Choice certainty and deliberative thinking in discrete choice experiments. A theoretical and empirical investigation

Dean A. Regier[a], Jonathan Sicsic[b,c]*, Verity Watson[d 1]

[a] Canadian Centre for Applied Research in Cancer Control, BC Cancer
School of Population and Public Health
University of British Columbia
675 West 10th Avenue, Vancouver, British Columbia V5Z 1L3, Canada

[b] LIRAES (EA 4470), University Paris Descartes, Sorbonne Paris Cité
Centre Universitaire des Saints-Pères, 45 rue des Saints-Pères, 75006 Paris, France
[c] CESP, Univ. Paris-Sud, UVSQ, INSERM, Univ. Paris-Saclay, 94805, Villejuif, France

[d] Health Economics Research Unit
University of Aberdeen
Institute of Applied Health Sciences
Aberdeen, AB25 2ZD, United Kingdom

[1] The Canadian Centre for Applied Research in Cancer Control (ARCC) is funded by the Canadian Cancer Society Research Institute (2015-703549). This paper developed from discussions between Verity Watson and Dean Regier that were funded by the Peter Wall Institute of Advanced Studies, University of British Columbia. Jonathan Sicsic acknowledges funding from the People Programme (Marie Curie Actions) of the European Union's Seventh Framework Programme (FP7/2007-2013) under REA grant agreement PCOFUND-GA-2013-609102, through the PRESTIGE programme coordinated by Campus France. He also benefited for this research from grants provided by the French National Institute for Cancer (Coordinator: Dr Nora Moumjid). The Health Economics Research Unit is funded by the Chief Scientist Office of the Scottish Government Health and Care Directorates. The usual disclaimer applies. We thank Aki Tsuchiya, Nicolas Krucien, Thijs Dekker, and all participants to the 5th workshop on non-market valuation for useful comments on previous drafts of the paper.

ABSTRACT

Resource allocation decisions require information about individuals' preferences for goods and services. Survey based stated preference methods, such as discrete choice experiments (DCEs), are used to elicit preferences for non-market goods. A critique of stated preference research is that respondents to hypothetical surveys may not provide careful and thoughtful responses that reveal rational preferences. Choice certainty has been used to measure survey respondents' task engagement. Researchers assume that respondents who are certain about their choices provide deliberative responses. In the case of DCE, we argue that the variability of choice certainty is also important. We present a novel framework to identify thoughtful / deliberative respondents. The framework combines respondents' certainty with their variability in certainty across a set of choice tasks. We test our framework empirically using data from two case studies. We find respondents with higher mean certainty and variability (i) seldom use decision heuristics, (ii) are more likely to have monotonic preferences, (iii) have longer response times, (iv) make choices that have higher interval validity, and (v) have higher choice consistency. We discuss the relevance of alternative ex-post calibration strategies with a view to improve the precision and accuracy of DCE-based welfare estimates.

Key words: Choice certainty; Discrete choice experiments; Hypothetical bias; Information processing; Stated preferences; Survey engagement

JEL codes: C35; D80; I12

1. Introduction

Public decision making is often concerned with the provision of non-market goods, such as environmental amenities or health care services. These resource allocation decisions require information about individuals' preferences for the non-market good. Yet, in such settings there exists little or no market data from which to infer preferences. Survey based stated preference methods are used to elicit preferences for non-market goods (Boxall et al., 1996; Carson et al., 2001). A popular stated preference method is the discrete choice experiment (DCE). DCEs describe the non-market good or service being valued by a set of attributes. The attributes are arranged into multi-attribute bundles, which are presented to individuals in choice sets of two or more bundles. The DCE task requires individuals to choose their preferred bundle. A criticism of stated preference research is that responses may differ from people's real choices - this is called the hypothetical bias problem (Blumenschein et al., 2008, 2001; Morrison and Brown, 2009; Murphy and Stevens, 2004). A related critique is that choice tasks are difficult to complete and respondents may not engage with the task. In this context, individuals do not provide the deliberative responses necessary for rational decision-making (Loomis, 2011; Luchini and Watson, 2014).

Survey respondents' task engagement has been measured by their choice certainty (Beck et al., 2013; Brouwer et al., 2010; Dekker et al., 2016; Lundhede et al., 2009; Olsen et al., 2011). Researchers assume that respondents who state they are certain about their decision have rational preferences and provide more reliable responses. As such, these respondents may be less subject to hypothetical bias and their choices are more consistent (Beck et al., 2016; Fifer et al., 2014; Ready et al., 2010). In this paper, we question this assumption and argue that respondents who are *always* certain about their choice in DCE tasks are more likely to make quick and intuitive decisions without much thought. Respondents who are engaged in the task and provide thoughtful responses will not always be certain about their decisions and their choice certainty will depend on the choices they face (Olsen et al., 2011; Sudman et al., 1996). We present a novel framework to identify engaged respondents that combines respondents' choice certainty with the *variability* in respondents' choice certainty across a set of choice tasks. We test our framework empirically. We use respondents' mean choice certainty and variability to separate respondents into two groups: quick (non-deliberative) processing and rational (deliberative) processing. Then, we test various conjectures that relate choice certainty patterns to the quality of DCE responses.

The remainder of the paper is structured as follows: Section 2 provides an overview of choice certainty in stated preference studies. Section 3 presents the theoretical link between choice certainty and deliberative thinking. Section 4 discusses our two case studies and provides details of our data. Our testable hypotheses on the link between deliberative thinking, choice certainty and certainty variability and empirical results are presented in Section 5. Section 6 discusses the results and present future research avenues.

2. Use of choice certainty in the stated preference literature

Researchers have long recognised that when individuals are asked their monetary value of non-market goods in a contingent valuation method (CVM) survey they are likely to face uncertainty. This uncertainty arises because individuals are either unfamiliar with the good in question or unfamiliar with assessing their willingness to pay (WTP) or willingness to accept (WTA) compensation for a change in the good (Li and Mattsson, 1995; Wang, 1997). Consequently, researchers have asked respondents to state their degree of certainty about their WTP or to state the range within which their WTP lies[2]. One implicit assumption of this approach is that respondents know the range in which their true WTP lies. Therefore, respondents are certain they will or will not pay amounts below or above the range, but are uncertain about whether or not they are willing to pay amounts within the range[3]. A further assumption is that respondents who are most certain about their WTP provide more reliable responses. This assumption has been verified in studies that compare CVM-derived valuations with real payments and find a closer correspondence between valuations and payments when valuations are based only on the responses of the most certain respondents (Murphy et al., 2005).

Researchers have also used respondents' choice certainty to calibrate CVM responses. This is done in two ways: 1) by recoding responses or 2) by re-weighting responses in empirical analysis based on choice certainty. In general, response recoding changes the data for

[2] Researchers have measured contingent valuation choice certainty in two ways: 1/ simultaneously with the CV responses or 2/ after respondents' WTP has been elicited (post-choice certainty). The choice certainty question frame has become standardised, but researchers use a range of different response options including descriptive certainty scales (definitely sure, probably sure, not sure) and numerical certainty scales (ranging from 1 to 5, or from 1 to 10). Numerical certainty scales can be polar-point labelled scales in which the end-points (1 and 10) are given a descriptive label such as 1=very uncertain to 10=very certain, but the intermediate points are not.

[3] Loomis and Eksland 1998 and others find evidence of such a `u-shaped' relationship between the monetary amount and respondent certainty (Loomis and Ekstrand, 1998).

respondents who are uncertain about their response: respondents who are uncertain and state ‘yes, I am willing to pay $x’ are re-coded as if they stated ‘no, I am not willing to pay $x’(Blumenschein et al., 2008). Studies that reweight the data place a greater empirical weight on respondents who are certain of their answer (Li and Mattsson, 1995; Martinez-Espineira and Lyssenko, 2012).

The findings from CVM studies have led researchers to elicit choice certainty alongside preferences in DCE tasks (Beck et al., 2016, 2013; Brouwer et al., 2010; Dekker et al., 2016; Fifer et al., 2014; Lundhede et al., 2009; Olsen et al., 2011; Regier et al., 2014). In these DCE studies, researchers usually assume that certain respondents are more engaged and provide more thoughtful responses. Ready et al (2010), Fifer et al (2014) and Beck et al (2016) investigate the relationship between certainty and hypothetical bias in DCE tasks. They hypothesise that more certain responses are less prone to hypothetical bias. The studies find mixed results. Ready et al (2010) recode responses that fail to meet a certainty threshold of 7 (out of 10). They find close correspondents between hypothetical and real WTP after recoding. Fifer et al (2014) aggregate certainty across all the DCE tasks a respondent completes and calculate the respondents’ median certainty. They find that hypothetical WTP is higher than real WTP and that certain respondents (median certainty≥8) have lower hypothetical bias. Beck et al (2016) use choice certainty and compare the effect on hypothetical bias of either recoding or reweighting with a certainty threshold of 9 (out of 10). They find that recoding increases hypothetical bias and reweighting has no impact on hypothetical WTP and therefore hypothetical bias. Beck et al (2013) also compare the effect of different recoding and reweighting approaches on WTP, but do not have comparable data on real valuations. They find recoding and reweighting substantially reduces estimated WTP.

Researchers have also tested the relationship between certainty and respondent engagement in DCE tasks by investigating the link between response certainty and consistency (Beck et al., 2013; Dekker et al., 2016; Lundhede et al., 2009). These studies include choice task certainty as an explanatory variable of error variance in econometric models of respondents’ DCE choices. All studies find the higher choice task certainty is associated with higher choice consistency (i.e., with lower error variance). Researchers have also explored the determinants of choice task certainty, and in particular how choice task certainty is affected by the utility balance of the alternatives in the choice task (Brouwer et al, 2010; Olsen et al, 2011; Regier et al, 2014; Lundhede et al, 2009). All studies find that choice task certainty increases with the utility difference across alternatives in the task. Olsen et al (2011) interpret this as evidence that response certainty measures preference certainty as proposed by Li and Mattsson (1995) and Wang (1997).

3. Choice certainty and deliberative thinking: theoretical background

Most researchers assume the relationship that holds between choice certainty and response quality in CVM tasks also holds for DCE tasks. But there is currently no theoretical framework that relates choice certainty to more reliable choices (Loomis, 2011). There are important differences between the CVM and DCE methods that affect the link between choice certainty and reliability (Olsen et al, 2011; Beck et al, 2013; Regier et al., 2014). CVM tasks ask respondents to report if their WTP for a good is above a stated monetary amount. In a CVM context, choice certainty contains information about whether the monetary amount lies within the range of respondents' WTP. DCE tasks ask respondents to choose one multi-attribute good from a set of goods in which the monetary cost is one attribute. This task requires the respondent to assess for each good if their monetary value (WTP or WTA) for the described bundle is higher or lower than the monetary cost. In the case that respondents are willing to pay for both goods, they then have to assess which good provides the highest utility.

In a DCE, choice certainty contains information about whether the respondents is able to distinguish between the utility of the two or more alternatives (Olsen et al, 2011; Lundhede et al., 2009). Some choice sets will include alternatives that provide very different utilities. In this situation, the alternatives are easy to distinguish, and therefore respondents are likely to be certain about their choice. While in other choice sets all the alternatives may have a similar utility. In this situation, the alternatives are hard to distinguish, and respondents are likely to be uncertain about their choice (Olsen et al, 2011; Regier et al., 2014). We should expect respondents’ certainty to vary across tasks. Respondents who are engaged with the task and who provide reliable responses will in some choice sets be certain of their choice (when one alternative dominates all others) and in other choice sets be uncertain about their choice. Regier et al (2014) find that respondents who are uncertain about their decision reveal more precise information about their preferences in the choices that they make.

The conceptual link between response certainty, task complexity, and imprecise preference data is supported in the judgement and decision-making literature (Kahneman, 2003; Simon, 1990, 1979; Stanovich and West, 2000). Stanovich and West (2000) and Kahneman (2003) suggest a dual-process theory of decisional thinking. Respondents who make intuitive choices without much thought are using System 1 processing, which is *associated with errors of intuition*. These errors include heuristics of accessibility, which is the amount of effort with which thoughts come to mind. Task complexity may induce System 1 thinking, which can act as a computational escape hatch for respondents who find the tasks too difficult to complete rationally. Our assertion is that System 1 processing may lead to less informative and noisier choices.

Simon (1990) suggests that choices are shaped by the task environment and individuals' computational capabilities. The task environment includes the complexity of the task (Simon, 1979) while computational capability accounts for an 'information-processing cognitive architecture'. This architecture includes short- and long-term memory and a 'production system' capable of problem-solving and learning from new information (Simon, 1990). Kahneman (2003) similarly states: "the accessibility of a thought is determined jointly by the characteristics of the cognitive mechanism that produce it and by the characteristics of the stimuli and events that evoke it" (p.699). As such, the choice task that creates the stimuli can work together with System 1 or System 2 type processing (or the cognitive mechanism or processing architecture) to influence the ease of a thought or judgment. These assertions leave scope for task complexity differently affecting respondents employing System 1 or System 2 type processing when completing choice tasks. Respondents who engage in System 1 processing may be less deliberative in their responses (a result of complexity), less likely to experience task complexity and less likely to be uncertain about their choices. System 2 thinkers who use all of the information in the choice set experience task complexity. As a consequence, their level of certainty should vary during the choice experiment as a result of varying task complexity (Swait and Adamowicz, 2001).

In this paper, we propose that in addition to response certainty, what matters to identify thoughtful responses is certainty *variability* across choice tasks. Previous studies have considered response certainty at the choice task level and have ignored potentially useful information contained in the *variability* of certainty across the tasks completed by the same individual. We argue that respondents' certainty variability across choices contains useful information about respondents' task engagement. Respondent's choice certainty should differ across the different choice sets if they are involved in the task, i.e. if they engage in deliberative rather than intuitive thinking. This assumption is supported by previous literature showing that choice task difficulty varies because of varying choice complexity, which subsequently affects stated choice certainty (Olsen et al., 2011; Regier et al., 2014; Swait and Adamowicz, 2001). As a consequence of varying difficulty, picking the alternative with highest perceived utility is more cognitively demanding in more complex choice tasks (Louviere et al., 2008) and respondents' stated certainty decreases (Regier et al., 2014). In line with these results, we posit that respondents who are *always* certain or uncertain of their choices across the experiment are actually putting less deliberative effort into answering the choice questions and thus provide lower quality data. In the next section we present two case studies used to test our assumptions.

4. The two case studies

Our empirical analysis is based on two case studies that use a DCE to elicit individuals' preferences for non-market goods and also collect post choice certainty after each choice task. Table 1 summarizes the characteristics of each study.

4.1. Women's preferences for breast cancer screening

This study used a DCE to elicit women's preferences and trade-offs between the benefits and risks of breast cancer screening. The data were collected in 2016 using an online survey of 812 women between 40 and 74 years of age and living in France (Sicsic et al., 2018). Breast cancer screening was described by seven attributes: 1) breast cancer mortality risk (10, 15, 20, 25, 30 deaths out of 1,000 women followed during 25 years); 2) false-positive mammography risk (0, 50, 100, 150, 200 false-positive results out of 1,000 women tested); 3) over-diagnosis risk (0, 10, 50, 100, 150 over-diagnosed cases out of 1,000 women tested); 4) type of screening referral (the local screening centre or the doctor); 5) travel time (0 minutes, 10 minutes, 30 minutes, 60 minutes, 90 minutes); 6) total number of tests over a lifetime (0, 6, 12, 18, 24), and 7) out of pocket cost (€0, €30, €60, €60 refunded later). A main effects D-efficient design was used to generate the choice tasks. In each choice task, the women had to choose between two screening alternatives and an opt-out option in which the levels were always the same. An example choice task is provided in Appendix A. The survey included 16 choice tasks, which were blocked to two questionnaire versions each with eight choice tasks. Respondents were randomly allocated to one questionnaire version and the order of choice tasks was randomly varied across respondents. One additional choice task was added to the questionnaire. This task was designed to test the monotonicity of preferences (i.e. the ability to choose a dominant alternative within a specific choice task). In the choice task testing for monotonicity, the dominant alternative included better levels for the first three attributes (i.e. lower breast cancer mortality, lower false-positive and over diagnosis risk) and equal levels of the four other attributes compared to the dominated screening alternative.

For each choice task, response certainty was elicited using a 10 polar point labelled Likert scale (0= "very uncertain", 10 = "very certain") and response time was measured automatically within the online survey.

4.2. Preferences for the return of incidental genomic findings

This study used a DCE to elicit citizen's preferences for the return of incidental genomic findings (Regier et al, 2015). The data were collected in 2014 using an online survey of 1200 Canadian citizens over 18 years of age. The DCE included five attributes: 1) risk of developing the disease(s) at some point in the future (5% lifetime risk or higher, 40% lifetime risk or higher, 80% lifetime risk or higher, 90% lifetime risk or higher); 2) disease treatability (no effective medical treatment or lifestyle change, recommended effective lifestyle change only, recommended effective medical treatment only, recommended effective medical treatment and lifestyle change); 3) disease severity (mild, moderate, severe, or very severe health consequences); 4) information on carrier status (yes, no); and 5) cost of receiving the results ($425, $750, $1000, $1500). Questionnaires were in French or English. The experimental design used D-efficient procedures to maximize the statistical efficiency of the statistical model (Regier et al., 2015). The approach resulted in 80 choice tasks. Each subject was randomly assigned to one of five blocks that included 16 choice tasks. An example choice task is in Appendix A.

Each choice task included two testing options and a "no information" (opt-out) option. For each choice task, certainty level on a 10 point Likert scale (0= "very uncertain", 10 = "very certain") were collected as well as total response time for the questionnaire.

5. Linking mean choice certainty and variability to survey response quality

5.1. Taxonomy of choice certainty and set of testable assumptions

In this section, we propose a certainty classification and a set of testable assumptions about the relationship between choice certainty and *variability* and deliberative thinking. We partition respondents into three certainty classes based on their mean choice certainty (see Table 2). Our thresholds for categorising certain respondents are consistent with groupings used in the literature (Blumenschein et al., 2008; Li and Mattsson, 1995; Loomis and Ekstrand, 1998). First, respondents with an average choice certainty strictly above 8 (on the 0 to 10 certainty scale) are considered to have offered relatively *certain* answers. Second, respondents with an average choice certainty below 6 are considered to have offered relatively *uncertain* answers. Third, respondents with a mean certainty between 6 and 8 (included) are considered *hesitant*.

We further partition respondents within these classes into two smaller groups based on observed certainty variability, as measured by the standard deviation in certainty over the sequence of choices: the "*constantly*-" prefixed group denotes respondents with strictly lower certainty variability compared to the average of their class, and the "*variably*-" prefixed group denotes respondents with higher certainty variability compared to the average of their class.

Table 3 presents descriptive statistics of our six-group certainty classification in both case studies. The *variably uncertain* (VU) group represents the smallest proportion of the sample (i.e., 9.6% in study 1 and 7.8% in study 2), and the *constantly hesitant* (CH) group has largest sample size (22.0% and 24.2%, respectively). Table 3 further describes the choice certainty distributions in the two case studies. As expected, the standard deviation, range, and 95% intervals of the certainty scores are systematically higher in the subgroups having higher certainty variability compared to the groups with lower certainty variability.

This partition reflects our assumption that both mean certainty and the variability in certainty contain useful information about respondents' engagement in the choice tasks, i.e. their use of deliberative versus intuitive thinking. Deliberative thinking cannot be directly inferred from the data, therefore we proxy it by various indicators of data quality. We posit that respondents using deliberative thinking provide higher data quality, that is, they provide more accurate and precise information about their preferences. Following previous research (e.g., Beck et al., 2013, 2016; Brouwer et al., 2010; Lundhede et al., 2009), we first assume that individuals using deliberative thinking should be engaged enough in the task to be sufficiently certain of their choices, thus the *uncertain* (CU and VU groups) are more likely to provide low (or moderate[4]) quality data (**H1**). Furthermore we assume that only individuals with sufficient certainty variability complete the DCE tasks with a high level of deliberative thinking (**H2**). This assumption directly follows on from the theoretical framework presented in section 3. Therefore the *variably hesitant* (VH) and *variability certain* (VC) groups are assumed to have provided the highest quality data. On the contrary, we assume that individuals who are always certain of their choices (e.g., the *constantly certain*) are less likely to have engaged in the task and are thus likely to provide lower quality data.

We test the relevance of our certainty classification based on five conjectures that connect respondents' choice behaviour in the DCE to data quality indicators that are indicative of deliberative thinking.

Conjecture 1: *The constantly uncertain (CU), constantly hesitant (CH) and constantly certain (CC) groups are more likely to adopt simplifying choice heuristics such as always selecting the same alternative across the choice tasks (e.g., the status quo)*[5].

> ***Rationale***. According to assumption H2, low certainty variability is associated with fast intuitive thinking, i.e. higher likelihood of adopting simplifying choice heuristics, such as deterministic choice patterns.

Conjecture 2: *The variably hesitant (VH) and variably certain (VC) groups use more rational decision behaviour and therefore are more likely to hold monotonic preferences*.

> ***Rationale***. A monotonic preference, which is defined as the ability to adopt utility maximizing rules, is an indicator of rational decision-making. According to assumption H1 and H2, only respondents with a sufficient level of certainty and certainty variability are assumed to engage in rational (deliberative) thinking.

---

[4] We recognize that our assumption is less clear for the *variably uncertain (VU)* group because individuals in this group are on average uncertain but at times offer relatively more certain responses.

[5] The process of always choosing the status quo option (i.e., serial non-participation) may be explained by different factors including (but not limited to) protest responses or lexicographical preferences (Haefen et al., 2005). We assume that people who always make the same decision whatever the content of the choice set are more likely to adopt simplifying choice heuristics. This is consistent with previous theoretical and empirical analyses (Dekker et al., 2016; Loomes et al., 2009; Swait and Adamowicz, 2001). While non-demanding behaviour is consistent with consumer theory, always choosing the status quo (or any other alternative) in repeated choice experiments can lead to unidentified preference parameters in standard random utility theory (RUT) models. RUT is not meant to deal with completely deterministic choices.

Conjecture 3: *The variably hesitant (VH) and variably certain (VC) groups offer more considered choices and therefore have longer response times*.

Rationale. By definition, deliberative thinking is associated with slow decision making. In a DCE context, one needs sufficient time to compute each alternative's utility and then to pick the alternative with highest perceived utility. Previous studies support the assumption of a link between response time and cognitive effort in the context of online surveys (Borger, 2016; Uggeldahl et al., 2016). Moreover, using eye-tracking methods, it was shown that more time is needed to process more complicated attributes (Krucien et al., 2017).

Conjecture 4: *The choices of the variably hesitant (VH) and variably certain (VC) groups have higher internal validity (i.e., logical consistency)*.

Rationale. Internal validity reflects choice behaviours in line with a priori assumptions, where individuals are expected to behave rationally. According to assumption H1 and H2, only respondent with a sufficient level of certainty and certainty variability are assumed to engage in rational (deliberative) thinking.

Conjecture 5: *In econometric models of choices, the variably hesitant and variably certain groups have a higher scale (i.e. lower error variance) as an indicator of higher choice consistency*.

Rationale. In discrete choice models, scale is related to choice consistency: higher scale is associated with lower error variance and is an indicator of higher choice consistency (DeShazo and Fermo, 2002; Hensher et al., 1998). Individuals who make more consistent choices are more likely to engage in deliberative thinking.

5.2. Empirical testing of behavioural assumptions

5.2.1 Choice certainty and use of simplifying heuristics (conjecture 1)

We measure respondents use of simplifying choice heuristics when completing the DCE tasks using two indicators: 1) *serial non-trading behaviour* in which individuals always select the same "testing" option (i.e., option A or option B) across the choice tasks and 2) *serial non-demanding behaviour* in which individuals always choose the "no testing" (i.e., opt-out) option across the choice tasks.

The association between the certainty groups and the use of decision heuristics is presented in Table 4. In case study 1, the results of Fishers' exact test show that there is a significant relationship between certainty groups and respondents being either a serial non-trader ($p<0.0001$) or serial non-demander ($p<0.0001$). We find that serial non-trading behaviour is more prevalent among the *constantly uncertain (CU)* respondents (10.2% against 3.3% overall), and serial non-demanding behaviour is more prevalent among *constantly certain (CC)* respondents (17.5% against 6.2% overall). *Constantly certain* respondents account for 58% of all serial non-demanders in the sample. These results provide empirical support for conjecture 1.

In case study 2, only 4 respondents (0.3%) are serial non-traders thus this indicator is not considered in the empirical analysis (statistical tests not computed). We find that *constantly certain (CC)* respondents are significantly more likely to be serial non demanders (51.1% against 18.3% overall, $p<0.0001$). These results provide partial empirical support for conjecture 1.

5.2.2. Choice certainty and monotonicity of preferences (conjecture 2)

In case study 1, we test the monotonicity of preferences using an additional choice task that included a dominant alternative with 'better' levels for the first three attributes (i.e. lower breast cancer mortality, lower false-positive and over diagnosis risk) compared to the alternative screening option. The results of the monotonicity test stratified by certainty groups are shown in Table 5 (case study 1 only). The respondents with higher certainty variability are more likely to have monotonic preferences than respondents with lower certainty variability ($p<0.0001$). This is particularly true for *variably hesitant (VH) and variably certain (VC)* respondents: in these two groups, respectively 90.3% and 94.7% pass the test, versus 83.0% of respondents overall. This result provides empirical support for conjecture 2[6].

5.2.3. Choice certainty and response time (conjecture 3)

[6] Note that the strong prevalence of serial non-demanders among the *constantly certain* (CC) group may explain why this group performs poorly in the monotonicity test, as serial non-demanders systematically fail to choose the dominant screening alternative in the monotonicity test.

Two indicators are used to analyse the association between choice certainty patterns and response time: 1) median response time to the questionnaire[7] and 2) the proportions of "speeders" (or "professional respondents") defined as respondents who rushed through the questionnaire without sufficiently considering the information provided (Borger, 2016). For both studies, a completion time of no less than 15-20 minutes was expected based on information provided from the pilot surveys. Therefore, a total response time strictly below 10 minutes was used as a cut-off to identify the so-called "speeders", who represented 16.8% (study 1) and 17.8% (study 2) of the sample.

The results are presented in Table 5. In both studies, we find the median response time per choice task is significantly higher for the groups with high certainty variability compared to the groups with lower certainty variability ($p<0.0001$). For instance the *variably certain (VC)* respondents have a median response time of 15.8 minutes (study 1) and 17.2 minutes (study 2), compared with 13.9 minutes (study 1) and 13.8 minutes (study 2), for the *constantly certain (CC)* group. In both studies, the proportion of "speeders" is higher among the subgroups with both low mean certainty and low certainty variability. For instance, in the *constantly uncertain (CU) group,* 27.4% (study 1) and 23.0% (study 2) have response times below 10 minutes, compared with 9.2% (study 1) and 11.8% (study 2) in the *variably certain (VC) group*. These results provide strong empirical support for conjecture 3.

5.2.4. Choice certainty and internal validity (conjecture 4)

We assess internal validity of the choices made by respondents in each certainty group by estimating i) stratified multinomial logit (MNL) models and ii) stratified mixed multinomial logit (MXL) models to describe the choice of each of the six certainty groups. In the MNL model, preferences are assumed homogeneous and thus all parameters of the utility function are fixed. This assumption is relaxed by fitting MXL models, assuming normally distributed random coefficients for all attributes. Moreover, the MXL includes a random generic intercept ($ASC_{optionA|B}$) as follows:

$$U_{njt} = \alpha_n ASC_{optionA|B} + \beta_n' x_{jt} + \epsilon_{njt} \quad (1)$$

[7] We use response time to the entire questionnaire because this was the only information available in both case studies. Median response time is used instead of the mean because it attenuates the impact of extreme response times that could be negatively associated with data quality. Indeed, in a similar study context (i.e. an online survey) Borger (2016) finds a non-linear relationship between response time and choice randomness.

As $ASC_{optionA|B}$ enters the utility function of the two testing (i.e. non opt-out) alternatives, it introduces correlation across these alternatives thus relaxing the independence from irrelevant alternatives (IIA) assumption (Train, 2003). Second, it accounts for serial non-demanding behaviours such as individuals who are always choosing to opt-out ($ASC_{optionA|B} \ll 0$). Thus, the results of the MXL may be more robust to possible confounding between constant reporting of high certainty and serial non-demanding. More generally, results of the MXL models may be more robust to potential confounding between choice certainty and preference heterogeneity[8].

The results of the MNL and MXL models are used to compute the proportion of estimated preference parameters that are in line with a-priori hypotheses[9]. In case study 1, all the quantitative attributes (i.e., mortality, false-positive risk, overdiagnosis risk, travel time, number of tests) are expected to have a negative impact on the utility derived from the screening programme. We have no *a priori* expectation about the impact of doctors' prescription as compared to receiving an invitation letter from the local screening centre, therefore, this parameter is not considered in the internal validity assessment. In total, there are six attributes (corresponding to eight preference parameters, because the cost attribute is categorical) for which we have *a priori* assumptions with respect to the sign of the coefficient. Our *a priori* expectations are that women will prefer a screening service with lower risks of 1) mortality, 2) false-positive results, 3) over-diagnosis, with 4) a shorter travel time, 5) lower number of lifetime tests, and 6) lower out-of-pocket cost.

In case study 2, there are five attributes (four of which are included as categorical variables), corresponding to a total of ten preference parameters for which we have *a priori* assumptions with respect to the sign of the coefficient. Our a-priori expectations are that individuals will prefer a test when: 1) they have a higher probability of developing a disease in future, 2) there are effective treatments for the disease, 3) the disease is moderately severe or severe, 4) they receive information on carrier status and 5) the cost is lower.

Table 6 reports the number and percentage of coefficients that are in line with a-priori assumptions in the two case studies. The results of the stratified MNL and MXL models on

[8] As noted by an anonymous reviewer, if a respondent has strong feelings about some alternatives or their features, she might find it easy to make the hypothetical choices and always report high certainty. In other words, there is possible confounding between choice certainty (and variability) and preference heterogeneity. Estimation of MXL models allows better accounting for this issue. If confounding exists, we expect the association between choice certainty patterns and internal validity to be less strong in MXL compared to MNL models.

[9] In both datasets, all the qualitative attributes are effects coded.

which these tests are based are provided in Appendix B (Table B1 to B4). In case study 1, in each certainty class, the internal validity is systematically higher for individuals with higher certainty variability. The choices of the *variably hesitant (VH)* and *variably certain (VC)* groups have the highest internal validity (100% of preference parameters have the expected sign in MNL models) whereas the choices of the *constantly uncertain (CU)* and *constantly certain (CC)* groups have the lowest internal validity (respectively 50% and 63% of preference parameters have the expected sign in the MNL models). In the MXL models, the internal validity is more similar across the six certainty classes. The results are consistent with those from the MNL and show our results are robust to potential confounding between choice certainty and preference heterogeneity: the choices of the *variably hesitant (VH)* and *variably certain (VC)* groups have 100% internal validity, versus 50% for the *constantly uncertain (CU)* and 88% for the *constantly certain (CC)* groups. Overall, our results provide empirical support for conjecture 4.

In case study 2, the six classes display less heterogeneity in terms of internal validity as compared to case study 1. The choices of the *constantly uncertain (CU)* group exhibit lowest internal validity (70% in the MNL model and 80% in the MXL model). The choices of the *variably hesitant (VH)* and *variably certain (VC)* groups have the highest internal validity (respectively 100% and 90% in the MNL model, and 100% in the MXL model). In comparison, the choices of the *constantly certain (CC)* exhibit lower internal validity in the MNL model (80%), but similar consistency in the MXL model (100%). Overall, these results provide less strong but still consistent evidence supporting conjecture 4.

5.2.5. Choice certainty and choice consistency (conjecture 5)

We estimate heteroskedastic (or scaled) multinomial logit models (HMNL/SMNL) to investigate the relationship between choice certainty and choice consistency. The HMNL allows for unequal variances across individuals (DeShazo and Fermo, 2002; Hensher et al., 1998). The utility function is specified as follows:

$$U_{njt} = \mu_n \beta' x_{jt} + \epsilon_{njt} \quad (2)$$

In equation (2), U_{njt} is the utility that respondent *n* derives from alternative *j* in choice situation *t*, x_{jt} is the vector of attributes levels for alternative *j* in choice situation *t*, β' is the vector of preference weights to be estimated, ϵ_{njt} is an error term assumed *iid* Gumbel, and

μ_n is a function of individual characteristics assumed to influence the magnitude of the scale parameter. This scale function is parameterised as $\mu_n = \exp(Z_n\gamma)$ where Z_n is a vector of individual characteristics and γ is a vector of parameters reflecting the influence of those characteristics on the scale (Hole, 2006). As the scale is inversely related to the variance of the error term, any characteristic having a positive (respectively, negative) impact on μ_n can be interpreted as associated to higher (respectively, lower) choice consistency (DeShazo and Fermo, 2002; Brouwer et al., 2010).

We estimate five HMNL models using the following individual characteristics as predictors of the error variance function:

- M1) Choice certainty at the task level (Dekker et al., 2016; Lundhede et al., 2009);
- M2) Total response time (Borger, 2016; Uggeldahl et al., 2016);
- M3) Five effect-coded indicators of certainty group membership (with the *constantly hesitant* group omitted for model identification);
- M4) Mean certainty and certainty variability (both mean-centred);
- M5) Mean certainty and certainty variability (mean-centred) plus an interaction term between mean certainty and certainty variability.

Models 1 and 2 are standard in the literature and thus are used as benchmarks. Model 3 directly aims at testing our conjecture 5. Models 4 and 5 aim to test additional assumptions. In M4, we test whether there is a linear relationship between choice consistency and mean certainty and certainty variability. In M5, we test the assumption that the relationship between choice certainty and deliberative thinking is moderated by the level of certainty variability. This model can be viewed as more flexible than M3, as it includes continuous (rather than categorical) effects. We expect a positive and significant effect of the interaction term on scale.

Table 7 reports the results of the HMNL models (M1 to M5) estimated in case study 1. M1 and M2 show that choice certainty and response time do not have a statistically significant impact on scale[10], and the model fit (log-likelihood) is unchanged after their inclusion. M3 shows that indicators based on our proposed certainty group partition have a statistically significant effect on scale and the model goodness of fit improves. The *constantly uncertain* and *constantly certain* groups have lower scale (higher response errors), i.e. lower choice

[10] The results remain unchanged when we include response time at the choice task level or the log of response time in order to attenuate the effect of extreme response times.

consistency. Conversely, the *variably hesitant (VH)* and *variably certain (VC)* groups have significantly higher scale indicating higher choice consistency. These results provide empirical support for conjecture 5.

In M4, only certainty variability has a positive and significant impact on scale ($p<0.0001$). In M5, as expected, there is a positive and significant interaction between mean certainty and certainty variability. This result validates the assumption that certainty variability moderates in the relationship between choice certainty and choice consistency. M5 has the highest goodness of fit, suggesting that a flexible, continuous representation of deliberative thinking could be better than the discrete representation as assumed in M3 using the six-class partition.

The results of the HMNL models estimated in case study 2 are reported in Table 8. In line with previous study results (Borger, 2016; Uggeldahl et al., 2016), response time has a positive and significant impact on scale. In line with case study 1, choice certainty (at the task level) does not have a statistically significant impact on scale. In M3, the *constantly certain (CC)* group has lower scale whereas the *variably hesitant (VH)* and *variably certain (VC)* groups have significantly higher scale. These results provide empirical support for conjecture 5. In M4, higher mean certainty is negatively associated with choice consistency. However, the effect disappears when interacting mean certainty with certainty variability in M5. In line with case study 1, we find a positive and significant interaction between mean certainty and certainty variability, thus showing the importance of accounting for choice certainty variability[11].

[11] Note that estimation of models accounting for both scale and preference heterogeneity can be achieved within the Generalized multinomial logit (GMNL) modelling framework. This model is controversial because scale and preference heterogeneity are confounded in discrete choice models as shown in Hess and Rose (2012). For this reason, we cannot allow for preference heterogeneity while estimating determinants of scale heterogeneity, because it would raise identification issues.

6. Discussion and extensions

In this paper, we proposed a novel framework to identify engaged, deliberate respondents in discrete choice experiments. The framework combined respondents' choice certainty with the *variability* in respondents' choice certainty across a set of choice tasks. We tested behavioural assumptions derived from the theoretical framework using two DCE datasets collected in different contexts. We found respondents with higher certainty variability seldom used decision heuristics, were more likely to have monotonic preferences, and had longer response times. We also found that econometric models of these respondent's choices had higher internal validity and lower error variance.

Information on choice certainty variability may be useful to improve the precision and accuracy of DCE-based welfare estimates. One way to do this is within an ex-post calibration framework similar to those that use choice certainty. The *scaling approach* has been the most used empirically (Beck et al., 2013; Brouwer et al., 2010; Fifer et al., 2014; Lundhede et al., 2009). Researchers estimate a scaled (or heteroskedastic) multinomial logit models (HMNL) and include the choice certainty as parameter of the scale function. The aim is to down-weight uncertain responses (Brouwer et al., 2010; Lundhede et al., 2009). We show in section 5.2.5. that the HMNL model can be extended to incorporate both mean choice certainty and certainty variability. However, if preference heterogeneity and choice certainty and certainty variability are confounded then down weighting responses with low variability may bias the results. This approach requires econometric advances that allow discrete choice models to separately identify scale and preference heterogeneity (Hess and Rose, 2012).

An alternative approach is to directly re-weight respondents in the likelihood function (Regier et al, 2014). The aim is the same as the *scaling approach*: to increase efficiency by down-weighing observations with high variance. In Appendix C, we show how this approach can be applied with certainty variability. We estimate weighted error components logit models and account for some preference heterogeneity in the alternative specific constants. We find reweighting in favour of respondents with higher mean certainty decreased the precision and plausibility of the welfare estimates. We find that including certainty variability in the re-weighting function improved the results. Further research is needed to investigate how this approach can be extended to incorporate preference heterogeneity for all attributes. For instance, if a mixed logit framework is used several questions remain: i) should parameters be estimated in preference or WTP-space? ii) which distributions should be chosen for random parameters (e.g., normal, lognormal, triangular)? iii) which type of estimation methods - simulated maximum likelihood or Bayesian Markov Chain Monte Carlo (MCMC) ? and iv) which statistics should be reported for accuracy (mean or median WTP) and precision (standard errors or standard deviations) of welfare estimates?

When designing DCE studies, researchers aim to present respondents with choice sets that elicit the maximum amount of information about respondents' preferences. Researchers often aim to maximise the overall statistical efficiency by balancing utility of the alternatives presented in a choice task using prior preference information estimated from a pilot study (Greiner et al., 2014; Sándor and Wedel, 2001). However this approach increases task complexity for respondents (Swait and Adamowicz, 2001; Viney et al., 2005) and decreases response certainty (Regier et al., 2014). Our results suggest if researchers then reweight data toward *certain* respondents they may lose some of the efficiency gain provided by their experimental design.

While reweighting places more emphasis on respondents engaged in deliberative thinking, it may compromise the representativeness of the results by down-weighting some respondents. Many studies include rationality tests and other measures of data quality and researchers may remove individuals who fail these tests from the analysis. Reweighting is similar but all individuals are retained. In both cases, data on actual choices are needed to allow researchers to test whether focussing on the responses of more engaged respondents increases the accuracy if welfare estimates.

We encourage future research to investigate the determinants of choice certainty variability and in particular, whether and how it is related to the experimental design (e.g., inclusion of a status quo or opt-out option), to the complexity of the choice tasks (number of attributes, number of choice options) and to the use of particular decision heuristics. For instance, we found that one heuristic (serial non-trading) was frequently used by "constantly uncertain" individuals while the other (serial non-demanding) was frequently used by "constantly certain" respondents. Intuitively, we may assume that "*constantly **uncertain***" respondents have either i) not well-formed preferences or ii) cognitive difficulties in answering the choice questions. This is consistent with the interpretation of choice uncertainty in other studies (see e.g., Dekker et al, 2016). Therefore, these respondents may not want to spend much time answering the questions (we find that they have lower response times) thus explaining they systematically choose the same option. On the contrary, we may assume that "*constantly*

certain" respondents have well-constructed preferences, i.e. strong feelings about some alternatives or their features, and find it easy to make the hypothetical choices such as systematically rejecting the cancer screening or genetic testing alternative[12].

Conclusion

We showed that higher certainty is associated with higher deliberative thinking **only for** individuals who vary in their certainty during completion of the choice experiment. In other words, certainty variability is a key variable to consider for optimal ex-post calibration of respondents' choices. We suggest applied researchers should use a certainty index in their reweighting function **only if** it is interacted with certainty variability. Further research is needed on the specification of econometric models incorporating information on choice certainty and variability and how it improves the precision and accuracy of welfare estimates in various contexts (e.g. valuation of health, environmental, or transportation amenities).

[12] We thank an anonymous reviewer for suggesting this interpretation.

References

Beck, M.J., Fifer, S., Rose, J.M., 2016. Can you ever be certain? Reducing hypothetical bias in stated choice experiments via respondent reported choice certainty. Transp. Res. Part B Methodol. 89, 149–167. https://doi.org/10.1016/j.trb.2016.04.004

Beck, M.J., Rose, J.M., Hensher, D.A., 2013. Consistently inconsistent: The role of certainty, acceptability and scale in choice. Transp. Res. Part E Logist. Transp. Rev. 56, 81–93. https://doi.org/10.1016/j.tre.2013.05.001

Blumenschein, K., Blomquist, G.C., Johannesson, M., Horn, N., Freeman, P., 2008. Eliciting Willingness to Pay Without Bias: Evidence from a Field Experiment*. Econ. J. 118, 114–137. https://doi.org/10.1111/j.1468-0297.2007.02106.x

Blumenschein, K., Johannesson, M., Yokoyama, K.K., Freeman, P.R., 2001. Hypothetical versus real willingness to pay in the health care sector: results from a field experiment. J. Health Econ. 20, 441–457.

Borger, T., 2016. Are Fast Responses More Random? Testing the Effect of Response Time on Scale in an Online Choice Experiment. Environ. Resour. Econ. 65, 389–413. https://doi.org/10.1007/s10640-015-9905-1

Boxall, P.C., Adamowicz, W.L., Swait, J., Williams, M., Louviere, J., 1996. A comparison of stated preference methods for environmental valuation. Ecol. Econ. 18, 243–253. https://doi.org/10.1016/0921-8009(96)00039-0

Brouwer, R., Dekker, T., Rolfe, J., Windle, J., 2010. Choice Certainty and Consistency in Repeated Choice Experiments. Environ. Resour. Econ. 46, 93–109. https://doi.org/10.1007/s10640-009-9337-x

Carson, R.T., Flores, N.E., Meade, N.F., 2001. Contingent Valuation: Controversies and Evidence. Environ. Resour. Econ. 19, 173–210. https://doi.org/10.1023/A:1011128332243

Dekker, T., Hess, S., Brouwer, R., Hofkes, M., 2016. Decision uncertainty in multi-attribute stated preference studies. Resour. Energy Econ. 43, 57–73. https://doi.org/10.1016/j.reseneeco.2015.11.002

DeShazo, J.R., Fermo, G., 2002. Designing Choice Sets for Stated Preference Methods: The

Effects of Complexity on Choice Consistency. J. Environ. Econ. Manag. 44, 123–143.

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? Transp. Res. Part Policy Pract. 61, 164–177. https://doi.org/10.1016/j.tra.2013.12.010

Greiner, R., Bliemer, M., Ballweg, J., 2014. Design considerations of a choice experiment to estimate likely participation by north Australian pastoralists in contractual biodiversity conservation. J. Choice Model. 10, 34–45. https://doi.org/10.1016/j.jocm.2014.01.002

Haefen, V., H, R., Massey, D.M., Adamowicz, W.L., 2005. Serial Nonparticipation in Repeated Discrete Choice Models. Am. J. Agric. Econ. 87, 1061–1076. https://doi.org/10.1111/j.1467-8276.2005.00794.x

Hensher, D., Louviere, J., Swait, J., 1998. Combining sources of preference data. J. Econom. 89, 197–221. https://doi.org/10.1016/S0304-4076(98)00061-X

Hess, S., Rose, J.M., 2012. Can scale and coefficient heterogeneity be separated in random coefficients models? Transportation 39, 1225–1239. https://doi.org/10.1007/s11116-012-9394-9

Hole, A.R., 2006. Small-sample properties of tests for heteroscedasticity in the conditional logit model. Econ. Bull. 3, 1–14.

Kahneman, D., 2003. A perspective on judgment and choice: mapping bounded rationality. Am. Psychol. 58, 697–720. https://doi.org/10.1037/0003-066X.58.9.697

Krucien, N., Ryan, M., Hermens, F., 2017. Visual attention in multi-attributes choices: What can eye-tracking tell us? J. Econ. Behav. Organ. 135, 251–267. https://doi.org/10.1016/j.jebo.2017.01.018

Li, C.-Z., Mattsson, L., 1995. Discrete Choice under Preference Uncertainty: An Improved Structural Model for Contingent Valuation. J. Environ. Econ. Manag. 28, 256–269. https://doi.org/10.1006/jeem.1995.1017

Loomes, G., Orr, S., Sugden, R., 2009. Taste uncertainty and status quo effects in consumer choice. J. Risk Uncertain. 39, 113–135. https://doi.org/10.1007/s11166-009-9076-y

Loomis, J., 2011. What’s to Know About Hypothetical Bias in Stated Preference Valuation

Studies? J. Econ. Surv. 25, 363–370. https://doi.org/10.1111/j.1467-6419.2010.00675.x

Loomis, J., Ekstrand, E., 1998. Alternative approaches for incorporating respondent uncertainty when estimating willingness to pay: the case of the Mexican spotted owl. Ecol. Econ. 27, 29–41. https://doi.org/10.1016/S0921-8009(97)00126-2

Louviere, J.J., Islam, T., Wasi, N., Street, D., Burgess, L., article., J.D. served as editor and T.E. served as associate editor for this, 2008. Designing Discrete Choice Experiments: Do Optimal Designs Come at a Price? J. Consum. Res. 35, 360–375. https://doi.org/10.1086/586913

Luchini, S., Watson, V., 2014. Are choice experiments reliable? Evidence from the lab. Econ. Lett. 124, 9–13. https://doi.org/10.1016/j.econlet.2014.04.005

Lundhede, T.H., Olsen, S.B., Jacobsen, J.B., Thorsen, B.J., 2009. Handling respondent uncertainty in Choice Experiments: Evaluating recoding approaches against explicit modelling of uncertainty. J. Choice Model. 2, 118–147. https://doi.org/10.1016/S1755-5345(13)70007-1

Martinez-Espineira, R., Lyssenko, N., 2012. Alternative approaches to dealing with respondent uncertainty in contingent valuation: A comparative analysis. J. Environ. Manage. 93, 130–139. https://doi.org/10.1016/j.jenvman.2011.09.008

Morrison, M., Brown, T.C., 2009. Testing the Effectiveness of Certainty Scales, Cheap Talk, and Dissonance-Minimization in Reducing Hypothetical Bias in Contingent Valuation Studies. Environ. Resour. Econ. 44, 307–326. https://doi.org/10.1007/s10640-009-9287-3

Murphy, J.J., Allen, P.G., Stevens, T.H., Weatherhead, D., 2005. A Meta-analysis of Hypothetical Bias in Stated Preference Valuation. Environ. Resour. Econ. 30, 313–325. https://doi.org/10.1007/s10640-004-3332-z

Murphy, J.J., Stevens, T.H., 2004. Contingent Valuation, Hypothetical Bias, and Experimental Economics. Agric. Resour. Econ. Rev. 33, 182–192. https://doi.org/10.1017/S1068280500005761

Olsen, S.B., Lundhede, T.H., Jacobsen, J.B., Thorsen, B.J., 2011. Tough and Easy Choices: Testing the Influence of Utility Difference on Stated Certainty-in-Choice in Choice Experiments. Environ. Resour. Econ. 49, 491–510. https://doi.org/10.1007/s10640-010-9443-

Ready, R.C., Champ, P.A., Lawton, J.L., 2010. Using Respondent Uncertainty to Mitigate Hypothetical Bias in a Stated Choice Experiment. Land Econ. 86, 363–381.

Regier, D.A., Peacock, S.J., Pataky, R., van der Hoek, K., Jarvik, G.P., Hoch, J., Veenstra, D., 2015. Societal preferences for the return of incidental findings from clinical genomic sequencing: a discrete-choice experiment. CMAJ Can. Med. Assoc. J. J. Assoc. Medicale Can. 187, E190-197. https://doi.org/10.1503/cmaj.140697

Regier, D.A., Watson, V., Burnett, H., Ungar, W.J., 2014. Task complexity and response certainty in discrete choice experiments: An application to drug treatments for juvenile idiopathic arthritis. J. Behav. Exp. Econ. 50, 40–49. https://doi.org/10.1016/j.socec.2014.02.009

Sándor, Z., Wedel, M., 2001. Designing Conjoint Choice Experiments Using Managers' Prior Beliefs. J. Mark. Res. 38, 430–444.

Sicsic, J., Pelletier-Fleury, N., Moumjid, N., 2018. Women's Benefits and Harms Trade-Offs in Breast Cancer Screening: Results from a Discrete-Choice Experiment. Value Health 21, 78–88. https://doi.org/10.1016/j.jval.2017.07.003

Simon, H.A., 1990. Invariants of Human Behavior. Annu. Rev. Psychol. 41, 1–20. https://doi.org/10.1146/annurev.ps.41.020190.000245

Simon, H.A., 1979. Rational Decision Making in Business Organizations. Am. Econ. Rev. 69, 493–513.

Stanovich, K.E., West, R.F., 2000. Individual differences in reasoning: implications for the rationality debate? Behav. Brain Sci. 23, 645–665; discussion 665-726.

Sudman, S., Bradburn, N., Schwarz, N., 1996. Thinking about answers: The application of cognitive processes to survey methodology, San Francisco: Jossey-Bass. ed.

Swait, J., Adamowicz, W., 2001. Choice Environment, Market Complexity, and Consumer Behavior: A Theoretical and Empirical Approach for Incorporating Decision Complexity into Models of Consumer Choice. Organ. Behav. Hum. Decis. Process. 86, 141–167. https://doi.org/10.1006/obhd.2000.2941

Uggeldahl, K., Jacobsen, C., Lundhede, T.H., Olsen, S.B., 2016. Choice certainty in Discrete Choice Experiments: Will eye tracking provide useful measures? J. Choice Model. 20, 35–48. https://doi.org/10.1016/j.jocm.2016.09.002

Viney, R., Savage, E., Louviere, J., 2005. Empirical investigation of experimental design properties of discrete choice experiments in health care. Health Econ. 14, 349–362. https://doi.org/10.1002/hec.981

Wang, H., 1997. Treatment of "Don't-Know" Responses in Contingent Valuation Surveys: A Random Valuation Model. J. Environ. Econ. Manag. 32, 219–232. https://doi.org/10.1006/jeem.1996.0965

Tables

Table 1
Summary of the two case studies

Characteristics	Case study 1	Case study 2
Setting	France	Canada
Topic	Breast cancer screening	Genomic sequencing
Language	French	English / French
Respondents	Women (age 40-74 years)	Citizens (age 18+ years)
Sample size	812	1200
Choice tasks per respondent	8 (+1)	16
Response certainty scale	0 to 10	0 to 10
Response time	yes	yes
Monotonicity test	yes	no

Table 2
Taxonomy of choice certainty

Group label	Choice certainty		Assumptions	
	Mean	SD	Engagement in deliberative thinking	Data quality
Constantly uncertain (CU)	<=6	$< \sigma_{uncertain}$	low	low
Variably uncertain (VU)	<=6	$>= \sigma_{uncertain}$	low/moderate	low/moderate
Constantly hesitant (CH)	]6 ; 8]	$< \sigma_{hesitant}$	low	low
Variably hesitant (VH)	]6 ; 8]	$>= \sigma_{hesitant}$	high	high
Constantly certain (CC)	>8	$< \sigma_{certain}$	low	low
Variably certain (VC)	>8	$>= \sigma_{certain}$	high	high

Note: σ denotes the mean certainty standard deviation of a class. For instance, $\sigma_{uncertain}$ is the mean SD of the *uncertain* class.

Table 3
Descriptive statistics of the certainty groups

Certainty group	**Case study 1** (Breast cancer screening)						**Case study 2** (Genomic sequencing)					
	Sample size		*Certainty distribution*				*Sample size*		*Certainty distribution*			
	N	%	Mean	SD	Range	[95% CI][a]	N	%	Mean	SD	Range	[95% CI][a]
Constantly uncertain (CU)	113	13.9%	4.97	0.98	[1 to 8]	[3 to 6]	126	10.5%	4.50	1.63	[0 to 7]	[0 to 6]
Variably uncertain (VU)	78	9.6%	4.87	1.93	[0 to 10]	[1 to 8]	94	7.8%	5.10	1.91	[0 to 10]	[2 to 8]
Constantly hesitant (CH)	179	22.0%	7.30	0.75	[5 to 9]	[6 to 8]	290	24.2%	7.19	0.78	[5 to 10]	[6 to 8]
Variably hesitant (VH)	145	17.9%	7.17	1.52	[0 to 10]	[5 to 10]	215	17.9%	7.05	1.64	[0 to 10]	[4 to 9]
Constantly certain (CC)	166	20.4%	9.31	0.74	[7 to 10]	[8 to 10]	264	22.0%	9.33	0.62	[7 to 10]	[8 to 10]
Variably certain (VC)	131	16.2%	8.82	1.28	[0 to 10]	[6 to 10]	211	17.6%	8.64	1.16	[0 to 10]	[7 to 10]
Overall	812	100%	7.38	1.97	[0 to 10]	[4 to 10]	1200	100%	7.44	1.99	[0 to 10]	[4 to 10]

[a] Based on the empirical distribution: 95% of certainty scores range between these two limits

Table 4
Test of conjecture 1: Choice certainty and deterministic choice patterns

Certainty group	Serial non-traders				Serial non-demanders			
	Case study 1		Case study 2		Case study 1		Case study 2	
	N	%	N	%	N	%	N	%
Constantly uncertain (CU)	12	10.2%	0	0.0%	6	5.3%	27	21.4%
Variably uncertain (VU)	4	5.1%	0	0.0%	1	1.3%	5	5.3%
Constantly hesitant (CH)	8	4.5%	2	0.93%	12	6.7%	25	8.6%
Variably hesitant (VH)	1	0.7%	1	0.34%	1	0.7%	5	2.3%
Constantly certain (CC)	1	0.6%	1	0.38%	29	17.5%	135	51.1%
Variably certain (VC)	1	0.8%	0	0.0%	1	0.8%	23	10.9%
Fisher' exact test (p-value)	p<0.0001				p<0.0001		p<0.0001	
Total	27	3.3%	4	0.3%	50	6.2%	220	18.3%

Table 5
Tests of conjectures 2 and 3: Choice certainty, monotonicity of preferences, and response time

Certainty group	Monotonicity of preferences				Response time			
	Case study 1		Case study 2		Case study 1		Case study 2	
	N	%	N	%	Median	% speeder[a]	Median	% speeder
Constantly uncertain (CU)	77	68.1%	NA	NA	12.8	27.4%	12.8	23,0%
Variably uncertain (VU)	68	87.2%	NA	NA	15.6	16.7%	15.4	22.3%
Constantly hesitant (CH)	147	68.1%	NA	NA	14.1	21.8%	14.8	18.9%
Variably hesitant (VH)	131	90.3%	NA	NA	16.3	8.9%	16.1	15.8%
Constantly certain (CC)	127	76.5%	NA	NA	13.9	16.7%	13.8	18.9%
Variably certain (VC)	124	94.7%	NA	NA	15.8	9.2%	17.2	11.8%
P-value of independence test	$p<0.0001$[b]				$p<0.0001$[c]	$p<0.0001$[b]	$p<0.0001$[c]	$p=0.080$[b]
Total	674	83,0%			14.8	16.8%	15.1	17.8%

NA: not available
[a] Percentage of respondents with a total response time strictly below 10 minutes.
[b] Chi-square independence test.
[c] Nonparametric equality-of-medians test.

Table 6
Test of conjecture 4: Choice certainty and internal validity

Certainty group	Internal validity: MNL results (homogeneous preferences)				Internal validity: MXL results (heterogeneous preferences)			
	Case study 1		Case study 2		Case study 1		Case study 2	
	N	%	N	%	N	%	N	%
Constantly uncertain (CU)	4/8	50%	7/10	70%	4/8	50%	7/10	80%
Variably uncertain (VU)	5/8	63%	8/10	80%	6/8	75%	8/10	80%
Constantly hesitant (CH)	7/8	88%	9/10	90%	7/8	88%	9/10	80%
Variably hesitant (VH)	8/8	100%	10/10	100%	8/8	100%	10/10	100%
Constantly certain (CC)	5/8	63%	8/10	80%	7/8	88%	10/10	100%
Variably certain (VC)	8/8	100%	9/10	90%	8/8	100%	10/10	100%
Total	8/8	100%	9/10	90%	8/8	100%	9/10	90%

MNL: Multinomial logit model ; MXL: Mixed multinomial logit model
[a] Number (N) and percentage (%) of preference parameters with expected sign.

Table 7
Test of conjecture 5: Choice certainty and choice consistency (case study 1)

	Model 1		Model 2		Model 3		Model 4		Model 5	
Attribute	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Breast cancer mortality	-0.749***	(0.148)	-0.822***	(0.102)	-0.830***	(0.094)	-0.837***	(0.097)	-0.871***	(0.102)
False-positive	-0.009	(0.005)	-0.010*	(0.006)	-0.014**	(0.006)	-0.012**	(0.006)	-0.013**	(0.006)
Overdiagnosis	-0.053***	(0.011)	-0.058***	(0.007)	-0.060***	(0.007)	-0.060***	(0.007)	-0.063***	(0.007)
Screening referral (doctor)	-0.059***	(0.017)	-0.064***	(0.016)	-0.066***	(0.014)	-0.065***	(0.016)	-0.067***	(0.016)
Travel time	-0.033**	(0.010)	-0.036***	(0.009)	-0.040***	(0.009)	-0.038***	(0.009)	-0.041***	(0.009)
Number of tests	-0.005	(0.047)	-0.006	(0.052)	-0.057	(0.049)	-0.035	(0.050)	-0.056	(0.052)
OOP refunded	0.018	(0.031)	0.019	(0.035)	0.039	(0.036)	0.032	(0.035)	0.047	(0.038)
OOP €30	-0.014	(0.034)	-0.015	(0.038)	-0.033	(0.040)	-0.03	(0.040)	-0.03	(0.042)
OOP €60	-0.164***	(0.043)	-0.180***	(0.041)	-0.201***	(0.040)	-0.188***	(0.040)	-0.224***	(0.042)
Scale function parameters[a]										
Certainty (choice level)	0.015	(0.024)	-	-	-	-	-	-	-	-
Response time (overall)	-	-	0.000	(0.002)	-	-	-	-	-	-
Constantly uncertain (CU)	-	-	-	-	-0.233*	(0.137)	-	-	-	-
Variably uncertain (VU)	-	-	-	-	-0.121	(0.139)	-	-	-	-
Constantly hesitant (CH) - *ref*	-	-	-	-	*ref*		-	-	-	-
Variably hesitant (VH)	-	-	-	-	0.277***	(0.098)	-	-	-	-
Constantly certain (CC)	-	-	-	-	-0.504***	(0.184)	-	-	-	-
Variably certain (VC)	-	-	-	-	0.508***	(0.095)	-	-	-	-
Mean certainty	-	-	-	-	-	-	0.048	(0.030)	0.068**	(0.031)
SD certainty	-	-	-	-	-	-	0.301***	(0.062)	0.325***	(0.075)
Mean certainty*SD certainty	-	-	-	-	-	-	-	-	0.211***	(0.049)
# observations	6496		6496		6496		6496		6496	
# respondents	812		812		812		812		812	
Log-Likelihood	-5951.7455		-5952.6306		-5855.2717		-5906.5256		-5848.2509	

[a] Scale function estimated from heteroskedastic multinomial logit models. Significance levels: ***<1%, **<5%, *<10%
OOP : out-of-pocket

Table 8
Test of conjecture 5: Choice certainty and choice consistency (case study 2)

	Model 1		Model 2		Model 3		Model 4		Model 5	
Attribute	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)	Estimate	(SE)
Disease risk: 40% lifetime risk of higher	0.330***	(0.066)	0.234***	(0.029)	0.363***	(0.038)	0.471***	(0.091)	0.380***	(0.038)
Disease risk: 80% lifetime risk or higher	0.398***	(0.071)	0.289***	(0.030)	0.453***	(0.037)	0.580***	(0.101)	0.469***	(0.038)
Disease risk: 90% lifetime risk or higher	0.379***	(0.071)	0.270***	(0.031)	0.423***	(0.039)	0.537***	(0.097)	0.438***	(0.040)
Recommended effective medical treatment only	-0.026	(0.033)	0.049**	(0.019)	0.110***	(0.028)	0.105**	(0.044)	0.085***	(0.031)
Recommended effective medical treatment and lifestyle change	0.212***	(0.043)	0.183***	(0.021)	0.279***	(0.028)	0.352***	(0.067)	0.284***	(0.030)
No recommended treatment or lifestyle change	-0.370***	(0.064)	-0.162***	(0.024)	-0.163***	(0.045)	-0.276***	(0.051)	-0.225***	(0.036)
Moderate QOL consequences	0.112***	(0.038)	0.128***	(0.022)	0.224***	(0.030)	0.263***	(0.064)	0.210***	(0.034)
Severe QOL consequences	0.066*	(0.038)	0.107***	(0.021)	0.175***	(0.030)	0.205***	(0.055)	0.168***	(0.034)
Very severe QOL consequences	0.039	(0.044)	0.082***	(0.024)	0.141***	(0.035)	0.158***	(0.056)	0.130***	(0.039)
Carrier status	0.481***	(0.073)	0.299***	(0.027)	0.404***	(0.045)	0.563***	(0.087)	0.455***	(0.029)
Cost	-0.001***	(0.000)	-0.001***	(0.000)	-0.001***	(0.000)	-0.001***	(0.000)	-0.001***	(0.000)
Scale function parameters[a]										
Certainty (choice level)	0.006	(0.022)	-	-	-	-	-	-	-	-
Response time (overall)	-	-	0.025***	(0.004)	-	-	-	-	-	-
Constantly uncertain (CU)	-	-	-	-	0.060	(0.109)	-	-	-	-
Variably uncertain (VU)	-	-	-	-	0.018	(0.146)	-	-	-	-
Constantly hesitant (CH) - *ref*	-	-	-	-	*ref*		-	-	-	-
Variably hesitant (VH)	-	-	-	-	0.226*	(0.136)	-	-	-	-
Constantly certain (CC)	-	-	-	-	-0.888**	(0.397)	-	-	-	-
Variably certain (VC)	-	-	-	-	0.238***	(0.092)	-	-	-	-
Mean certainty	-	-	-	-	-	-	-0.038**	(0.019)	-0.007	(0.022)
SD certainty	-	-	-	-	-	-	0.062	(0.049)	0.118**	(0.051)
Mean certainty*SD certainty	-	-	-	-	-	-	-	-	0.090***	(0.027)
# observations	19200		19200		19200		19200		19200	
# respondents	1200		1200		1200		1200		1200	
Log-Likelihood	-19978,414		-19928,974		-19904,959		-19970,004		-19956,269	

[a] Scale function estimated from heteroskedastic multinomial logit models.
Significance levels: ***<1%. **<5%. *<10

APPENDICES

Appendix A. Description of choice tasks in the two case studies

Figure A1

Example of a breast cancer screening DCE choice task (case study 1)

	Screening option A	Screening option B	No screening option
BC mortality	10	25	30
False-positive	200	50	0
Overdiagnosis	150	10	0
Type of screening referral	invitation letter	your doctor	none
Travel time	10 min	90 min	0 min
Number of tests	18	12	0
Out-of-pocket cost	€ 60	€ 30	€ 0
Which option would you choose ?	□	□	□

Note: In each choice task, respondents could attain a more detailed presentation of the attributes by clicking on the attribute label.

A more thorough description of attributes and levels is available in (Sicsic et al., 2018) : Women's Benefits and Harms Trade-Offs in Breast Cancer Screening: Results from a Discrete-Choice Experiment. Value Health 21, 78–88. doi.org/10.1016/j.jval.2017.07.003

Figure A2

Example of return of genomic information DCE choice task (case study 2)

	Option A	Option B	No information
Disease Risk *More diseases will be identified if the lifetime risk is lower*	Diseases with a **5% lifetime risk or higher**	Diseases with a **90% lifetime risk or higher**	No information
Disease Treatability	Recommended effective medical treatment only	Recommended effective lifestyle change only	No information
Disease Severity *Health consequences of the diseases you may develop*	Very severe health consequences	Severe health consequences	No information
Carrier Status *Disease risk not affecting you but can affect your family*	Does not provide information on carrier status	Information on if your family members could be affected	No information
Cost to you	$1500	$750	$0
	Option A □	**Option B** □	**No Information** □

A more thorough description of attributes and levels is available in (Regier et al., 2015) : Societal preferences for the return of incidental findings from clinical genomic sequencing: a discrete-choice experiment. CMAJ 187, E190-197. doi.org/10.1503/cmaj.140697

Appendix B. Results of the stratified MNL and MXL models investing internal validity across choice certainty groups

Table B1

Results of the multinomial logit (MNL) models stratified by choice certainty groups (**case study 1**)

Attribute	Expected sign	Constantly Uncertain	Variably Uncertain	Constantly Hesitant	Variably Hesitant	Constantly Certain	Variably Certain
		MLE	MLE	MLE	MLE	MLE	MLE
Breast cancer mortality	$\beta_1 < 0$	-0.164	-0.412	-1.034***	-1.177***	-0.617***	-1.496***
False-positive	$\beta_2 < 0$	0.020	-0.007	-0.026**	-0.031**	0.008	-0.024
Overdiagnosis	$\beta_3 < 0$	-0.014	-0.044*	-0.067***	-0.097***	-0.031**	-0.102***
Screening referral (doctor)	$\beta_4 = ?$	0.023	-0.018	-0.041	-0.070**	-0.111***	-0.176***
Travel time	$\beta_5 < 0$	0.009	-0.008	-0.065***	-0.072***	-0.002	-0.072***
Number of tests	$\beta_6 < 0$	0.179	0.020	-0.105	-0.246*	0.208**	-0.150
OOP refunded	$\beta_7 > 0$[a]	-0.026	-0.001	0.026	0.072	0.010	0.340**
OOP €30	$\beta_8 < 0$	-0.062	0.013	0.054	-0.072	0.016	-0.251
OOP €60	$\beta_9 < 0$	-0.121	-0.019	-0.099	-0.199*	-0.181**	-0.660***
# observations		904	624	1432	1160	1328	1048
# respondents		113	78	179	145	166	131
Log-Likelihood		-874.1021	-587.1207	-1281.5399	-959.8106	-1328.3926	-769.4856

MLE: Maximum likelihood estimates. OOP : out-of-pocket

Significance levels: ***<1%, **<5%, *<10%

Coefficients highlighted in red do not have the expected sign.

[a] The β_7 coefficient is expected to be positive because the OOP attribute is effect coded, so β_7 is interpreted in relation to the mean expected impact of OOP over all its levels. A rational woman is expected to experience relatively less disutility of having to advance fees (despite being refunded lately) compared to having to pay €30 or €60 for a mammogram out of pocket (i.e., with no reimbursement).

Table B2

Results of the mixed multinomial logit (MXL) models stratified by choice certainty groups (**case study 1**)

	Expected sign	Constantly Uncertain	Variably Uncertain	Constantly Hesitant	Variably Hesitant	Constantly Certain	Variably Certain
Attribute	MLE	MLE	MLE	MLE	MLE	MLE	MLE
Breast cancer mortality	$\beta_1 < 0$	-0.235	-0.548	-2.121***	-2.255***	-2.017***	-2.924***
False-positive	$\beta_2 < 0$	0.021	-0.014	-0.070***	-0.070***	-0.023	-0.068**
Overdiagnosis	$\beta_3 < 0$	-0.018	-0.065**	-0.135***	-0.170***	-0.096***	-0.192***
Screening referral (doctor)	$\beta_4 = ?$	0.029	0.006	-0.067	-0.086*	-0.223***	-0.249***
Travel time	$\beta_5 < 0$	0.006	-0.013	-0.140***	-0.134***	-0.065	-0.161***
Number of tests	$\beta_6 < 0$	0.187	0.050	-0.406**	-0.506**	0.160	-0.363
OOP refunded	$\beta_7 > 0$[a]	0.035	0.057	0.058	0.095	0.813*	0.782**
OOP €30	$\beta_8 < 0$	-0.135	0.001	0.084	-0.031	-0.749*	-0.688*
OOP €60	$\beta_9 < 0$	-0.239	-0.071	-0.337	-0.473**	-1.255***	-1.407***
# observations		904	624	1432	1160	1328	1048
# respondents		113	78	179	145	166	131
Log-Likelihood		-719.917	-503.712	-947.689	-809.926	-747.123	-640.202

MLE: Maximum likelihood estimates. OOP : out-of-pocket.

Significance levels: ***<1%, **<5%, *<10%

Coefficients highlighted in red do not have the expected sign.

[a] The β_7 coefficient is expected to be positive because the OOP attribute is effect coded, so β_7 is interpreted in relation to the mean expected impact of OOP over all its levels. A rational woman is expected to experience relatively less disutility of having to advance fees (despite being refunded lately) compared to having to pay €30 or €60 for a mammogram out of pocket (i.e., with no reimbursement).

Table B3
Results of the MNL models stratified by choice certainty groups (**case study 2**)

	Expected sign	Constantly Uncertain	Variably Uncertain	Constantly Hesitant	Variably Hesitant	Constantly Certain	Variably Certain
Attribute		MLE	MLE	MLE	MLE	MLE	MLE
40% lifetime risk of higher	β1 > 0	0.1353*	0.1007	0.0865*	0.0680	0.0485	-0.0123
80% lifetime risk or higher	β2 > 0	0.0022	-0.0400	0.0998**	0.2105***	0.1057**	0.1920***
90% lifetime risk or higher	β3 > 0	-0.0226	0.0776	0.0934**	0.1427***	0.0274	0.1760***
Recommended effective medical treatment only	β4 > 0	-0.0373	0.0286	0.0293	0.0384	-0.0039	0.0467
Recommended effective medical treatment and lifestyle change	β5 > 0	0.2000***	0.1785***	0.2489***	0.1835***	0.2062***	0.3151***
No recommended treatment or lifestyle change	β6 < 0	-0.2373***	-0.2746***	-0.3027***	-0.2803***	-0.1899***	-0.3956***
Moderate QOL consequences	β7 > 0	0.0447	0.1116*	0.1158***	0.0017	-0.0025	0.0203
Severe QOL consequences	β8 > 0	-0.0437	-0.1152*	-0.0518	0.1607***	0.0464	0.012
Very severe QOL consequences	β9 = ? [a]	-0.0107	-0.0792	-0.0692	0.0262	-0.0133	0.0364
Carrier status	β10 > 0	0.1804***	0.1221***	0.2599***	0.1217***	0.2374***	0.2945***
Cost	β11 < 0	-0.0015***	-0.0012***	-0.0011***	-0.0011***	-0.0005***	-0.0009***
# observations		2016	1504	4640	3440	4224	3376
# respondents		126	94	290	215	264	211
Log-Likelihood		-2042.0962	-1527.9215	-4545.3372	-3429.7786	-4082.464	-3429.8822

MLE: Maximum likelihood estimates

Significance levels: ***<1%, **<5%, *<10%

Coefficients highlighted in red do not have the expected sign.

[a] There is no clear theoretical assumption concerning the sign of this effect because on the one hand, individuals may want to have information about the risk of developing a severe disease but on the other hand, they may be anxious when thinking about developing a very severe disease.

Table B4
Results of the mixed multinomial logit (MXL) models stratified by choice certainty groups (**case study 2**)

Attribute's levels	Expected sign	Constantly Uncertain MLE	Variably Uncertain MLE	Constantly Hesitant MLE	Variably Hesitant MLE	Constantly Certain MLE	Variably Certain MLE
40% lifetime risk of higher	β1 > 0	0.382*	0.235	0.415***	0.243**	0.350**	0.627***
80% lifetime risk or higher	β2 > 0	0.595***	0.270*	0.755***	0.931***	0.923***	1.178***
90% lifetime risk or higher	β3 > 0	0.244	0.293**	0.590***	0.685***	0.596***	1.086***
Recommended effective medical treatment only	β4 > 0	-0.179	-0.056	-0.064	0.005	0.127	0.061
Recommended effective medical treatment and lifestyle change	β5 > 0	0.009	0.101	0.304***	0.135	0.456***	0.357***
No recommended treatment or lifestyle change	β6 < 0	-0.751***	-0.597***	-0.598***	-0.665***	-0.465***	-0.817***
Moderate QOL consequences	β7 > 0	0.166	0.130	0.190**	0.353***	0.411***	0.249**
Severe QOL consequences	β8 > 0	-0.006	-0.264*	-0.003	0.530***	0.281**	0.174*
Very severe QOL consequences	β9 = ? [a]	-0.212	-0.267	-0.095	0.259**	-0.007	0.194
Carrier status	β10 > 0	0.727***	0.354***	0.794***	0.293***	0.954***	0.890***
Cost	β11 < 0	-0.005***	-0.002***	-0.002***	-0.002***	-0.005***	-0.002***
# observations		2016	1504	4640	3440	4224	3376
# respondents		126	94	290	215	264	211
Log-Likelihood		-1016.943	-1197.2052	-2969.3629	-2732.985	-1593.1628	-2252.795

MLE: Maximum likelihood estimates

Significance levels: ***<1%, **<5%, *<10%

Coefficients highlighted in red do not have the expected sign.

[a] There is no clear theoretical assumption concerning the sign of this effect because on the one hand, individuals may want to have information about the risk of developing a severe disease but on the other hand, they may be anxious when thinking about developing a very severe disease.

Appendix C. Using mean choice certainty and variability to derive more accurate welfare estimates: ex-post calibration strategy based on reweighting

In this Appendix, we propose one strategy to incorporate information on respondent' choice certainty and variability in discrete choice models including an opt-out. First, we present the benchmark model used to estimate preferences at the sample level using data from our two cases studies. This model can be applied to any DCE including at least 2 alternatives and one opt-out. Then, we compare the results of models that differ in the way data are weighted based on respondent' certainty. In our comparison we focus on the precision and the magnitude of welfare estimates derived from the choice models.

C.1. Econometric models

C.1.1. The error components model

In order to account for the non-independence of the data provided by the same respondent and the non-independence of alternatives in choice data including 2 or more alternatives and an opt-out, the benchmark model to be estimated is an error component logit (ECL) model specified as follows (Scarpa et al., 2005; Train, 2003):

$$U_{njt} = \alpha_n ASC_{optionA|B} + \beta' x_{jt} + \epsilon_{njt} \quad (1a)$$

Where U_{njt} is the utility individual *n* derives from choosing alternative *j* in choice scenario *t*, $ASC_{optionA|B}$ is a generic intercept entering the utility function of the two testing (i.e. non opt-out) alternatives, with α_n the associated coefficient assumed normally distributed representing the individual systematic tendency to choose a testing alternative (e.g., breast cancer screening or genomic testing, respectively). Thus α_n allows accounting for various non-trading behaviours such as individuals who are always choosing to opt-in ($ASC_{optionA|B} \gg 0$) or individuals who are always choosing to opt-out ($ASC_{optionA|B} < 0$).

The simulated log-likelihood (SLL) of the sample associated to the ECL model is written as:

$$SLL = \sum_{n=1}^{N} ln\left\{\frac{1}{R}\sum_{r=1}^{R} P_n(\alpha^r)\right\} \quad (1b)$$

Where R is the number of Halton draws used for simulation, α^r is the rth draw from the (normal) distribution of α_n, and $P_n(\alpha^r)$ is the unconditional probability of individual *n*'s sequence of choices evaluated at the value of the r^{th} draw. These unconditional probabilities are then averaged over the R draws, and the SLL is maximized at convergence.

In can be seen from Eq. (1b) that the SLL can be manipulated to re-weight categories of respondents, i.e. give some respondents more influence in the estimation process. The weighted simulated log-likelihood (wSLL) function is written as follows:

$$wSLL = \sum_{n=1}^{N} w_n * ln\left\{\frac{1}{R}\sum_{r=1}^{R} P_n(\alpha^r)\right\} \quad (1c)$$

With w_n denoting individual weights that are defined below.

C.1.2. The re-weighting models

We describe the different weights that enter the expression of SLL function in Eq. (1c). Four weighted error component logit (WECL 1 to 4) models are defined by specifying different w_n.

$(\mathbf{1})$ **WECL1**: $\boldsymbol{w_n = 1}$, with *n=1…..N* denoting each respondent. Each respondent is given an equal weight =1. WECL1 collapses with the ECL model presented in Eqs. (1b).

$(\mathbf{2})$ **WECL2 :** $\boldsymbol{w_n = \overline{certainty_{nt}}}$, where $\overline{certainty_{nt}}$ denotes the mean certainty of respondent *n* over the T choices.

$(\mathbf{3})$ **WECL3:** $\boldsymbol{w_n = \sigma_{certainty,n}}$, where $\sigma_{certainty.n}$ denotes the certainty's standard deviation calculated for respondent *n* over the T choices. In WECL3, we assume that deliberative thinking is an increasing function of choice certainty variability.

$(\mathbf{4})$ **WECL :** $\boldsymbol{w_n = \overline{certainty_{nt}} * \sigma_{certainty,n}}$. In line with the results of the HMNL models estimated in Section 5 of the paper, we assume that giving more weight to individuals with higher mean certainty and certainty variability could reduce response error and thus improve the efficiency of the econometric model (i.e., lower standard errors). In WECL4, we assume that deliberative thinking is an increasing function of both mean choice certainty and certainty variability.

For all the WECL models, 5000 Halton draws are used to simulate the log-likelihood of the sample. Prior to estimating the models, the weights are normalized so that the sum of the individual specific weights are equal to the number of respondents in each sample population (Regier et al., 2014). The standardization of weights allows for the correct calculation of parameter estimates' standard errors and also (as sample size remains unchanged) allows for inter-model comparisons of weighted simulated log-likelihood (wSLL) at convergence[13].

In case study 1, the ECL model results are used to calculate marginal rates of substitution (MRS) between breast cancer mortality and the four other quantitative attributes (false-positive risk, over-diagnosis risk, travel time, and number of screening tests over a lifetime). The four MRS are interpreted as willingness-to-accept (WTA) and represent women's benefit and harm trade-offs. For instance, the MRS between over-diagnosis and breast cancer mortality is interpreted as the number of over diagnosed cases women are willing to accept, on average, to avoid one breast cancer -related death (Sicsic et al., 2018). In case study 2, the ECL model results are used to calculate willingness to pay (WTP) to receive information on risk of disease, disease treatability, and carrier status (Regier et al., 2015).

To investigate the impact of the re-weighting models on statistical efficiency, we first compare the standard errors (SEs) and width of the 95 percent confidence interval (95% CI) around WTP/WTA estimates in WECL2-4 compared to WECL1 (benchmark model). The standard errors provide a direct measure of statistical efficiency and thus precision of welfare estimates. We also investigate precision by computing D-error, which is the determinant of the inverse of the variance-covariance matrix scaled by the number of estimated parameters. D-error estimates that are lower indicate greater model efficiency. We expect: 1) an increase in the SEs and D-error in WECL2 (synonymous of lower efficiency) as this model gives highest weights to *constantly certain* individuals who have higher response error, and 2) a decrease in the SEs and D-error in WECL3-4 (synonymous of higher efficiency) as these models place more weights in respondents who answer the DCE in line with deliberative thinking.

In WECL 3-4, we expect lower WTA estimates (case study 1) and lower WTP estimates (case study 2), because we give more weight to more thoughtful respondents who may be less subject to hypothetical bias (Loomis, 2011; Ready et al., 2010).

[13] We expect that the wSLL in WECL2 should be higher than that of WECL1 because in WECL2, the "*constantly certain*" individuals are given the highest weights. In section 4, we found that *constantly certain* individuals included a high proportion of individuals with deterministic preferences (e.g. serial non-traders or serial non-demanders). It can be seen from Eq. (1b)-(1c) that as the proportion of individuals with more deterministic preferences (i.e. with higher P_n) increases, so does the wSLL function. On the contrary, we assume that the wSLL in WECL3 to 5 should decrease because individuals with more deterministic preferences (higher P_n) are given less weights.

C.2. Results of the re-weighting models

C.2.1. Case study 1

The results of the WECL models (WTA, SEs and 95%CI) estimated from case study 1 are presented in Table C1. As expected, the SLL increases in WECL2 compared to WECL1 (benchmark model). This is because we place more weight on individuals with more deterministic preferences (the "*constantly certain*") thus increasing the predictive value of the model, whereas the SLL decreases in WECL3-4 as a result of giving lower weights to respondents with more deterministic preferences.

The D-error and the SEs around two WTA estimates (namely, travel time and number of screening tests) are higher in WECL2 compared to WECL1, which is consistent with our theoretical predictions. In WECL 3 and 4, the SEs of all WTA estimates are lower as is the D-error, which is synonymous of higher statistical precision. The model with greatest impact on statistical efficiency is WECL4. In WECL4, statistical efficiency improvements range from 23% to 69%. These improvements are particularly important for two WTA measures, false-positive results and screening tests: there is a 41% decrease in the SEs for WTA false-positives and a 69% decrease in the SEs for WTA screening tests.

In WECL3 and 4, there is a decrease in the WTA estimates for false-positives and screening tests. In WECL1, women are willing to accept on average 50.27 false-positives and 9.89 additional screening tests to save one (statistical) life from breast cancer, compared to 43.74 (-13%) false-positives and 6.38 (-35%) screening tests in WECL4.

C.2.2. Results of the re-weighting models (case study 2)

The results of the WECL models (WTP, SEs and 95%CI) estimated from case study 2 are presented in Table C2. The results are in line with a-priori expectations and consistent with those from case study 1. The SLL increases in WECL2 compared to WECL1, and decreases in WECL3 and 4. We find the SEs around WTP estimates are all higher (from 7% to 9%) in WECL2 compared to WECL1, and the D-error is higher. Conversely, in WECL3 and 4, all the SEs around WTP estimates as well as the D-error are lower. WECL 3 provides the greatest improvement in statistical efficiency. Compared to case study 1, there is lower impact of the re-weighting function on the precision of welfare estimates (with reduction in SEs ranging from 2% to 4%), but a higher impact on the overall statistical efficiency as measured through D-error (with reduction in D-error of 15 to 16%).

Similarly to case study 1, we observe higher WTP estimates in WECL2 and lower WTP estimates in WECL3-4 for all but one WTP parameter (risk of disease for diseases with 90% lifetime risk or higher). The reduction in WTP are largest in WECL3. For instance, the individuals are willing to pay on average \$390.53 to have information on risk of disease affecting his/her family (carrier status) in WECL1 as compared to \$345.31 (-12%) in WECL4.

In both case studies, when we followed current practice and re-weighted to favour respondents with higher mean certainty, we found detrimental impacts on both the precision and plausibility of welfare estimates. However, including certainty variability in the re-weighting function improved the precision of welfare estimates and decreased the willingness-to-pay (accept) estimates. As willingness-to-pay are usually overestimated in stated preference research (Blumenschein et al., 2001; Morrison and Brown, 2009; Murphy et al., 2005), this would suggest our ex-post calibration strategy effectively reduced hypothetical bias. However, caution is needed in interpreting this result due to the absence of data on revealed preference. At least, our results indicate that the re-weighting functions based on certainty variability provided more plausible welfare estimates.