Harnessing a large observational database to interpret the results of a stated preference survey: the case of blood donation

Kaat De Corte[1], Sarah Willis[1], Silvia Perra[1], Mark Pennington[2], Neil Hawkins[1], John Cairns[1], Richard Grieve[1] for the HEMO group

[1] Department of Health Services Research & Policy, London School of Hygiene and Tropical Medicine
[2] King's Health Economics, King's College London

This project was funded by the National Institute for Health Research Health Services and Delivery Research (project number 13/54/62). The views and opinions expressed therein are those of the authors and do not necessarily reflect those of the Health Services and Delivery Research Programme, NIHR, NHS or the Department of Health.

Abstract

Stated preference (SP) methods are used extensively in health economics to elicit preferences from individuals to provide policy-relevant information. The validity of these methods lies in the extent to which hypothetical choices mirror real preferences. Recent progress has been made in externally validating stated preferences. However, attempts to do so in a health context are limited, and have yet to exploit the growing availability of large, longitudinal datasets.

This paper describes a large online SP survey to inform the cost-effectiveness of alternative policy initiatives proposed by the National Blood and Transplant service (NHSBT). NHSBT supply blood to the NHS at an annual cost of £180 million, and wish to maintain the current blood supply, but at a lower cost. Some policies under consideration are novel, such as the provision of a health report; others have been partially adopted, such as change of opening hours. The latter can be considered through the analysis of both the SP survey (5000 invitees) and a linked longitudinal dataset of 1.2 million blood donors in England (the PULSE database).

The SP survey was developed iteratively. Five attributes were used: travel time, opening hours, total donation time, provision of a health report and the number of permitted donations per year. Choices surrounding these attributes and their levels were governed by policy-relevance and preliminary findings of qualitative research. We asked donors "*how many times a year would you donate?"* to estimate relative frequencies of blood donation. To incorporate the feeling of 'belonging' expressed by some donors about their usual panel, donors were first asked to think about the *"last place you gave blood"* and later, a *"different place"*. Donors were invited to complete two independently orthogonal designs, based on the need to estimate the marginal rate of substitution (MRS) between attributes with reasonable precision.

The choice data were analysed to estimate the MRS across each potential pair of service attributes. The predictions of donation frequency from the SP survey were contrasted with observed donation frequency from the PULSE dataset. We found that overall, the average predicted frequencies of blood donation from the SP survey, exceeded those observed from PULSE. These discrepancies were small for some donor subgroups (e.g. older blood donors).

The paper discusses how revealed preference data from a large longitudinal dataset can inform the interpretation of a SP study for policy purposes and outlines the implications for the design and analysis of future SP surveys.

1. Introduction

Stated preference (SP) methods are widely used in health economics, for example to estimate the relative value of alternative service changes (De Bekker-Grob et al., 2012), and to elicit the willingness to pay for health gain. Progress has been made in the development of innovative SP designs (Street et al., 2001), and analytical methods that respect the structure of SP data (Louviere, 2006). However, a barrier to the use of results from SP surveys in decision-making is that actual behaviour does not reflect the preferences stated (Viney et al., 2002). In the general literature on SP methods, comparisons of stated and revealed preferences (RP) have shown that individuals surveyed tend to overstate their valuation of a particular good, service or outcome, which can lead to misleading estimates of relative value (Fifer et al., 2014). In the health context, the opportunities for contrasting SP and RP have been limited, with little attention given to understanding the reasons for any differences (Ryan et al., 2010). These attempts have not exploited the growing availability of large, longitudinal datasets, to estimate the magnitude of the discrepancy between SP and RP, the factors that predict this discrepancy and the implications for predicting the effects of policy change.

Previous comparisons of SP and RP have aimed to establish if responses to SP surveys represent underlying preferences, that is whether the survey results have 'external validity' (Ryan and Gerard, 2003). The bias from individuals overstating their preferences for a particular set of choices has also been termed 'hypothetical bias' (Johansson-Stenman and Svedsater, 2012). To reduce this bias, the survey design can attempt to replicate the constraints that individuals face, but there is no consensus on how to minimise this bias, nor on whether some level of bias is inevitable. List and Gallet (2001) and Murphy et al. (2005) conducted meta-analyses of the discrepancies between SP and RP reported in the contingent valuation literature, including privately and publicly funded goods. Both studies reported that individuals overstate their preferences in hypothetical settings, and that the magnitude of this discrepancy was higher for publicly funded goods. Murphy et al. also found that the magnitude of the discrepancy differed according to the study design. Those studies which used RP and SP for the same individuals (within-subject design) generally reported smaller estimates of discrepancy, versus those that contrasted RP and SP across different samples (between-subject design), where the potential for hidden bias in the estimated discrepancy, is greater, due to unobserved differences between the individuals.

The general literature on SP modelling outside the health economics context has postulated reasons for hypothetical bias. Johnson-Stenman and Svedsater (2012) draw on evidence from psychology and behavioural economics, which suggests that people have a positive self-image, and have an incentive to overstate their preferences, particularly for what has been termed 'moral goods'. People derive positive utility from expressing certain attitudes, including those that show social responsibility (Taylor and Brown, 1994), especially when the expression of those attitudes is not binding (Kahneman and Knetsch, 1992). However, as Ryan et al. (2010) note, there has been limited work on examining hypothetical bias (or external validity) in preference modelling within areas of direct relevance to health economics, and further research to improve predictions from SP surveys is needed.

The aim of this paper is to illustrate an approach for estimating the magnitude of the discrepancy between SP and RP. Our running example is a study that estimates the effect of alternative blood donation service configurations on the frequency with which individuals are willing to donate blood. The paper proceeds as follows: the next section introduces the motivating example, section 3 provides an overview of the design of the SP survey, section 4 defines the source and collection of data for RP, section 5 describes an overview of the analytical methods for analysing the responses to the SP survey and for undertaking the discrepancy analysis, section 6 contains the results, and section 7 a discussion including an agenda for further research.

2. The motivating example: a SP survey of blood donors

2.1 Overview

National Health Service Blood and Transplant (NHSBT) supply blood to the NHS in England at an annual cost of £180 million, and wish to maintain the current supply of whole blood, but at a lower cost. To inform future policy, NHSBT require estimates of the relative costs and consequences of future service changes, some of which are novel, such as the provision of a health report; others have been partially adopted, such as changes to opening hours of the blood collection centres. For NHSBT, the key consequence of interest is the frequency with which donors are willing to donate whole blood in the future. A key objective of the study was therefore to survey donors to estimate the frequency of blood donation under alternative service configurations. A discrete choice experiment (DCE), where donors indicate which of two or more blood donation opportunities they prefer, would have allowed the estimation of the marginal rate of substitution between attributes but would not have provided estimates of the donation frequencies required for the decision model evaluating the relative value of alternative collection strategies. Instead, we undertook a SP survey, which allowed us to estimate the importance of the alternative blood service attributes on frequency of donation (predicted frequencies of donation). In other words, the SP survey was designed to estimate the frequency with which blood donors are willing to donate blood under alternative future blood service configurations. For each set of attribute levels, the donor was asked to state the frequency with which they would be willing to donate blood. The survey included an opt-out ("*I would probably not donate*"). An example question is shown in Figure 1.

A major concern with using SP to predict actual blood donation frequency according to alternative changes to the blood service is that, individuals may overstate the frequency with which they are willing to donate blood. Indeed the study recognised *a priori* that for an altruistic activity such as blood donation, individual's stated preferences would overstate their revealed preferences.
To address this concern, the study was designed to contrast the preferences for alternative frequencies of blood donation with actual donation activity in the past twelve months. The donors invited to complete the SP survey were from a large longitudinal dataset, the PULSE database, which contains information on the actual frequency of blood donations for all 1.2 million

registered blood donors in England, with a median period of follow-up of 5 years. The database also contains details on the characteristics of the blood donors and of the blood collection centres. This design enabled estimation of the discrepancies between predicted donation frequencies derived from the SP survey, and the observed frequency for individual donors. The estimated discrepancies can then be analysed by donor characteristics.

3. SP survey design

3.1 Choice of attributes and levels

Even though a blood donation is at least partly an altruistic act (Rapport and Maggs, 2002, Titmuss, 1970), it is assumed that the frequency at which donors give blood still depends on the convenience of the opportunities they have to donate and the costs of donation (Schreiber et al., 2006). As the costs of donation increase (decrease) donors will donate less (more). Our choice of attributes to describe the opportunities to donate and costs of donation was informed by a process of rapid literature review, preliminary findings from qualitative research with blood donors involved in a large randomised controlled trial (INTERVAL) and input from policy makers at NHSBT. We adopted five attributes related to strategies of immediate policy relevance: travel time, opening hours, total donation time, provision of a health report and the maximum number of donations per year (see Table 1). The first three attributes influence the cost to the donor of the blood donation, whereas the provision of a health report is anticipated to increase the benefit to the donor (Mews, 2013, Mortimer et al., 2013, Goette et al., 2009, Ringwald, 2010). The ongoing INTERVAL trial is investigating the safety of increasing the maximum frequency of donation (from 3 to 4 times per year for women, from 4 to 6 for men). This attribute is therefore included to understand how donors might respond if the limits were altered. The appropriate levels for each attribute were defined according to summary estimates from the PULSE database, NHSBT market research and consultation with blood donors. Donation venue was not included as an attribute but the survey had two sections: the first, which asks the donor to think about a donation opportunity at the “last place you gave blood” (LP), and the second, donation at a “different place” (DP). This was in order to capture more of the context not included in the attributes (e.g. familiarity of staff).

3.2 Questionnaire development

Six stated preference questions were included in the survey (2 LP and 4 DP questions), an example of which is shown in Figure 1. Taking into account current guidelines and possible changes following from the INTERVAL trial, the maximum number of donations permitted per year differed for males and females. Men and women therefore received different surveys. To streamline the presentation, only the results for women are presented but those for men can be found in the appendix.

3.4 Sample size and efficient design

Sample size calculations for stated preferences surveys are not straightforward and the response rate was uncertain. The full factorial design for men would have involved 96 possible last place (1^1 x 2^2 x 3^1 x 8^1) and 384 different place scenarios (2^2 x 3^1 x 4^1 x8^1). For women 64 possible last place

(1^2 x 2^3 x 8^1) and 256 possible different place scenarios (2^3 x 4^1 x 8^1) would have been required. An efficient design was adopted based on the need to estimate the marginal rate of substitution between attributes with reasonable precision (main effects only). Ngene™ was used to establish an efficient design by considering each section of the survey as one choice set compared to an 'opt out', resulting in 8 LP and 12 DP scenarios for women (12 versions of the survey) and 24 LP and 24 DP scenarios for men (72 versions of the survey). Reflecting the greater number of male scenarios, two men were invited for every woman.

3.5 Survey administration

The survey received favourable ethical approval from both NHS (REC reference 16/YH/0023) and LSHTM ethics committee (reference 10384) in January 2016. 5,016 donors were randomly selected from the PULSE dataset to be invited to participate, according to the following criteria: 17-70 years old; donation of at least one unit of whole-blood in the past 12 months; an email address held by NHSBT; and residence in mainland England. Donors were excluded from the survey if they were temporarily suspended from giving blood (e.g. donors who had recently had a tattoo), had previously stated they did not want to participate in surveys or had received a request to participate in a survey or research from NHSBT (including the INTERVAL trial) in the preceding six months. Due to recent NHSBT communication policy, females with AB+ blood were also excluded. Selected donors were sent an email invitation from NHSBT with a link to the online survey built using FluidSurveys™. Donors who did not complete the survey (excluding those who refused consent) were re-contacted by email 5 days later. The survey closed 7 days after the reminder email. To determine the generalizability of the survey results, the characteristics of the donors who completed the survey were compared to a larger PULSE sample of donors who have donated at least once in the last five years.

4. PULSE: source of revealed preference data

4.1 Analysis of observational data from PULSE

The following classifications of donor characteristics used in PULSE are followed in this study: ethnicity categorized into white, BAME (Black, Asian, Minority Ethnic), and other (mostly undisclosed); blood group categorized into vulnerable (O-, B-, AB-) and not vulnerable as per blood group compatibility. Donors are also classified by their reliability score based on their past behaviour (from 1 – the most reliable – to 5) and the number of donations in the 12 months following 15 March 2015.

For the analysis, an additional classification of blood donation sessions is created. Blood donation sessions can be characterized as constrained or unconstrained. Donor centres are thus unconstrained whereas temporary venues (e.g. a community hall or university session) are constrained. In addition, the service configuration experienced by the individual (henceforth described as the 'baseline service characteristics') can be characterised using PULSE variables. Using planned session opening times, the opening hours experienced by the donors can be mapped to the levels in the survey. Information collected on the duration of a donation is very sparse so the baseline was randomly assigned as 45min or 90min. Estimates of distance and travel time were calculated using the venue postcodes and donor postcode sectors and an algorithm which uses the shortest path on a road map. Table 2 summarizes all these characteristics by sex for the survey respondents and the PULSE population.

This ‘mapping’ of PULSE variables to attributes of the survey thus allows for a comparison between the observed frequency given the baseline characteristics and the predictions from the SP survey for those who responded.

5. SP analysis and comparison to RP

5.1 Analysis of SP survey using the Multinomial Logit

The stated preference survey design results in ordered categorical responses (I would probably not donate, I would donate once per year, two times per year, etc.). However, accepting some loss of information, the most commonly used regression model to analyse choice data – the multinomial logit (MNL) model – is used to analyse the survey results (Jones, 2007, Ryan et al., 2010). This model relies on the assumption that the error terms in the model are independent and identically distributed following the extreme value type I (Gumbel) distribution with mode 0 and variance $\frac{\mu\pi^2}{6}$, $\mu \in \mathbb{R}^+$.

Given our attributes, the model takes the following form:

$$y = constant + \beta_1 * travel\ time + \beta_2 * opening\ hours + \beta_3 * duration + \beta_4 * health\ report + \beta_5 * \max number\ of\ donations + \beta_6 * last\ centre$$

Given individual n and an alternative i in the set of all the possible choices available to that individual, c_n, the probability of choosing alternative i is:

$$P_{in} = \frac{\exp(\mu\boldsymbol{\beta}^T\boldsymbol{X}_{in})}{\sum_{j \in C_n} \exp(\mu\boldsymbol{\beta}^T\boldsymbol{X}_{in})}$$

where $\boldsymbol{\beta}$ is the vector of the regression coefficients and $\boldsymbol{X}$ is the covariate matrix.
For more details, please see Ryan et al. (2010). The mlogit command was used to implement the model in Stata^{TM} (Rabe-Hesketh and Skrondal, 2012). Using post-estimation, the predicted probabilities can be obtained for different service configurations for all stated frequencies. These predicted probabilities, given certain service characteristics, are an important input into the decision model to be developed to evaluate the effect of policy changes on blood volume.

5.2 Comparison of stated and revealed preference

Due to the availability of data to compare both stated preferences and revealed preferences expressed by this sample of donors, concerns of external validity, i.e. hypothetical bias, can be addressed. The emphasis in this analysis was to understand the variability in this discrepancy between what people say they will do and what they actually do.

First, to confirm the presence of any hypothetical bias, the distributions of the probabilities of donations at each frequency were compared by plotting the probability of donating at a given frequency from the observed data next to probability predicted by the MNL model. This comparison shows, at an aggregate level, whether the stated and revealed preference of donors correspond when SP levels are set to modal baseline levels.

Next, the discrepancy between SP and RP for each individual was estimated as follows. The MNL coefficients corresponding to the baseline service characteristics of an individual (as mapped from PULSE to the survey attributes) are used to predict the probability of each feasible donation frequency. The probabilities of one, two, three, four donations per year are rescaled by (1-Pr(I would probably not donate)) because these donors *did* donate in the past year. The expected value of these rescaled predicted probabilities corresponds to the mean predicted number of donations per year for that individual given they continue donation and is compared to the observed number of donations in the past year obtained from PULSE. Thus for a given individual (who completed the survey) actual behaviour given baseline service characteristics can be compared to the stated preference model.

5.3 Characterization of the SP/RP discrepancy

Using these individual discrepancy values, a linear regression was run to evaluate the impact of several *a priori* variables on the size of the discrepancy. Three different regression models were compared: an intercept only model, a model with main effects, and a model with main effects plus one interaction term (blood group * reliability score). The variables included in the model were chosen based on the hypothesis that there may have been constraints on the observed number of donations that are independent of the preference of a donor to give at a certain frequency. For example, we hypothesized that a session constrained with regard to the venue may prevent donors from donating at their desired frequency thus increasing the discrepancy between the observed number and stated preference frequency. In this regard, the following variables were included in the model: constraint, blood group, reliability score, age, ethnicity, and opening hours. The interaction was included because NHSBT modulates its interaction with donors depending on blood group and reliability such as the number of invitations to donate a donor receives.

6. Results

Due to the difference in safe maximum frequencies of donation for men and women, all analyses were split by gender. The results for women are reported below but the analogous results for men can be found in the appendix.

6.1 Survey Response

Just under a quarter of the PULSE population met the eligibility criteria for the survey (23%). A random sample of 5,016 donors who met the eligibility criteria for the survey was invited to complete the survey. 1,254 of invited donors provided consent; a 25% response rate. Response rate was similar for males (24%) and females (27%), but noticeably higher in the older age groups, 45% in 61-70 age group compared to 12% in donors aged 17-30 (see Appendix). Consenting donors had a very high completion rate (see appendix). However, the respondents are not representative of the PULSE population as shown in Table 2. This is not surprising as the eligibility criteria were different and those invitees who completed the survey were a non-random sample of those invited. The subsequent results for the SP survey are for those who responded to the questionnaire.

6.2 Multinomial logit model results

The multinomial logit model estimates a model for each possible frequency of donation. Table 3 provides the results for women (the corresponding table for men is in the appendix). The interpretation of the coefficients is illustrated by the following example. When donor centre opening hours are extended to include evenings (i.e. a change from Mon-Fri 9-12pm & 2-5pm to Mon-Fri 9-12pm & 2-8pm), the risk that a woman donates four times per year was 55% higher than that she donates three times per year holding all other attributes constant. Most coefficients indicate preferences in the expected direction. For example, when travel time increases, there is a higher risk that a woman will decrease her frequency of donation; when opening hours are expanded or a health report introduced, the risk that a woman would stop donating decreases.

Using these MNL model coefficients, predicted probabilities at each frequency of donation and an expected mean value can be calculated for a given set of attribute levels as illustrated in Table 4. The attribute levels have been set according to the service characteristics the donor experienced at their last donation. Individual probabilities have been added and thus an expected mean value for the population of female survey respondents calculated. The table thus states that according to the preferences expressed by these women, they are predicted to donate at a mean frequency of 2.62 times per year.

6.3 Preference validation results

As can be seen in Figure 2, the aggregate distributions of annual donation frequencies differ between the RP and SP data. The distribution of predicted frequencies varies substantially from the observed frequencies: there is an over-prediction of the probability that the donor would stop donating as well as donate three times a year or more.

However, an individual level comparison is more appropriate to characterize the variability in the discrepancy. Table 5 reports the observed and predicted frequency and the discrepancy between them according to different observed characteristics. Overall, there is a mean discrepancy of 0.62 donations which, considering the observed mean is 2 donations per year, is a 31% over-prediction. The table also shows variability in the magnitude of the difference between subgroups. These results however do not all correspond to the a priori hypothesis of situational constraints. For example, the discrepancy for unconstrained donors was expected to be smaller than that for constrained donors. This is contradicted as the discrepancy for unconstrained donors is twice that for constrained donors. With regard to blood group, the trend is again not as expected as vulnerable blood groups are predicted to overestimate 0.1 donations per year on top of the 0.61 donations per year discrepancy reported for not vulnerable blood group donors but this is not a significant difference. Although, the overall trend in the discrepancy with regard to reliability and age does correspond with the a priori hypothesis that more reliable and older donors have smaller discrepancy scores. Very reliable donors' observed and predicted frequencies only differ by 0.49 donations per year on average, whereas the discrepancy of the least reliable is more than three times larger. Similarly, the youngest group of donors have a discrepancy almost three times that of the oldest group (0.90 compared to 0.32 donations per year). There is no clear trend in the discrepancy with regard to opening hours. The discrepancy is larger when evening hours are included without lunch. However, once lunchtime is also included the observed and predicted means are again more congruent.

The next step in characterizing the variability in the discrepancy is a linear regression on the same covariates as reported in Table 5. Three different models were estimated. Model 1, without any covariates has an intercept of 0.62 corresponding to the average difference between individual observed and predicted frequencies. Model 2 contains all the covariates in Table 5 but without any interactions whereas model 3 also contains an interaction between blood group and reliability score. The constant term in models 2 and 3 is the discrepancy at the baseline value of all the covariates in the model rather than the average discrepancy. Thus according to model 2, a female donor who is 41 years old, white, does not have a vulnerable blood type, a reliability score of 4 and a constrained venue with opening hours of Mon-Fri 9am-5pm, will overstate her donation frequency by 1.74 donations per year. The other coefficients in the table indicate how a non-baseline level of the covariate impact this discrepancy: for example, being a more reliable donor significantly decreases the discrepancy by 1.26, 0.64, and 0.76 donations for reliability scores of 1, 2, and 3 respectively, while donating in an unconstrained centre increases the discrepancy by 0.39 donations. An increase in age of one year above the mean of 41 consistently decreases the discrepancy by 0.01 donations.

Table 6 also reports that the models including covariates do explain some of the variation in discrepancy with adjusted R^2 values of 18.98% and 19.08% for model 2 and 3 respectively but the interaction term does not explain much variation as also suggested by the similar MSE values between models 2 and 3. The insignificance of the interaction term implies that NHSBT's donor invitation policy does not have a big impact on the discrepancy between donors' intentions and their behaviour given the other covariates in the model.

7. Discussion

The analysis has fulfilled the aim of this paper to estimate the discrepancy between SP and RP for blood donors who responded to the SP survey. On average, the MNL models estimates that these women over-predict their donation frequency by 31%. This estimated hypothetical bias of 0.62 donations compared to a mean of 2 observed donations per year is lower than expected considering the commonly used calibration factor in the literature of "divide by 2" (List and Gallet, 2001). Especially, since previous studies suggest that hypothetical bias is more pronounced for moral goods due to social desirability bias (Johansson-Stenman and Svedsater, 2012). On the other hand, the fact that this was a within-sample comparison may have diminished the observed hypothetical bias (Hensher, 2010).

The estimated linear regression models were able to explain 19% of the variation in the discrepancy using observed characteristics. In all subgroup stratifications except blood type and type of centre, the largest discrepancy was at least threefold the smallest discrepancy. It is important to consider that the discrepancy will only differ between strata of a subgroup if the frequency is actually determined by convenience and other modifiable characteristics captured by the attributes. If the frequency is independent of these attributes, a discrepancy may still be observed but it would be independent of subgroups. In addition, the possible effect of the nature of blood donation as a non-market good should not be overlooked. However, the fact that donors do express preferences for different attributes included in the survey as shown through the MNL model suggests that at least some part of their donation behaviour is driven by convenience and is thus modifiable.

An interesting observation is that the female donors who currently already have more opportunity to donate – they belong to a vulnerable blood group, donate at an unconstrained centre with longer opening hours – are predicted by the model to overestimate their frequency of donation more than women who currently have less opportunity to donate. In other words, less convenience is associated with more accurate predictions. This is the opposite of what was expected suggesting a possible alternative hypothesis that increased costs lead the donor to think more carefully about their response. This could also mean that there is unobserved confounding by attributes not currently included in the survey and definitely emphasizes the importance of reflecting constraints in the design of the survey. One possible such attribute is appointment availability which will be considered in future surveys. The other design elements, i.e. ex ante techniques, considered in the literature are cheap talk, certainty scales and incentive alignment (Fifer et al., 2014, Champ et al., 2009). In this study, the information provided to the donor before answering the questions already emphasizes the importance of accurate responses so the effect of cheap talk is considered. However, due to the scale of the implications of any change to current NHSBT policy regarding blood service configurations, incentive alignment is very difficult. Moreover, adding complexity to the question through a certainty scale might overburden the respondent.

However, due to the rich data available on these donors in the PULSE database, ex post methods of hypothetical bias mitigation are more applicable to this context. There are two strands of ex post methods in the literature: data fusion and calibration (Hensher et al., 1999, Hensher, 2010, Norwood, 2005, Mark and Swait, 2004). Calibration is most relevant in this context because the aim is to understand the drivers of the discrepancy so that it can be adjusted not only for changes in policy which may have already occurred for some donors but also for novel interventions for which no data yet exists. The calibration factor (or 'bias function' as defined by Blackburn et al. (1994)) could thus be tested through analysis of PULSE data on recently implemented policy changes before extrapolation to new ones. However, this requires a good prediction model for this calibration factor and currently the linear regression models only capture some variation in the discrepancy. It is thus advisable that different modelling approaches are explored such as selection of covariates through machine learning instead of *a priori*. This is definitely possible considering the PULSE dataset contains many more variables that characterise a donation than are currently included.

It is also important to recognize the variation in the literature between the sources of revealed preference data used. Hensher (2010) gives a good overview but in essence emphasizes the difference between revealed preference data from actual observed behaviour (whether experimental or not) and revealed preference data based on surveys of current or recent (possibly non-experienced) alternatives. Currently, this study does not delve into this detail but there is potential to do so in the future because the survey included self-reported donations in the past year. A comparison of the actual behaviour and the stated behaviour thus lends itself to a future investigation of the balance struck by an individual between social desirability bias and honesty when describing past behaviour, i.e. recall bias in the context of moral goods. An individual level estimate of the balance between these two opposing attitudes could thus give an indication as to how this individual keeps a consistent self-image and could refine the calibration of the stated preferences.

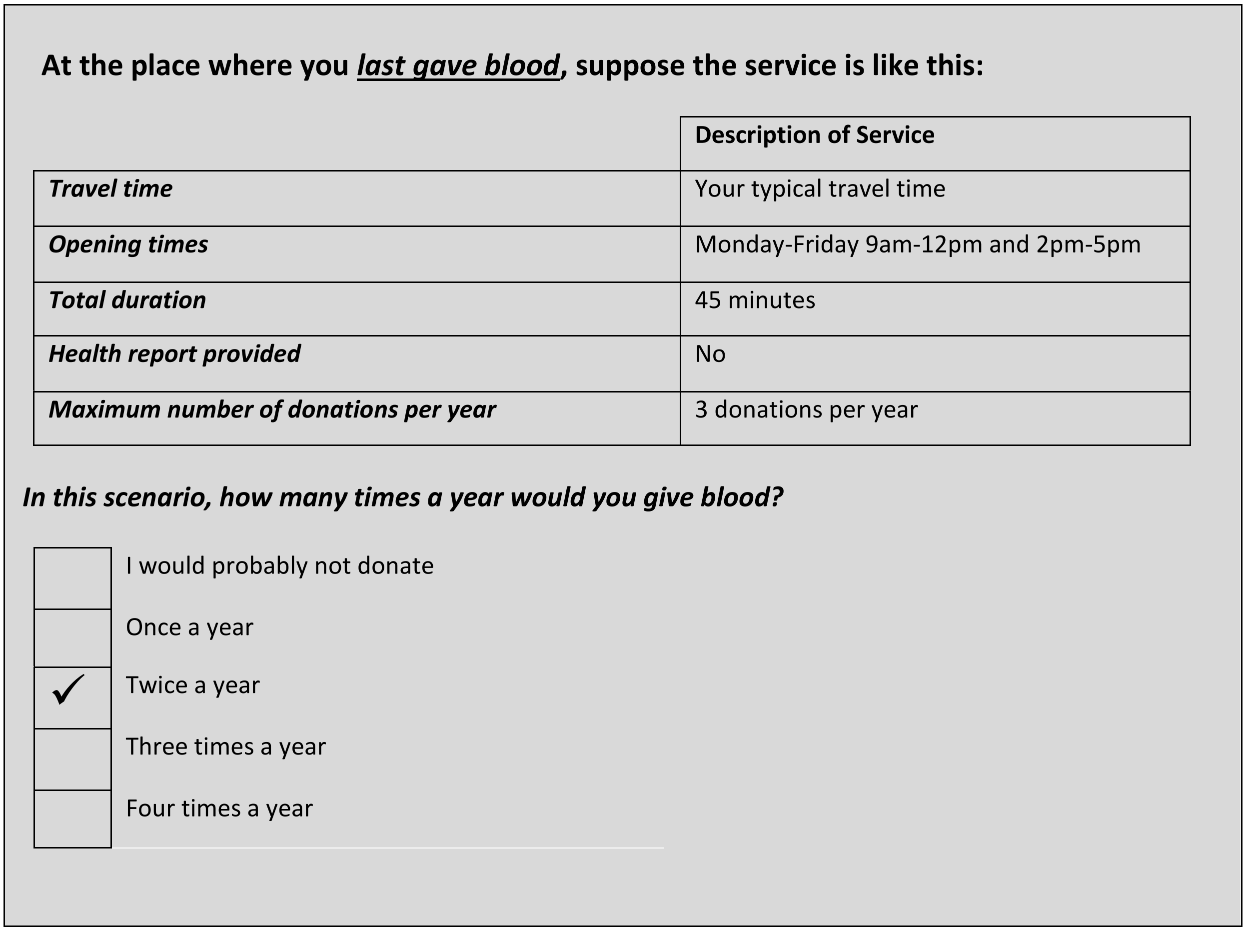
In general, this research advances the application of the hypothetical bias literature within the health economics context which currently uses many valuation techniques but very limited validation techniques (Blumenshein et al., 2001, Ryan et al., 2010). The results in this paper can only underscore the importance of the accuracy of a calibration factor as the heterogeneity of preferences and discrepancy is very evident and therefore invalidates a standard calibration ratio (Norwood, 2005). At least subgroup adjustment is required and preferably reweighting of data to reflect the characteristics of the population of interest considered. In this regard, the analysis performed could be improved by considering the heterogeneity of stated preferences in the MNL model. In the future, MRS could be calculated by subgroup and this heterogeneity investigated. This would make the analysis more congenial throughout as subgroups were considered at the discrepancy stage. This does not invalidate stated preference methods but underscores the distinction between “informativeness” (do stated preference methods provide information that is useful to estimate the real commitment an individual would make?) and “reliability” (how accurately do hypothetical scenarios reflect actual behaviour) as pointed out by Blackburn et al. (1994). A consistent approach to developing robust calibration or bias functions would actually corroborate the worth of stated preference methods.

Another limitation, but which is more fundamental to the process of validation is that the observed frequency is historical. So, the analysis is predicated on the assumption that preferences are stable over time and that there have been no changes in the context of blood donation experienced by the donor other than the factors included in the model. Alternatively, just the PULSE dataset could be used to estimate changes in donation frequency after a service change. Further research is also suggested by the importance of study design on the magnitude of hypothetical bias as described in the literature (List and Gallet, 2001, Hensher, 2010). The effect of the unique SP design where the answer choices take the form of a frequency on hypothetical bias could be contrasted to a conventional DCE design. The predictive value of models estimated from different study designs could then be contrasted and inform future SP data collection.

In conclusion, this paper shows how harnessing a large dataset can inform ex post calibration methods of SP data. Such calibration increases the predictive value of SP data thereby validating its use in decision modelling.

8. Tables and Figures

Figure 1: An example stated preference question



At the place where you *last gave blood*, suppose the service is like this:

	Description of Service
Travel time	Your typical travel time
Opening times	Monday-Friday 9am-12pm and 2pm-5pm
Total duration	45 minutes
Health report provided	No
Maximum number of donations per year	3 donations per year

In this scenario, how many times a year would you give blood?

- ☐ I would probably not donate
- ☐ Once a year
- ☑ Twice a year
- ☐ Three times a year
- ☐ Four times a year

Table 1: Attributes and levels used in the stated preference survey

Attribute	Definition	Levels
1. Donors travel time to blood donation	This is the time it would take you to travel to the place where you donate blood.	10 minutes shorter than your typical travel time
		Your typical travel time*
		15 minutes longer than your typical travel time
		30 minutes longer than your typical travel time
2. Opening times at blood donation venue	These are the days of the week, and times of the day when you can give blood.	Monday – Friday: 9am-12pm and 2pm-5pm
		Monday – Friday: 9am-5pm
		Monday – Friday: 9am-12pm and 2pm-8pm
		Monday – Friday: 9am-8pm
		Monday – Sunday: 9am-12pm and 2pm-5pm
		Monday – Sunday: 9am-5pm
		Monday – Sunday: 9am-12pm and 2pm-8pm
		Monday – Sunday: 9am-8pm
3. Total time to donate blood	This is the overall time it takes to give blood, from when you arrive until you are free to leave.	45 minutes
		90 minutes
4. Availability of health report	A health report is not currently provided. In the future if a health report were to be provided it might give measurements such as your blood pressure and cholesterol.	Yes, provided after each blood donation
		Not provided
5. Maximum no. of blood donations per year	This is the number of times each year that you are allowed to donate blood. A clinical trial is looking at the impact of donors giving blood more often. Depending on the results of the trial, donors might be allowed to give blood more often in future.	3 donations per year (females only)
		4 donations per year
		5 donations per year (males only)
		6 donations per year (males only)

* In all scenarios referring to the last place a donor gave blood, the travel time attribute is represented by one level only, "Your typical travel time"

Table 2: Characteristics of the respondent population and the PULSE population by sex

		Male		**Female**	
Donor Characteristics		**Respondents**	**PULSE**	**Respondents**	**PULSE**
Mean age (sd)		49 (13)	43 (14)	44 (14)	41 (14)
Mean distance from home to place of last donation, in km (range)		9.31 (0.02; 205.5)	12.56 (0, 987.3)	6.59 (0.05; 146.4)	11.07 (0, 1164)
Mean travel time from home to place of last donation, in minutes (range)		13.21 (0.07; 237.4)	16.44 (0, 1008)	10.23 (0.14; 146.1)	14.82 (0, 1179)
Blood Group	Vulnerable[*]	88 (11%)	63,715 (13%)	61 (14%)	93,969 (15%)
	Not vulnerable[†]	709 (89%)	433,829 (87%)	387 (86%)	543,367 (85%)
Ethnicity	White	729 (91%)	433,903 (87%)	414 (92%)	569,262 (89%)
	BAME[‡]	56 (7%)	47244 (10%)	29 (7%)	52,210 (8%)
	Other	12 (2%)	16,397 (3%)	5 (1%)	15,864 (3%)
Reliability Score[§]	1 (most reliable)	685 (86%)	259,657 (52%)	362 (81%)	316,929 (50%)
	2	83 (10%)	83,448 (17%)	60 (13%)	114,730 (18%)
	3	20 (3%)	57,298 (11%)	16 (4%)	78,861 (12%)
	4	9 (1%)	43,903 (9%)	10 (2%)	58,454 (9%)
	5 (least reliable)	0 (0%)	53,238 (11%)	0 (0%)	68,362 (11%)
Number of donations in the last 12 months	0	1 (0.1%)	151,097 (30%)	1 (0.2%)	209,612 (33%)
	1	177 (22%)	123,218 (25%)	149 (33%)	176,541 (28%)
	2	243 (30%)	106,225 (21%)	160 (36%)	144,156 (23%)
	3	319 (40%)	96,399 (19%)	124 (28%)	96,571 (15%)
	4	50 (6%)	19,104 (4%)	14 (3%)	9,957 (2%)
	5	6 (0.8%)	1,302 (0.26%)	0 (0%)	454 (0.07%)
	6 or more	1 (0.1%)	199 (0.04%)	0 (0%)	45 (0.01%)
	Mean (sd) [‖]	2.31 (0.93)	1.43 (1.23)	2 (0.87)	1.25 (1.12)
	Median (IQR) [‖]	2 (2,3)	1 (0,2)	2 (1,3)	1 (0,2)
Centre	Unconstrained[¶]	58 (7.28%)	54,244 (10.90%)	34 (7.59%)	62,034 (9.73%)
	Constrained[¤]	739 (92.72%)	443,300 (89.10%)	414 (92.41%)	575,302 (90.27%)
Opening times of last place of donation	Monday - Friday: 9am-5pm	0	0	0	0
	Monday - Friday: 9am-12pm & 2pm-5pm	29 (3.64%)	26,121 (5.25%)	18 (4.02%)	25,065 (3.93%)
	Monday - Friday: 9am-12pm & 2pm-8pm	59 (7.40%)	30,526 (6.13%)	29 (6.47%)	40,595 (6.37%)
	Monday - Friday: 9am-8pm	574 (72.02%)	348,936 (70.13%)	331 (73.88%)	457,044 (71.71%)
	Monday - Sunday: 9am-12pm & 2pm-5pm	0	0	0	0
	Monday - Sunday: 9am-5pm	35 (4.39%)	20,862 (4.19%)	17 (3.79%)	27,521 (4.32%)
	Monday - Sunday: 9am-12pm & 2pm-8pm	0	126 (0.03%)	1 (0.22%)	184 (0.03%)
	Monday - Sunday: 9am-8pm	99 (12.42%)	69,669 (14%)	52 (11.61%)	85,700 (13.45%)

[*]Vulnerable blood types refer to O-, B- and AB- blood; [†]Not vulnerable' blood groups refer to all blood types other than O-, B- and AB- blood. [‡]Black, Asian and minority ethnic groups; [§]Reliability score for each donor calculated by NHSBT (1 - most reliable; 5 - least reliable); [‖]Observed donations obtained from PULSE; [¶] Constrained centre refers to a temporary centre as the last place a donor gave blood. [¤]Unconstrained centre refers to a fixed centre as the last place a donor gave blood.

Table 3: Results of the multinomial logit model using stated preference responses, for women

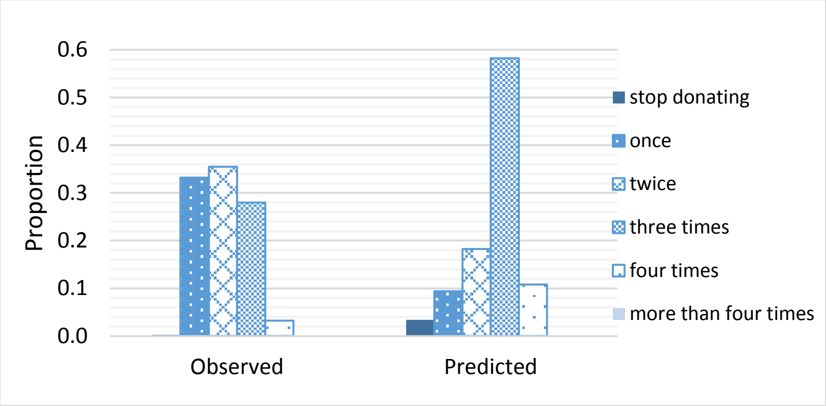
		β (SE)				
Attribute	**Level**	**stop**	**once**	**twice**	**three**	**four**
travel time	-10'	1.31 (.47)	1.04 (.51)	0.93 (.33)	**baseline**	1.46 (.4)
	typical	**baseline**				
	+15'	2.8 (1.38) *	3.21 (1.7) *	2.18 (.76) *		1.22 (.44)
	+30'	9.53 (4.57) ***	6.32 (3.09) ***	5.13 (1.45) ***		0.76 (.25)
opening times	**Mon – Fri: 9am-12pm & 2pm-5pm**	**baseline**				
	Mon – Fri: 9am-5pm	1.32 (.69)	0.61 (.32)	0.95 (.33)		0.97 (.28)
	Mon – Fri: 9am-12pm & 2pm-8pm	0.15 (.09) ***	0.37 (.23)	0.88 (.3)		1.70 (.54) *
	Mon – Fri: 9am-8pm	0.18 (.05) ***	0.67 (.19)	0.82 (.17)		0.98 (.24)
	Mon – Sun: 9am-12pm & 2pm-5pm	0.71 (.4)	0.54 (.28)	0.80 (.29)		0.90 (.23)
	Mon – Sun: 9am-5pm	0.29 (.13) **	0.84 (.39)	0.66 (.17)		1.35 (.32)
	Mon – Sun: 9am-12pm & 2pm-8pm	0.30 (.06) ***	0.79 (.28)	0.63 (.14) *		1.55 (.37) *
	Mon – Sun: 9am-8pm	0.39 (.21) *	0.27 (.17) *	0.39 (.16) *		1.20 (.4)
duration	45'	1.26 (.69)	4.56 (2.23) **	1.62 (.54)		0.85 (.19)
	90'	**baseline**				
health report	provided at each donation	0.65 (.11) **	0.85 (.18)	0.70 (.1) **		0.98 (.12)
	not provided	**baseline**				
maximum frequency	**3 times per year**	**baseline**				
	4 times per year	2.52 (.42) ***	3.06 (.74) ***	3.16 (.55) ***		11.50 (2.16) ***
location	**last place**	**baseline**				
	different place	1.17 (.49)	0.79 (.31)	0.95 (.22)		0.67 (.14) *

** <0.1; ** <0.01; ***< 0.001*

Table 4: Predicted annual probabilities of donating for women

Stop donating	3.39%
Once	9.36%
Twice	18.25%
Three times	58.20%
Four times	10.79%
Expected number of donations per year (SD)	2.63 (0.29)

Figure 2: Histogram of observed and predicted donation frequencies per year, for women



Observed values from PULSE refer to the 12 months prior to the stated preference survey. Predicted values are based on responses to the stated preference survey with attributes set to individual baseline characteristics.

Table 5: Comparison of the observed and predicted donation frequencies for women (predicted frequencies calculated at an individual level)

	Mean (SD)		
	Observed	**Predicted**	**Discrepancy**
Overall	2 (0.87)	2.63 (0.29)	0.62 (0.84)
Blood type			
Vulnerable*	2.05 (0.88)	2.76 (0.28)	0.71 (0.82)
Not vulnerable[†]	2.00 (0.86)	2.61 (0.29)	0.61 (0.84)
Age			
17-30	1.76 (0.86)	2.65 (0.34)	0.90 (0.86)
31-45	1.83 (0.89)	2.57 (0.33)	0.75 (0.82)
46-60	2.11 (0.81)	2.64 (0.27)	0.53 (0.81)
61-70	2.37 (0.81)	2.69 (0.16)	0.32 (0.78)
Ethnicity			
White	2.04 (0.87)	2.64 (0.27)	0.60 (0.85)
BAME[‡]	1.59 (0.73)	2.59 (0.33)	1.00 (0.66)
Other	1.75 (0.96)	1.98 (0.93)	0.23 (0.41)
Reliability score[§]			
1	2.21 (0.82)	2.70 (0.20)	0.49 (0.84)
2	1.2 (0.45)	2.36 (0.34)	1.16 (0.53)
3	1 (0)	2.01 (0.52)	1.01 (0.52)
4	1 (0)	2.83 (0.16)	1.83 (0.16)
Type of centre			
Unconstrained[‖]	1.67 (0.74)	2.82 (0.18)	1.15 (0.68)
Constrained[¶]	2.03 (0.87)	2.62 (0.29)	0.58 (0.84)
Opening hours			
Monday - Friday: 9am-5pm	0	0	0
Monday - Friday: 9am-12pm and 2pm-5pm	1.39 (0.85)	1.64 (0.34)	0.25 (0.80)
Monday - Friday: 9am-12pm and 2pm-8pm	2.10 (0.77)	2.83 (0.13)	0.73 (0.75)
Monday - Friday: 9am-8pm	2.04 (0.87)	2.67 (0.19)	0.62 (0.84)
Monday - Sunday: 9am-12pm and 2pm-5pm	0.00	0.00	0.00
Monday - Sunday: 9am-5pm	2.24 (0.97)	2.66 (0.14)	0.43 (0.99)
Monday - Sunday: 9am-12pm and 2pm-8pm	1 (0)	2.80 (0)	1.80 (0)
Monday - Sunday: 9am-8pm	1.88 (0.79)	2.63 (0.26)	0.75 (0.79)

**Vulnerable blood types refer to O-, B- and AB- blood.*

[†] *Not vulnerable' blood groups refer to all blood types other than O-, B- and AB- blood.*

[‡] *Black, Asian and minority ethnic groups*

[§] *Reliability score for each donor calculated by NHSBT (1- most reliable; 5- least reliable).*
Note: none of the responders to our survey had a reliability score of 5.

[‖] *Unconstrained centre refers to a fixed centre as the last place a donor gave blood.*

[¶] *Constrained centre refers to a temporary centre as the last place a donor gave blood.*

Table 6: Results of three linear regression models on the difference between the mean predicted number of blood donations per year (at the individual level) and the observed number of blood donations per year, for women

	Model 1	**Model 2**	**Model 3**
	β (SE)	**β (SE)**	**β (SE)**
Constant	0.62 (0.04)**	1.74 (0.33)**	1.73 (0.34)**
Constraint (baseline: Constrained[†])			
Unconstrained[‡]		0.39 (0.15)**	0.37 (0.15)**
Blood group (baseline: Not vulnerable [§])			
Vulnerable[‖]		0.18 (0.11)*	-0.06 (0.80)
Reliability factor (baseline: 4)			
1		-1.26 (0.24)**	-1.28 (0.26)**
2		-0.64 (0.26)**	-0.74 (0.28)**
3		-0.76 (0.31)**	-0.80 (0.32)**
Age (baseline: mean)		-0.01 (0.002)**	-0.01 (0.003)**
Ethnicity (baseline: white)			
BAME [¶]		0.14 (0.15)	0.16 (0.15)
Other		-0.31 (0.39)	-0.26 (0.39)
Opening hours (baseline: Monday - Friday: 9am-5pm)			
Monday - Friday: 9am-12pm and 2pm-5pm			
Monday - Friday: 9am-12pm and 2pm-8pm		0.67 (0.23)**	0.68 (0.23)**
Monday - Friday: 9am-8pm		0.48 (0.19)**	0.52 (0.19)**
Monday - Sunday: 9am-12pm and 2pm-5pm			
Monday - Sunday: 9am-5pm		0.40 (0.26)	0.41 (0.26)
Monday - Sunday: 9am-12pm and 2pm-8pm		1.57 (0.78)**	1.59 (0.78)**
Monday - Sunday: 9am-8pm		0.48 (0.21)**	0.51 (0.22)**
Blood group * Reliability score[¤] (baseline: not vulnerable * reliability score 4)			
vulnerable*reliability score 1			0.15 (0.81)
vulnerable*reliability score 2			0.70 (0.84)
vulnerable*reliability score 3			0.58 (1.11)
Number of observations	437	437	437
F	(0, 436) 0	(13,423) 8.86	(16, 420) 7.42
Prob > F	0.0000	0.0000	0.0000
R-squared	0.0000	0.2139	0.2205
Adj R-squared	0.0000	0.1898	0.1908
Root Mean Squared Error	0.8380	0.7543	0.7538

*Donation frequency was predicted according to a different baseline for each donor, using the attribute levels that most closely describe their last blood donation. *p <0.1; ** p<0.01; ***p< 0.001. [†]Constrained centre refers to a temporary centre as the last place a donor gave blood [‡]Unconstrained centre refers to a fixed centre as the last place a donor gave blood. [§] Not vulnerable' blood groups refer to all blood types other than O-, B- and AB- blood. [‖]Vulnerable blood types refer to O-, B- and AB- blood. [¶]Black, Asian and minority ethnic groups. [¤]Reliability score for each donor calculated by NHSBT (1 - most reliable; 5 - least reliable). Note: none of the responders to our survey had a reliability score of 5.*

9. References

BLACKBURN, M., HARRISON, G. W. & RUTSTROM, E. 1994. Statistical bias functions and informative hypothetical surveys. *American Journal of Agricultural Economics,* 76**,** 1084-1088.

BLUMENSHEIN, K., JOHANNESON, M., YOKOYAMA, K. K. & FREEMAN, P. R. 2001. Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. *Journal of Health Economics,* 20**,** 441-457.

CHAMP, P. A., MOORE, R. & BISHOP, R. C. 2009. A comparison of approaches to mitigate hypothetical bias. *Agricultural and Resource Economics Review,* 38**,** 166-180.

DE BEKKER-GROB, E. W., RYAN, M. & GERARD, K. 2012. Discrete choice experiments in health economics: a review of the literature. *Journal of Health Economics,* 21**,** 145-172.

FIFER, S., ROSE, J. & GREAVES, S. 2014. Hypothetical bias in Stated Choice Experiments: Is it a problem? And if so, how do we deal with it? *Transportation Research Part A,* 61**,** 164-177.

GOETTE, L., STUTZER, A., YAVUZCAN, G. & FREY, B. M. 2009. Free cholesterol testing as a motivation device in blood donations: evidence from field experiments. *Transfusion,* 49**,** 524-531.

HENSHER, D., LOUVIERE, J. & SWAIT, J. 1999. Combining sources of preference data. *Journal of Econometrics,* 89**,** 197-221.

HENSHER, D. A. 2010. Hypothetical bias, choice experiments and willingness to pay. *Transportation Research Part B,* 44**,** 735-752.

JOHANSSON-STENMAN, O. & SVEDSATER, H. 2012. Self-image and valuation of moral goods: Stated versus actual willingness to pay. *Journal of Economic Behavior & Organization,* 84**,** 879-891.

JONES, A. 2007. *Applied Econometrics for Health Economists,* Oxford, UK, Radcliffe Publishing Ltd.

KAHNEMAN, D. & KNETSCH, J. L. 1992. Valueing public goods: the purchase of moral satisfaction. *Journal of Environmental Economics and Management,* 22**,** 57-70.

LIST, J. A. & GALLET, C. A. 2001. What experimental protocol influence disparities between actual and hypothetical stated values? Evidence from a meta-analysis. *Environmental and Resource Economics,* 20**,** 241-254.

LOUVIERE, J. 2006. What you don't know might hurt you: some unresolved issues in the design and analysis fo discrete choice experiments. *In:* ADAMOWICZ, W. & DESHAZO, J. R. (eds.) *Special issue: frontiers on stated choice methods.* Environmental and Resource Economics.

MARK, T. L. & SWAIT, J. 2004. Using stated preference and revealed preference modeling to evaluate prescribing decisions. *Health Economics,* 13**,** 563-573.

MEWS, M. 2013. To pay or not to pay - the role of monetary incentives in an optimal blood doatnion service bundle. *International Journal of Nonprofit and Voluntary Sector Marketing,* 18**,** 192-202.

MORTIMER, D., GHIJBEN, P., HARRIS, A. & HOLLINGSWORTH, B. 2013. Incentive-based and non-incentive-based interventions for increasing blood donation. *The Cochrane Library*.

MURPHY, J. J., ALLEN, P. G., STEVENS, T. H. & WEATHERHEAD, D. 2005. A meta-analysis of hypothetical bias in stated preference valuation. *Environmental and Resource Economics,* 30**,** 313-325.

NORWOOD, F. B. 2005. Can calibration reconcile stated and observed preferences? *Journal of Agricultural and Applied Economics,* 37**,** 237-248.

RABE-HESKETH, S. & SKRONDAL, A. 2012. *Multilevel and Longitudinal Modeling Using Stata Volume II: Categorical Responses, Counts, and Survival,* College Station, Texas, State Press.

RAPPORT, F. L. & MAGGS, C. J. 2002. Titmuss and the gift relationship: altruism revisited. *Journal of Advanced Nursing,* 40**,** 495-503.

RINGWALD, J. 2010. Established ways to keep donor's interest alive. *ISBT Science Series,* 5.

RYAN, M. & GERARD, K. 2003. Using discrete choice experiments to value health care programmes: current practice and future research reflections. *Applied Health Economics and Health Policy,* 2**,** 55-64.

RYAN, M., GERARD, K. & AMAYA-AMAYAK, M. 2010. *Using Discrete Choice Experiments to Value Health and Health Care,* Dordrecht, The Netherlands, Springer.

SCHREIBER, G. B., SCHLUMPF, K. S., GLYNN, S. A., WRIGHT, D. J., TU, Y., KING, M. R., HIGGINS, M. J., KESSLER, D., GILCHER, R., NASS, C. C., GUILTINAN, A. M. & STUDY, N. H. L. B. I. R. E. D. 2006. Convenience, the bane of our existence, and other barriers to donating. *Transfusion,* 46**,** 545-553.

STREET, D. J., BUNCH, D. S. & MOORE, B. J. 2001. Optimal designs for 2k paired comparison experiments. *Communications in statistics, theory and methods,* 30**,** 2149-2171.

TAYLOR, S. E. & BROWN, J. E. 1994. Positive illusions and well-being revisited: separating fact from fiction. *Psychological Science,* 10**,** 80-83.

TITMUSS, R. M. 1970. *The gift relationship: from human blood to social policy,* London, Allen & Unwin.

VINEY, R., LANCSAR, E. & LOUVIERE, J. 2002. Discrete choice experiments to measure consumer preferences for health and health care. *Expert Review of Pharmacoeconomics & Outcomes Research,* 2**,** 319-326.

Appendix

Table 1: Proportion of responders who completed each part of the survey

	All responders (Donors who gave consent = 1254)		Men (N = 804)		Women (N = 450)	
SP last place Q1	1234	98.4%	790	98.3%	444	98.7%
SP last place Q2	1230	98.1%	787	97.9%	443	98.4%
SP different place Q1	1223	97.5%	783	97.4%	440	97.8%
SP different place Q2	1217	97.0%	779	96.9%	438	97.3%
SP different place Q3	1215	96.9%	778	96.8%	437	97.1%
SP different place Q4	1211	96.6%	775	96.4%	436	96.9%

Table 2: Response rates to the stated preference survey, N (%)

Response Rates		**N (%)**	**p-value** [†]
Overall response rate*		1254 (25%)	
Sex	Male	797 (24%)	0.03
	Female	448 (27%)	
Age	17-30	179 (12%)	<0.001
	31-40	171 (19%)	
	41-50	294 (27%)	
	51-60	337 (34%)	
	61-70	264 (45%)	
Blood Group	Vulnerable[‡]	149 (24%)	0.42
	Not vulnerable[§]	1096 (25%)	
Ethnicity	White	1143 (26%)	<0.001
	BAME[‖]	85 (18%)	
	Other	17 (19%)	
Reliability Score[¶]	1 (most reliable)	1047 (30%)	<0.001
	2	143 (15%)	
	3	36 (10%)	
	4	19 (18%)	
	5 (least reliable)	0 (0%)	
Number of donations in the last 12 months	0	2 (8%)	<0.001
	1	326 (16%)	
	2	403 (26%)	
	3	443 (39%)	
	4	64 (35%)	
	5	6 (50%)	
	6 or more	1 (100%)	

** Defined as the proportion of respondent population over the proportion of invited population*

[†] *The p-value for the chi-squared test on the proportions between groups*

[‡] *Vulnerable blood types refer to O-, B- and AB- blood*

[§] *Blood types other than O-, B- and AB- are considered 'not vulnerable'*

[‖] *Black, Asian and minority ethnic groups*

[¶] *Reliability score for each donor calculated by NHSBT (1- most reliable; 5- least reliable)*

Table 3: Results of multinomial logit model using stated preference responses, for men.

		β (SE)						
Attribute	**Level**	**stop**	**once**	**twice**	**three**	**four**	**five**	**six**
travel time	-10'	0.47 (.09) ***	0.58 (.17) *	0.93 (.15)	1.17 (.17)	**baseline**	1.02 (.24)	1.95 (.39) ***
	typical	**baseline**						
	+15'	1.71 (.34) **	2.01 (.63) *	1.67 (.29) **	1.61 (.27) **		1.10 (.24)	1.62 (.34) *
	+30'	4.73 (.85) ***	3.23 (.88) ***	2.42 (.46) ***	1.72 (.33) **		0.48 (.13) **	1.12 (.27)
opening times	**Mon – Fri: 9am-12pm & 2pm-5pm**	**baseline**						
	Mon – Fri: 9am-5pm	0.87 (.14)	1.44 (.44)	0.90 (.18)	0.88 (.16)		0.69 (.18)	0.93 (.21)
	Mon – Fri: 9am-12pm & 2pm-8pm	0.33 (.07) ***	0.90 (.33)	0.83 (.19)	0.82 (.15)		1.56 (.42) *	1.00 (.19)
	Mon – Fri: 9am-8pm	0.28 (.06) ***	0.76 (.25)	0.78 (.16)	0.73 (.13) *		1.36 (.34)	1.24 (.24)
	Mon – Sun: 9am-12pm & 2pm-5pm	0.45 (.08) ***	0.68 (.23)	0.73 (.14) *	0.70 (.12) *		1.37 (.39)	0.92 (.18)
	Mon – Sun: 9am-5pm	0.52 (.12) **	1.24 (.47)	1.33 (.32)	1.31 (.27)		1.47 (.4)	1.37 (.3)
	Mon – Sun: 9am-12pm & 2pm-8pm	0.16 (.04) ***	0.24 (.1) ***	0.61 (.12) *	0.63 (.11) *		0.95 (.26)	1.04 (.21)
	Mon – Sun: 9am-8pm	0.23 (.04) ***	0.27 (.1) ***	0.61 (.12) *	0.75 (.13) *		1.75 (.41) *	1.19 (.22)
duration	45'	0.43 (.05) ***	0.34 (.06) ***	0.52 (.06) ***	0.80 (.08) *		0.89 (.14)	1.10 (.13)
	90'	**baseline**						
health report	provided at each donation	0.68 (.07) ***	0.52 (.1) ***	0.73 (.08) **	0.79 (.08) *		1.04 (.18)	0.87 (.09)
	not provided	**baseline**						
maximum frequency	**4 times per year**	**baseline**						
	5 times per year	1.89 (.31) ***	2.70 (.69) ***	2.33 (.38) ***	2.68 (.44) ***		72.78 (23.18) ***	1.84 (.46) *
	6 times per year	2.11 (.31) ***	2.06 (.52) **	2.09 (.33) ***	2.51 (.38) ***		9.22 (2.78) ***	12.83 (2.45) ***
location	**last place**	**baseline**						
	different place	1.89 (.28) ***	1.95 (.55) *	1.32 (.2) *	0.85 (.11)		0.94 (.15)	0.55 (.09) ***

**** <0.1; ** <0.01; ***< 0.001.***

Table 4: Comparison of the observed and predicted donation frequencies for men (predicted frequencies calculated at an individual level)

	Mean (SD)		
	Observed	**Predicted**	**Discrepancy**
Overall	2.31 (0.93)	3.25 (0.35)	0.93 (0.94)
Blood type			
Vulnerable*	2.62 (0.96)	3.24 (0.36)	0.62 (0.95)
Not vulnerable†	2.28 (0.92)	3.25 (0.35)	0.97 (0.93)
Age			
17-30	1.90 (0.89)	3.26 (0.39)	1.36 (0.92)
31-45	2.17 (0.89)	3.19 (0.40)	1.02 (0.88)
46-60	2.37 (0.91)	3.26 (0.32)	0.88 (0.93)
61-70	2.53 (0.94)	3.28 (0.32)	0.75 (0.96)
Ethnicity			
White	2.34 (0.93)	3.25 (0.35)	0.91 (0.95)
BAME‡	1.91 (0.85)	3.21 (0.36)	1.31 (0.85)
Other	2.92 (0.67)	3.37 (0.23)	0.45 (0.73)
Reliability score§			
1	2.49 (0.87)	3.28 (0.33)	0.79 (0.91)
2	1.34 (0.52)	3.21 (0.35)	1.86 (0.63)
3	1.26 (0.45)	3.09 (0.39)	1.83 (0.60)
4	1 (0)	2.25 (0.36)	1.24 (0.34)
Type of centre			
Unconstrained‖	2.25 (1.03)	3.48 (0.25)	1.22 (1.03)
Constrained¶	2.32 (0.92)	3.23 (0.35)	0.91 (0.93)
Opening hours			
Monday - Friday: 9am-5pm	0.00	0.00	0.00
Monday - Friday: 9am-12pm and 2pm-5pm	2 (0.80)	2.42 (0.38)	0.42 (0.88)
Monday - Friday: 9am-12pm and 2pm-8pm	2.12 (0.87)	3.10 (0.30)	0.98 (0.88)
Monday - Friday: 9am-8pm	2.35 (0.93)	3.30 (0.28)	0.95 (0.94)
Monday - Sunday: 9am-12pm and 2pm-5pm	0.00	0.00	0.00
Monday - Sunday: 9am-5pm	2.23 (0.84)	2.88 (0.35)	0.65 (0.76)
Monday - Sunday: 9am-12pm and 2pm-8pm	0.00	0.00	0.00
Monday - Sunday: 9am-8pm	2.39 (1.02)	3.46 (0.22)	1.07 (1.04)

**Vulnerable blood types refer to O-, B- and AB- blood.*

[†] *Not vulnerable' blood groups refer to all blood types other than O-, B- and AB- blood.*

[‡] *Black, Asian and minority ethnic groups*

[§] *Reliability score for each donor calculated by NHSBT (1- most reliable; 5- least reliable). Note: none of the responders to our survey had a reliability score of 5.*

[‖] *Unconstrained centre refers to a fixed centre as the last place a donor gave blood.*

[¶] *Constrained centre refers to a temporary centre as the last place a donor gave blood.*

Table 5: Results of linear regression on the difference between the mean predicted number of blood donations per year (at the individual level) and the observed number of blood donations per year, for men

	Model 1	Model 2	Model 3
	β (SE)	**β (SE)**	**β (SE)**
Constant	0.93 (0.03)**	1.13 (0.34)**	1.05 (0.37)**
Constraint (baseline = constrained)			
Unconstrained		0.01 (0.13)	0.01 (0.13)
Blood group (baseline = not vulnerable)			
vulnerable		-0.22 (0.10)**	0.15 (0.68)
Reliability factor (baseline = 4)			
1		-0.43 (0.29)	-0.34 (0.32)
2		0.57 (0.30)*	0.63 (0.33)*
3		0.57 (0.34)*	0.65 (0.38)*
Age (baseline = mean)		-0.01 (0.002)**	-0.01 (0.002)**
Ethnicity (baseline = white)			
BAME		0.25 (0.13)**	0.26 (0.13)**
Other		-0.38 (0.25)	-0.37 (0.25)
Opening hours (baseline = Monday - Friday: 9am-5pm)			
Monday - Friday: 9am-12pm and 2pm-5pm		empty	empty
Monday - Friday: 9am-12pm and 2pm-8pm		0.62 (0.19)**	0.62 (0.19)**
Monday - Friday: 9am-8pm		0.60 (0.16)**	0.60 (0.16])**
Monday - Sunday: 9am-12pm and 2pm-5pm		empty	empty
Monday - Sunday: 9am-5pm		0.28 (0.21)	0.26 (0.22)
Monday - Sunday: 9am-12pm and 2pm-8pm		empty	empty
Monday - Sunday: 9am-8pm		0.66 (0.18)**	0.66 (0.19)**
Blood group * Reliability factor			
(baseline not vulnerable * reliability 4)			
vulnerable*reliability 1			-0.39 (0.69)
vulnerable*reliability 2			0.18 (0.84)
vulnerable*reliability 3			empty
Number of obs	773	773	773
F	(0, 772) 0	(13,759) 15.21	(15,757) 13.28
Prob > F	0	0	0
R-squared	0	0.2067	0.2083
Adj R-squared	0	0.1931	0.1926
Root Mean Squared Error	0.94	0.84687	0.84714

*Donation frequency was predicted according to a different baseline for each donor, using the attribute levels that most closely describe their last blood donation. *p <0.1; ** p<0.01; ***p< 0.001. [†]Constrained centre refers to a temporary centre as the last place a donor gave blood [‡]Unconstrained centre refers to a fixed centre as the last place a donor gave blood. [§] Not vulnerable' blood groups refer to all blood types other than O-, B- and AB- blood. [‖]Vulnerable blood types refer to O-, B- and AB- blood. [¶]Black, Asian and minority ethnic groups. [¤]Reliability score for each donor calculated by NHSBT (1 - most reliable; 5 - least reliable). Note: none of the responders to our survey had a reliability score of 5.*