes DK, Philips Z, Frew E. Think of a number… any number? *Health Economics.* 2005;14(11):1191-1195.

Wright SJ, Vass CM, Sim G, Burton M, Fiebig DG, Payne K. Accounting for Scale Heterogeneity in Healthcare-Related Discrete Choice Experiments when Comparing Stated Preferences: A Systematic Review. *The patient.* 2018;11(5):475-488.

Wuepper D, Clemm A, Wree P. The preference for sustainable coffee and a new approach for dealing with hypothetical bias. *Journal of Economic Behavior & Organization.* 2018.