



A Picture is Worth a Thousand Words: The Role of Survey Training Materials in Stated-Preference Studies

Caroline M. Vass^{1,2} · Niall J. Davison^{1,3} · Geert Vander Stichele⁴ · Katherine Payne¹

© The Author(s) 2019

Abstract

Background Online survey-based methods are increasingly used to elicit preferences for healthcare. This digitization creates an opportunity for interactive survey elements, potentially improving respondents' understanding and/or engagement.

Objective Our objective was to understand whether, and how, training materials in a survey influenced stated preferences.

Methods An online discrete-choice experiment (DCE) was designed to elicit public preferences for a new targeted approach to prescribing biologics ("biologic calculator") for rheumatoid arthritis (RA) compared with conventional prescribing. The DCE presented three alternatives, two biologic calculators and a conventional approach (opt out), described by five attributes: delay to treatment, positive predictive value, negative predictive value, infection risk, and cost saving to the national health service. Respondents were randomized to receive training materials as plain text or an animated storyline. Training materials contained information about RA and approaches to treatment and described the biologic calculator. Background questions included sociodemographics and self-reported measures of task difficulty and attribute non-attendance. DCE data were analyzed using conditional and heteroskedastic conditional logit (HCL) models.

Results In total, 300 respondents completed the DCE, receiving either plain text ($n = 158$) or the animated storyline ($n = 142$). The HCL showed the estimated coefficients for all attributes aligned with a priori expectations and were statistically significant. The scale term was statistically significant, indicating that respondents who received plain-text materials had more random choices. Further tests suggested preference homogeneity after accounting for differences in scale.

Conclusions Using animated training materials did not change the preferences of respondents, but they appeared to improve choice consistency, potentially allowing researchers to include more complex designs with increased numbers of attributes, levels, alternatives or choice sets.

1 Introduction

Stated-preference methods are a group of approaches used to elicit and then quantify individuals' preferences for health states, goods, or services [1, 2]. Stated-preference methods are more popular in areas where markets are imperfect and consumers cannot easily be observed (such as the environment or healthcare) or for forecasting demand for new

technologies [3, 4]. Most stated-preference methods involve rating, ranking, or making a choice between hypothetical options presented as questions in a survey [5]. In healthcare, stated-preference methods such as time trade-off, standard gamble, and contingent valuation have been used to understand people's time preferences, risk tolerance, and willingness to pay, respectively. However, the hypothetical nature of stated-preference methods has resulted in some criticism [6, 7].

Discrete-choice experiments (DCEs) are an increasingly popular type of stated-preference method [4]. In a DCE, respondents select their preferred alternative from a set in a series of hypothetical choices in a survey. Respondents are expected to make trade-offs between different attributes of the good or service to make their decision. In healthcare, where market data rarely exist, the quantification of preferences through DCEs allows decision makers to understand which aspects of an intervention provide the most benefit. Interest has also increased in using the results of DCEs to

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s40271-019-00391-w>) contains supplementary material, which is available to authorized users.

✉ Katherine Payne
Katherine.payne@manchester.ac.uk

¹ Manchester Centre for Health Economics, The University of Manchester, Manchester, UK

² RTI Health Solutions, Manchester, UK

³ BresMed, Manchester, UK

⁴ MindBytes, Ghent, Belgium

Key Points for Decision Makers

This study found the error variance reduced when an animated storyline was used to inform respondents about the disease area and the intervention before completing a discrete-choice experiment.

As reduced error variance is related to choice consistency, the results suggest respondents were more able to complete the elicitation tasks, but the survey materials did not affect the stated preferences.

Having engaged and informed respondents is beneficial in all stated-preference studies, but the advantages may be particularly pronounced in research relating to complex healthcare interventions, with lesser reached populations (e.g., those with lower literacy), or when the research question requires a complex experiment.

inform regulatory decisions [8, 9]. However, for DCEs to be used in decision making, they must be robust and produce data that minimize bias from either their hypothetical nature or other sources [10].

To ensure the hypothetical choices reflect real-life behavior, respondents to stated-preference surveys usually receive information and explanations in the form of “training materials” before completing the valuation tasks. The importance of training materials has received little attention in the DCE literature; they are rarely described or presented in published articles or made available through online appendices. Where guidelines for general best practice [11, 12] and more specifically for the identification of attributes, experimental design, and econometric analysis are available [13–16], guidance about how to design and frame the survey training materials that are presented before the choice sets is lacking. However, psychology literature surrounding choice making in health behavior emphasizes the importance of individuals’ “capability” [17], defined as the psychological capacity to engage in the necessary thought processes to make a choice or change behavior. In the context of a healthcare DCE, this could relate to respondents’ understanding of the disease and the treatment forming the basis of the valuation exercise, their ability to retain the information presented, and their ability to make decisions or choices based on this information, all of which can, and should, be addressed in the training materials at the start of a DCE.

Communicating large volumes of complex information is notoriously difficult, as people struggle to retain the information or stay engaged to read all of it [18]. Louviere [19] highlighted “information acceleration methods,” a concept developed in the 1990s, as a way to rapidly inform individuals about new technologies and their associated benefits and

harms. The information acceleration literature was developed in marketing and management to improve strategic management decisions involving a new alternative not currently in the market [20]. Much of the information acceleration literature was produced in the 1990s and focuses on visual materials using videos rather than interactive materials [21]. More recently, “serious games,” which can be defined as “a game in which education (in its various forms) is the primary goal, rather than entertainment” [22] (p. 17) have been developed to help train and/or motivate individuals to learn about new, often complex and abstract, concepts [18]. The rationale for using a serious game is to achieve better learning outcomes by immersing the participant in an educational and enjoyable environment, which is intrinsically motivating, through the use of interactive technology. Serious games have been used in a variety of applications and examples, including educating individuals about genetics or improving their mathematical achievements [23, 24]. A published systematic review and meta-analysis found the degree to which serious games aided learning (compared with conventional text-based approaches) was positive [18]. It has been argued that, in addition to learning, serious games could keep participants interested and engaged in a task [25], improving completion rates and the quality of data collected.

Recent systematic reviews of healthcare DCEs have shown a large increase in the number of these surveys administered online. Between 2001 and 2008, only 11% of DCEs were web surveys, but between 2013 and 2017, some 57% of healthcare DCEs were online [4, 26]. This digitization of DCE surveys provides scope to incorporate animated or interactive training materials. However, the extent to which training materials affect respondents’ choices or heuristics in DCEs remains unclear. This study aimed to investigate whether, and how, the format of presenting training materials influenced the choice data collected in an example DCE. This study used the case study of preferences for a new prescribing algorithm to guide the treatment of rheumatoid arthritis (RA) with a first-line biologic (a “biologic calculator”) as an example of a complex topic requiring substantive training materials.

2 Methods

A DCE designed to elicit preferences for a “biologic calculator” compared with a conventional prescribing practice was used as the basis for this study. Respondents were then randomized to complete the survey with training materials presented as plain text or as an animated storyline. Approval for the study was obtained from The University of Manchester’s Research Ethics Committee.

2.1 Discrete-Choice Experiment Design

The DCE was designed and is reported in line with published recommendations [11, 12]. In brief, an iterative process involving clinical experts and patient representatives and supported by systematic reviews was used to identify the relevant attributes (five) and their plausible levels (four for each attribute) described in Appendix A. Extensive piloting involving qualitative and quantitative methods resulted in the choice set shown in Fig. 1, which presented the choice question using a non-labelled format for two alternatives (of biologic calculator) with an opt out (representing current prescribing practice). The selection of relevant choice sets, four blocks of five, was informed using the software Ngene and aimed to minimize D-error [27]. An internal validity check for monotonic preferences was added, so each respondent completed six choice sets.

2.2 The Survey

The survey was uploaded online using Sawtooth SSiWeb [28]. The final survey comprised the training materials to explain the purpose of the DCE followed by the six questions presenting the choice sets and questions about the individual (including sociodemographics and questions to ascertain their level of understanding).

2.3 Training Materials

Two formats for presenting training materials were developed: plain text and an animated storyline. The content presented in each format was the same. The content and text for the training materials were developed via consultation with three clinical experts in stratified medicine.

The plain-text version of the training materials (see Appendix B) was presented on 15 separate webpages to avoid the need for scrolling on a standard computer screen. The text was supported by icon arrays illustrating probabilities. The respondent was required to click through and read each page of text.¹

The animated storyline was developed with assistance from a company that developed and applied their theory-driven, evidence-based approach to develop interactive educational tools such as serious games (MindBytes, <http://www.mindbytes.be> [29]). As advised by Reeve [30], a storyline was created and a narrative developed for an avatar (a figure representing a person). The framework aims to

ensure that the narratives and animations (game mechanics) enhance the educational objectives without creating a bias. Easy-to-follow stories are proposed to be useful motivators to keep respondents engaged with a subject and process or make sense of information [30]. Although personification is important [30], evidence also suggests that people identify most with avatars like themselves [31]; to minimize this bias, a green genderless, ageless stickman avatar with a gender-neutral name (Alex) was designed. The setting was also dynamic, with different backgrounds using archetypical visuals and a simple design to indicate the location e.g., “in hospital” or “at home,” allowing us to convey these concepts and avoid both information overload and bias [32, 33]. A “linear traditional narrative” [34], the most simplistic structure, was used in which the central character (Alex) was followed along a pathway that started with a description of RA. The story then explained how first-line treatments may fail, which then requires switching to a biologic. The story then explained that the choice of which biologic and relevant dose will be made by a clinician who may decide to use a biologic calculator to guide this decision. The last elements of the story explained the relevant attributes that describe the biologic calculator and that some trade-offs must be made when choosing a prescribing approach. Each attribute was explained in the storyline with the help of graphics and visuals using the learning mechanic-game approach [35], thereby ensuring that these visuals explicitly addressed the learning objectives without introducing potentially bias-inducing elements.

2.4 Background Questions

In the last section of the survey, respondents were asked to complete a series of background questions about themselves, including quality of life (the EuroQol Five-Dimension, Five-Level instrument [EQ-5D-5L] [36]) and sociodemographics. Although “quiz” questions were considered as a measure of respondents’ understanding of the training materials, the authors decided against the approach out of concern that some respondents may exit the survey if they could not provide a correct answer or felt they were being tested or examined. Such dropout of respondents may induce a selection bias, where only informed respondents (regardless of the training materials) proceeded to the choice sets. Instead, respondents were asked: “On a scale of 1–5, how confident are you that you would make the same choices if faced with the situations in real life?” and “On a scale of 1–5, how easy or difficult did you find making choices between the alternatives?” These questions were included to understand whether there was any difference in their choice-making ability as a result of the training materials. Self-reported attribute non-attendance (ANA) was also collected. Initially, respondents were presented with a screening question: “Did you find

¹ Files for each survey are available from the authors on request. A video demonstrating the storyline is available in the Electronic supplementary material.

If these were the only approaches to prescribing biologics, which, if any, would you choose?
Choose by clicking one of the buttons at the bottom of the page:












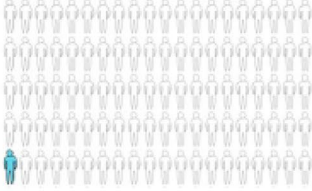
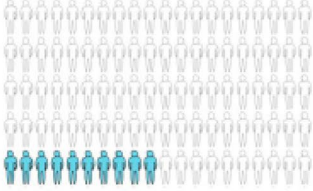
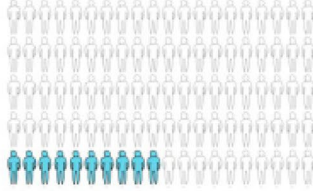


	Biologic Calculator A	Biologic Calculator B	Conventional Approach (no Biologic Calculator)
Delay to start treatment 	7 days 	14 days 	No delay 
Ability to predict who will respond 	40% Of 100 people predicted to respond, 40 respond 	80% Of 100 people predicted to respond, 80 respond 	No predictive ability
Ability to predict who will not respond 	95% Of 100 people predicted <u>not</u> to respond, 5 would have  5 people miss effective treatment	90% Of 100 people predicted <u>not</u> to respond, 10 would have  10 people miss effective treatment	No predictive ability
Risk of infection 	1% 	10% 	10% 
Annual cost saving to the NHS 	£0 a patient	£1,500 a patient 	No cost saving
	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Fig. 1 Example choice set. *NHS* national health service

yourself making choices based on one or two characteristics rather than the option as a whole?”; respondents who answered “yes” then saw five follow-up questions regarding their attention to each of the attributes.

2.5 Study Sample

The relevant study population for this survey was defined as members of the public aged ≥ 18 years. Respondents were recruited through an internet panel provider, ResearchNow[®]. Respondents were sent a link to the online survey, and individuals were randomly allocated to receive one of the two training materials (plain text or animated storyline) upon clicking the link to enter the survey.

2.6 Data Analysis

Descriptive statistics for the answers to the background questions were produced and used to summarize the respondents who completed the DCE. In addition, a logistic regression model was estimated to confirm that the randomization to training materials was successful by testing whether any sociodemographic variables predicted the survey version received (the dependent variable).

Choice data were analysed within a random utility maximization framework [37], where individuals are said to choose the alternative that provides them with the most utility. An individual's (n) utility (U_{nj}) for an alternative (j) is said to be made up of an observable component (V_{nj}) and

a random component (ε_{nj}). In this study, panel data methods were used to account for the same individual making multiple choices with fixed-effects conditional logit models. These models assume that the random component of utility has a Weibull form [38].

In this DCE, all attributes were continuous variables (time, cost, probabilities) and therefore entered the utility function as single attributes in the preliminary analysis (Eq. 1):

$$U_{njt} = \beta_{0t} + \beta_{1t}\text{Delay}_{nj} + \beta_{2t}\text{PPV}_{nj} + \beta_{3t}\text{NPV}_{nj} + \beta_{4t}\text{Risk}_{nj} + \beta_{5t}\text{Cost}_{nj} + \varepsilon_{njt}, \quad (1)$$

where β_{1-5} are the parameters associated with each of the attributes for each version of the materials, t . The functional form of preferences was also investigated by introducing square terms for each variable [39]. The term β_{0t} is an alternative-specific constant (ASC) for the opt out, which captures differences in the mean of the distribution of the unobserved effects in the random component, ε_{njt} , between the opt out (conventional approach) and the other alternatives (biologic calculators). Equation (1) was estimated separately for respondents who received the plain-text version and the animated storyline.

In the standard conditional logit model, the estimated coefficients reflect both preference weights and the variance of the unobservable element of utility (the variance of ε_{nj}). This means that differences in estimated coefficients may be due to differences in preferences or differences in the relative variance of the error term (differences in “scale”). The variance of the error term can be interpreted as a measure of the randomness (or consistency) in choices. In this example, the consistency in choices may also depend on the training material received. To understand whether the choice behavior of respondents who received the plain-text version and those who received the animated storyline differed, a heteroskedastic conditional logit (HCL) model [40] was also estimated using the pooled data from both groups:

$$U_{nj} = \lambda_n \beta_0 + \lambda_n \beta_1 \text{Delay}_{n,j} + \lambda_n \beta_2 \text{PPV}_{n,j} + \lambda_n \beta_3 \text{NPV}_{n,j} + \lambda_n \beta_4 \text{Risk}_{n,j} + \lambda_n \beta_5 \text{Cost}_{n,j} + \varepsilon_{n,j}. \quad (2)$$

In the HCL model, the scale parameter, λ , a measure of error variance, was permitted to vary by the training materials received and is modelled as follows:

$$\lambda_n = \exp(\gamma \text{TEXT}_n), \quad (3)$$

where TEXT_n is equal to one when respondent n received plain text. Testing the significance of the parameter γ is therefore a test of whether the training materials affected choice consistency (the scale parameter, λ) [41].

The results of the conditional logit models were used to estimate marginal rates of substitution, which reveal the

amount of one attribute (e.g., time to starting treatment) individuals are, on average, willing to exchange for another (e.g., predictive value). The associated confidence intervals (CIs) for the marginal rates of substitution were estimated using the delta method [42]. Even if the results of the HCL model suggest significant scale heterogeneity, marginal rates of substitution (ratios of coefficients) are unaffected by the heteroskedasticity in the error.

3 Results

In total, 300 members of the public completed the DCE, with 158 people receiving training materials as plain text and 142 people receiving the animated storyline. Table 1 shows the sample characteristics for a number of key variables for all respondents and the two subsamples. Appendix C shows the results of a logistic regression model that suggests no observable characteristics predicted the training materials respondents received, which suggests that randomization worked. A total of 37 respondents clicked on the link and consented to the survey but did not complete the questions, and 28 participants left during the training materials; most of these [$n=23$ (82.1%)] were randomized to receive the animated storyline.

In this study, 16.7% ($n=50$) of the total sample failed the internal test for monotonicity. Slightly fewer failures were observed for respondents who received the animated storyline [$n=22$ (15.5%)] than for those who received the plain text [$n=28$ (17.7%)], but the difference was not statistically significant ($p=0.605$). All respondents, whether they failed or passed the monotonicity test, were included in the final analyses of the choice data. Appendix D contains the results of a split sample analysis for each version of training materials with (1) respondents who “passed” the internal validity test and (2) all respondents.

Table 1 also shows the responses to self-reported task difficulty and confidence in choices. When asked to rate their confidence on a scale of one to five, respondents who received the plain-text version of the training materials reported a lower average confidence score (mean 2.50) than those who received the animated storyline (mean 2.59), although this difference was not statistically significant ($p=0.404$). Similarly, when asked to self-report their ease of choice making on a scale of one to five, those who received the plain-text version reported a lower average ease score (mean 2.68) than those who received the animated storyline, although, again, this difference was not statistically significant ($p=0.353$).

The respondents who were randomized to the animated storyline spent on average of 50.1 s (95% CI 41.3–59.0) reading the training materials. Appendix E shows kernel density estimates for average time spent reading and clicking

Table 1 Summary of study sample characteristics

Characteristic	Overall	Plain text	Animated storyline
Age group (years)			
18–24	3 (1.0)	2 (1.3)	1 (0.7)
25–34	45 (15.0)	22 (13.9)	23 (16.2)
35–44	76 (25.3)	48 (30.4)	28 (19.7)
45–54	89 (29.7)	43 (27.2)	46 (32.38)
55–64	68 (22.7)	35 (22.2)	33 (23.2)
≥ 65	19 (6.3)	8 (5.1)	11 (7.8)
Female sex	114 (38.0)	52 (32.9)	62 (43.66)
Occupational status			
Employed full time	234 (78.3)	105 (74.5)	129 (81.7)
Employed part time	37 (12.4)	22 (15.6)	15 (9.5)
Self-employed	5 (1.7)	3 (2.1)	2 (1.3)
Unemployed	2 (0.7)	1 (0.7)	1 (0.6)
Retired	14 (4.7)	9 (6.4)	5 (3.2)
Looking after home/ family	4 (1.3)	1 (0.7)	3 (1.9)
Student	1 (0.3)	0 (0.0)	1 (0.6)
Freelance/temping	1 (0.3)	0 (0.0)	1 (0.6)
Long-term sickness	1 (0.3)	0 (0.0)	1 (0.6)
Choices in real life			
Very confident same	39 (13.0)	16 (10.1)	23 (16.2)
Quite confident same	127 (42.3)	75 (47.5)	52 (36.6)
Neither confident/not	75 (25.0)	44 (27.9)	31 (21.8)
Quite confident dif- ferent	49 (16.3)	18 (11.4)	31 (21.8)
Very confident dif- ferent	10 (3.3)	5 (3.2)	5 (3.5)
Task difficulty			
Very easy	32 (10.7)	18 (11.4)	14 (9.9)
Easy	109 (36.3)	60 (38.0)	49 (34.5)
Neither easy nor difficult	74 (24.7)	38 (24.1)	36 (25.4)
Difficult	78 (26.0)	39 (24.7)	39 (27.5)
Very difficult	7 (2.3)	3 (1.9)	4 (2.8)

Data are presented as *n* (%)

through the training materials for respondents who received the animated storyline. The average (mean) time spent on each choice set in the DCE, regardless of randomization, was 2.82 s (95% CI 1.65–3.99). Respondents randomized to the animated storyline completed the choice sets slightly more quickly (2.08 s; 95% CI 1.80–2.37) than those who received plain-text training materials (2.79 s; 95% CI 1.97–3.61), but this difference was not statistically significant ($p=0.124$). Appendix F shows kernel density estimates for average time spent completing a choice set by training materials received.

Figure 2 shows the difference in rates of self-reported ANA for each attribute. In all attributes except for cost, people who received the plain-text training materials were more likely to report ANA. Only 15 people (19.0%) who received the

animated storyline compared with 26 (42.9%) who received plain-text material reported ANA to the risk attribute.

3.1 Results of the Discrete-Choice Models

The estimated coefficients for all attributes had signs consistent with a priori expectations about the direction of impact of the attribute on preferences (Table 2). Respondents disliked increases in the delay to the start of treatment or risk of infections but liked increases in positive predictive values (PPVs) and negative predictive values (NPVs) and cost savings to the healthcare system. The ASC was large, negative, and statistically significant, suggesting that individuals derived utility from the “biologic calculator” over and above that derived from the attributes. Alternative specification of the utility function was investigated by including quadratic terms for each attribute, but no quadratic terms were statistically significant ($p>0.01$).

The presence of scale heterogeneity was confirmed in the estimated HCL model (Table 2). The estimated scale term of -0.216 was statistically significant ($p<0.01$), suggesting that the error variance differed between the two groups. The negative sign suggests that the scale parameter was smaller for the sample who received plain text, implying that this group had a larger error variance and were therefore, on average, less consistent than those who received the animated storyline. The scale parameter (exponential of the scale term of the HCL model) was estimated at 0.805.

To test whether the format of the training materials affected the choices and estimated preferences from the DCE, a likelihood ratio test was used to compare the conditional logit models from each sample and the HCL model. The likelihood ratio test suggested that, conditional upon there being differences in scale, the hypothesis of preference homogeneity could not be rejected ($p=0.282$), meaning there were no statistically significant differences in average preferences between the two groups. This is also illustrated in the marginal rates of substitution presented in Table 3. Further analyses on the sample who received the serious game found the respondents who spent longer reading the materials also made more consistent choices (see Appendix G).

4 Discussion

The aim of this study was to investigate whether, and how, the format of training information affected respondents’ ability to complete DCE choice sets rather than using training materials to prime or bias respondents. The results of this study found that preferences estimated from choice data collected in the DCE were, on average, unaffected by the format of the training materials. Importantly, there were statistically significant differences in the error variance between the sample of respondents randomized to receive plain-text rather

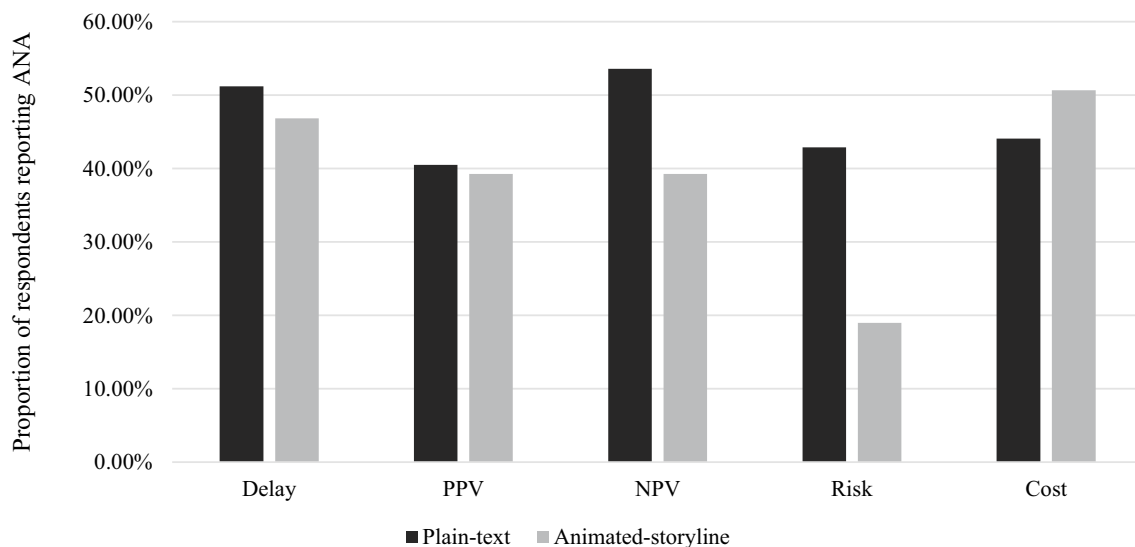


Fig. 2 Proportion of respondents self-reporting attribute non-attendance for each attribute by training materials received. *NPV* negative predictive value, *PPV* positive predictive value

Table 2 Pooled and split-sample estimates of discrete-choice data using different model specifications

Attribute	Animated storyline conditional logit	Plain text conditional logit	Heteroskedastic conditional logit
ASC (conventional prescribing)	-1.063*** (0.29)	-1.122*** (0.26)	-1.205*** (0.22)
Delay	-0.006 (0.00)	-0.010** (0.00)	-0.009** (0.00)
PPV ^a	0.100*** (0.01)	0.089*** (0.01)	0.104*** (0.01)
NPV ^a	0.153** (0.06)	0.073 (0.05)	0.126** (0.04)
Risk	-0.060*** (0.01)	-0.032** (0.01)	-0.051*** (0.01)
Cost ^b	0.003 (0.01)	0.019** (0.01)	0.012* (0.01)
Scale term (plain text)			-0.216** (0.08)
Log-likelihood	-650.54746	-805.78534	-1459.4602
Observations (<i>N</i>)	2556	2844	5400

Data are presented as estimated coefficient (standard error) unless otherwise indicated

ASC alternative-specific constant, *NPV* negative predictive value, *PPV* positive predictive value

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

^aAttribute scaled so 1% = 10% so coefficients represent the effect of a 10% change in the predictive value

^bAttribute scaled so £1 = £100 so coefficient represents the effect of a £100 change in the cost saving

than animated-storyline training materials. In other words, the respondents randomized to read the animated storyline were less affected by factors that are unobservable (such as omitted attributes), which resulted in their choices appearing to be statistically significantly less random than those of respondents who received the plain-text training materials. A potential interpretation of this result could be that the respondents were better informed when the information in the training materials was conveyed in an animated storyline.

The observed difference in error variance identified in this research has implications for researchers conducting DCEs in complex areas that require many attributes, alternatives, or choice sets or those with a small population of

interest. Reduced error variance could be a signal of reduced cognitive burden, implying that those who received the interactive training materials could answer more, or more difficult, choice sets.

The rate of self-reported ANA to all attributes, except cost, was lower for respondents who received the animated storyline. The difference was most notable in the more complex attributes that involved some element of risk, such as NPV and PPV. A wealth of literature suggests that, even in well-educated populations, people find risk and probabilities hard to comprehend [43–45]. Evidence also suggests that risk is not always communicated effectively in healthcare DCEs [46]. The finding that fewer respondents

Table 3 Marginal rates of substitution

Unit of exchange	Training material	For a £100 saving	For a 10% increase in PPV	For a 10% increase in NPV	For a 1% decrease in risk	For a 1-day reduction in the delay
Willingness to delay treatment (days)	Storyline	0.45 ^{a,b} (− 1.82 to 2.71)	15.45 ^a (− 5.50 to 36.4)	23.64 ^a (− 15.31 to 62.60)	9.29 ^a (− 3.42 to 21.99)	–
	Plain text	1.86 (0.13 to 3.60)	8.68 (2.11 to 15.25)	7.08 ^b (− 5.29 to 19.45)	3.08 (0.13 to 6.02)	–
Willingness to accept risk (%)	Storyline	0.14 ^b (− 0.02 to 0.31)	1.66 (0.92 to 2.41)	2.55 (0.33 to 4.76)	–	0.11 ^b (− 0.04 to 0.25)
	Plain text	0.61 (0.22 to 1.19)	2.82 (0.74 to 4.90)	2.30 ^b (− 1.59 to 6.19)	–	0.32 (0.01 to 0.64)

Figures in parentheses are 95% confidence intervals

NPV negative predictive value, PPV positive predictive value

^aDenominator not statistically significant

^bNumerator not statistically significant

reported ANA could suggest an increased understanding of these attributes when training materials were presented as an animated storyline.

This study focused on training materials that were used to inform a DCE, but the findings may also be relevant to researchers using other stated-preference methods (such as time trade-off, standard gamble, and contingent valuation) in either health or non-health settings, such as studies eliciting preferences for environmental goods or services [47]. Stated-preference surveys are increasingly digitized [4], suggesting that scope exists for interactive or more engaging materials. Using an animated storyline or a more sophisticated serious game may be useful when eliciting preferences for complex interventions or when current practice is difficult to explain. A recent systematic review of DCEs showed a rise in the number of studies conducted in lower-income or developing countries [4], where literacy rates might be lower or the subject matter less familiar to respondents, particularly if access to healthcare is low. Furthermore, researchers are also using stated-preference methods to elicit utilities for health states from challenging samples such as children [48] or those with conditions related to cognitive impairment [49]. Using interactive materials that do not overwhelm the respondent with text but also do not change preferences may be a way to improve survey respondents' understanding and thus the confidence of researchers and policy makers in the derived valuations. This study adds to the growing body of evidence [50] on the value of using this theory-driven, evidence-based approach to developing health-related educational applications that are able to realize the desired outcomes.

Researchers have also used other methods to reduce hypothetical bias in healthcare DCEs, including “cheap talk” and “time to think.” In “cheap talk,” respondents are led through a script explaining hypothetical bias and its consequences in economic valuation [51]. In “time to think,” respondents are encouraged to deliberate before stating their choices

[52]. The ability of “cheap talk” and “time to think” methods to reduce hypothetical bias is debatable [53–58], but these approaches may be used in addition to serious games to improve the validity of the elicited stated preferences.

This study represents a preliminary investigation into the potential influence of the format of training materials on respondents' choices and response efficiency. Future research may wish to consider more complex econometric models to understand how personality traits or attitudes may impact the influence of training materials. However, it has been noted that considering preferences and attitudes simultaneously introduces an endogeneity issue: as they are latent, they could be jointly correlated with unobservable factors [59]. Researchers employing serious games or interactive survey materials may wish to consider allowing respondents to choose the format of the information they receive.

This study used an online panel to recruit members of the public to complete the DCE. The sampling approach meant it was not feasible to conduct a study to gain qualitative insights into respondents' views about the format of training materials [60, 61]. Given the observed dropout rate of respondents randomized to receive the animated storyline, it is possible that the animated storyline acted as a filter to remove less serious or inattentive respondents. If more attentive respondents watched the storyline materials, the difference in error variance could be attributed to sample selection rather than improvements in communication. The mechanism by which the animated storyline improved choice consistency requires further research. Further research may wish to use an alternative recruitment strategy to enable the collection of qualitative data to illuminate why an animated storyline, specifically, or serious games, more generally, appear to influence response efficiency but not observed preferences.

The largest limitation of this study relates to the sample size, which limited our ability to understand two key aspects: self-reported difficulty/confidence and speed of completion.

Although differences in self-reported difficulty/confidence and failure of the internal validity test were identified, these differences were not statistically significant in our sample. Respondents who received the animated storyline also answered the choice questions slightly quicker than those who received the plain text, but this difference was not statistically significant, which may also be an artefact of the sample size. It should also be noted that response time was also automatically recorded using the page timer in Sawtooth software and may reflect differences in browsers or computing power and thus should not be compared with other studies with different recording methods. As effect sizes were unknown in advance of the research, no power calculation was conducted to detect differences between the survey versions. Instead, the sample of 150 participants was based on estimating preference coefficients. Further research with a larger sample size is warranted to understand whether there is a statistical difference and whether the speed of completion is due to better acquisition or retention of information in the survey. Future research may seek to compare investing in training materials for the survey to other aspects of the study design, e.g., increasing the sample size or pre-testing to acquire priors.

5 Conclusion

This study found that providing animated information about the disease area and intervention being valued had a positive influence on the quality of choice data collected in a DCE in terms of the variance of the error term. The results of this study may have particular relevance to researchers conducting surveys about complex issues or those completed by small sample sizes. Stated-preference researchers should pay close attention and carefully develop the training materials to ensure respondents can make informed decisions when presented with the subsequent valuation exercise, such as a choice set. Researchers reporting the results of stated-preference methods should provide survey materials in online appendices so readers may consider them alongside the study findings. Further research is required to establish the generalizability of these results in a larger sample size, and in other settings, using alternative stated-preference methods, or for specific subgroups of respondents.

Acknowledgements The authors are grateful for feedback received from participants of the MATURA Patient Advisory Group (MPAG), the Leeds Musculoskeletal Biomedical Research Unit (LMBRU) arthritis patient involvement group, and the National Rheumatoid Arthritis Society (NRAS) support group in Glasgow. The authors are also grateful for feedback received when this research was presented at the Society for Medical Decision Making's European Meeting and the Academic Unit of Health Economics at the University of Leeds.

Compliance with Ethical Standards

Funding CV completed this work while based at The University of Manchester for the project “Mind the Risk” funded by Riksbankens Jubileumsfond. ND completed this work while based at The University of Manchester. Katherine Payne received financial support for completing this work for the project “Mind the Risk” funded by Riksbankens Jubileumsfond and for the project “Maximising Therapeutic Utility in Rheumatoid Arthritis” (MATURA) funded by the Medical Research Council (Grant ref MR/K015346/1). MindBytes (Ghent, Belgium) were paid a fee to animate a predefined storyline for the training materials. The views expressed in the manuscript are those of the authors and not necessarily those of the funding bodies.

Ethics This paper was prepared in compliance with ethical standards. Ethical approval for the study was granted by the University of Manchester's Research Ethics Committee.

Conflict of interest CV, NJD, GVS, and KP have no conflicts of interests that are directly relevant to the content of this article.

Data availability The datasets generated during and/or analysed during the current study are still undergoing further analyses and are unavailable.

Open Access This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (<http://creativecommons.org/licenses/by-nc/4.0/>), which permits any noncommercial use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Louviere J, Hensher D, Swait J. Stated choice methods: analysis and application. Cambridge: Cambridge University Press; 2000.
2. Smith RD, Sach TH. Contingent valuation: what needs to be done? Health Econ Policy Law. 2010;5:91–111.
3. Carson RT, Louviere JJ. A common nomenclature for stated preference elicitation approaches. Environ Resour Econ. 2011;49:539–59.
4. Soekhai V, de Bekker-Grob EW, Ellis AR, Vass CM. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics. 2019;37:201–26.
5. Boxall PC, Adamowicz W, Swait J, Williams M, Louviere J. A comparison of stated preference methods for environmental valuation. Ecol Econ. 1996;18:243–53.
6. Ozdemir S. Improving the validity of stated-preference data in health research: the potential of the time-to-think approach. Patient. 2015;8:247–55.
7. Carlsson F, Martinsson P. Do hypothetical and actual marginal willingness to pay differ in choice experiments? J Environ Econ Manag. 2001;41:179–92.
8. Muhlbacher AC, Juhnke C, Beyer AR, Garner S. Patient-focused benefit–risk analysis to inform regulatory decisions: the European Union perspective. Value Health. 2016;19:734–40.
9. Vass CM, Payne K. Using discrete choice experiments to inform the benefit–risk assessment of medicines: are we ready yet? Pharmacoeconomics. 2017;35:1–21.

10. Reed SD, Lavezzari G. International experiences in quantitative benefit–risk analysis to support regulatory decisions. *Value Health*. 2016;19:727–9.
11. Bridges JF, Hauber AB, Marshall D, Lloyd A, Prosser L, Regier DA, et al. Conjoint analysis applications in health—a checklist: a report of the ISPOR Good Research Practices for Conjoint Analysis Task Force. *Value Health*. 2011;14:403–13.
12. Lancsar E, Louviere J. Conducting discrete choice experiments to inform healthcare decision making: a user’s guide. *Pharmacoeconomics*. 2008;26:661–77.
13. Coast J, Al-Janabi H, Sutton E, Horrocks SA, Vosper J, Swancutt DR, et al. Using qualitative methods for attribute development for discrete choice experiments: issues and recommendations. *Health Econ*. 2012;21:730–41.
14. Johnson F, Lancsar E, Marshall D, Kilambi V, Mulbacher A, Regier D, et al. Constructing experimental designs for discrete-choice experiments: report of the ISPOR conjoint analysis experimental design good research practices task. *Value Health*. 2013;16:3–13.
15. Louviere J, Lancsar E. Choice experiments in health: the good, the bad, the ugly and toward a brighter future. *Health Econ Policy Law*. 2009;4:527–46.
16. Hauber AB, González JM, Groothuis-Oudshoorn CGM, Prior T, Marshall DA, Cunningham C, et al. Statistical methods for the analysis of discrete choice experiments: a report of the ISPOR Conjoint Analysis Good Research Practices Task Force. *Value Health*. 2016;19:300–15.
17. Michie S, van Stralen MM, West R. The behaviour change wheel: a new method for characterising and designing behaviour change interventions. *Implement Sci*. 2011;6:42.
18. Wouters P, van Nimwegen C, van Oostendorp H, van der Spek ED. A meta-analysis of the cognitive and motivational effects of serious games. *J Educ Psychol*. 2013;105:249–65.
19. Louviere J. What you don’t know might hurt you: some unresolved issues in the design and analysis of discrete choice experiments. *Environ Resour Econ*. 2006;34:173–88.
20. Urban GL, Hauser JR, Roberts JH. Prelaunch forecasting of new automobiles. *Manag Sci*. 1990;36:401–21.
21. Weinberg B. An information-acceleration-based methodology for developing preproduction forecasts for durable goods: design, development, and initial validation. Massachusetts Institute of Technology, Sloan School of Management; 1993. <https://dspace.mit.edu/handle/1721.1/12547>
22. Michael DR, Chen SL. Serious games: games that educate, train, and inform. New York: Muska and Lipman/Premier-Trade; 2005.
23. Annetta LA, Minogue J, Holmes SY, Cheng M-T. Investigating the impact of video games on high school students’ engagement and learning about genetics. *Comput Educ*. 2009;53:74–85.
24. Bai H, Pan W, Hirumi A, Kebritchi M. Assessing the effectiveness of a 3-D instructional game on improving mathematics achievement and motivation of middle school students. *Br J Educ Technol*. 2012;43:993–1003.
25. Garris R, Ahlers R. Games, motivation, and learning: a research and practice model. *Simul Gaming*. 2002;33:441–67.
26. Clark M, Determann D, Petrou S, Moro D, de Bekker-Grob EW. Discrete choice experiments in health economics: a review of the literature. *Pharmacoeconomics*. 2014;32:883–902.
27. Choice Metrics. Ngenue User Manual 1.1.1; 2012.
28. Sawtooth. Sawtooth Software SSI Web 8.3.8; 2012.
29. Verschuere S, Buffel C, Vander Stichele G. Developing theory-driven, evidence-based serious games for health: framework based on research community insights. *JMIR Serious Games*. 2019;7:e11565.
30. Reeve C. Narrative-based serious games. In: *Serious games on the move*; 2009. p. 73–89.
31. Nowak KL, Rauh C. The influences of the avatar on online perceptions of anthropomorphism, androgyny, credibility, homophily, and attraction. *J Comput Commun*. 2006;11:153–78.
32. Zikmund-Fisher B, Ubel PA, Smith DM, Derry HA, McClure JB, Stark A, et al. Communicating side effect risks in a tamoxifen prophylaxis decision aid: the debiasing influence of pictographs. *Patient Educ Couns*. 2008;73:209–14.
33. Feldman-Stewart D, Brundage MD. Challenges for designing and implementing decision aids. In: *Patient education and counseling*, vol. 54; 2004. p. 265–273.
34. Bateman C. Diversity in game narrative. *Only a game*; 2005. http://onlyagame.typepad.com/only_a_game/2005/08/diversity_in_ga.html.
35. Callaghan M, Savin-Baden M, McShane N, Gomez Eguiluz A, Arnab S, Lim T, et al. Mapping learning and game mechanics for serious games analysis. *Br J Educ Technol*. 2015;46:391–411.
36. Devlin N, Shah K, Feng Y, Mulhern B, van Hout B. Valuing health-related quality of life: an EQ-5D-5L value set for England. *Health Econ*. 2017;27:7–22.
37. Thurstone L. A law of comparative judgment. *Psychol Rev*. 1927;34:273–86.
38. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. *Frontiers in econometrics*. Academic Press: New York; 1973. pp. 105–142.
39. van der Pol M, Currie G, Kromm S, Ryan M. Specification of the utility function in discrete choice experiments. *Value Health*. 2014;17:297–301.
40. Hole AR. Small-sample properties of tests for heteroscedasticity in the conditional logit model. *Econ Bull*. 2006;3:1–14.
41. Vass CM, Wright S, Burton M, Payne K. Scale heterogeneity in healthcare discrete choice experiments: a primer. *Patient*. 2018;11:167–73.
42. Hole AR. A comparison of approaches to estimating confidence intervals for willingness to pay measures. *Health Econ*. 2007;16:827–40.
43. Lipkus I, Samsa G, Rimer B. General performance on a numeracy scale among highly educated samples. *Med Decis Mak*. 2001;21:37–44.
44. Gigerenzer G, Hertwig R, van den Broek E, Fasolo B, Katsikopoulos KV. “A 30% chance of rain tomorrow”: how does the public understand probabilistic weather forecasts? *Risk Anal*. 2005;25:623–9.
45. Vass C, Rigby D, Payne K. “I Was Trying to Do the Maths”: exploring the impact of risk communication in discrete choice experiments. *Patient*. 2019;12:113–23.
46. Harrison M, Rigby D, Vass CM, Flynn T, Louviere JJ, Payne K. Risk as an attribute in discrete choice experiments: a systematic review of the literature. *Patient*. 2014;7:151–70.
47. Rakotonarivo OS, Schaafsma M, Hockley N. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. *J Environ Manag*. 2016;183:98–109.
48. Ratcliffe J, Couzner L, Flynn T, Sawyer M, Stevens K, Brazier J, et al. Valuing Child Health Utility 9D health states with a young adolescent sample: a feasibility study to compare best-worst scaling discrete-choice experiment, standard gamble and time trade-off methods. *Appl Health Econ Health Policy*. 2011;9:15–27.
49. Bridges JFP, Kinter ET, Schmeding A, Rudolph I, Mühlbacher A. Can patients diagnosed with schizophrenia complete choice-based conjoint analysis tasks? *Patient*. 2011;4:267–75.
50. Buffel C, van Aalst J, Bangels A-M, Toelen J, Allegaert K, Verschuere S, et al. CliniPup®, a web-based serious game for health to reduce perioperative anxiety and pain in children: a pilot study. *JMIR Serious Games*. 2019;7:e12431.

51. Cummings RG, Taylor LO. Unbiased value estimates for environmental goods: a cheap talk design for the contingent valuation method. *Am Econ Rev.* 1999;89:649–65.
52. Rigby D, Vass CM, Payne K. Opening the “Black Box”: an overview of methods to investigate the decision-making process in choice-based surveys. *Patient.* 2019. <https://doi.org/10.1007/s40271-019-00385-8>.
53. Ozdemir S, Johnson FR, Hauber AB. Hypothetical bias, cheap talk, and stated willingness to pay for health care. *J Health Econ.* 2009;28:894–901.
54. Whittington D, Smith VK, Okorafor A, Liu JL, McPhail A. Giving respondents time to think in contingent valuation studies: a developing country application. *J Environ Econ Manag.* 1992;22:205–25.
55. Bosworth R, Taylor LO. Hypothetical bias in choice experiments: Is cheap talk effective at eliminating bias on the intensive and extensive margins of choice? *BE J Econ Anal Policy* 2012;12.
56. Alpizar F, Carlsson F, Johansson-Stenman O. Does context matter more for hypothetical than for actual contributions? Evidence from a natural field experiment. *Exp Econ.* 2008;11:299–314.
57. Silva A, Nayga RM, Campbell BL, Park JL. Can perceived task complexity influence cheap talk’s effectiveness in reducing hypothetical bias in stated choice studies? *Appl Econ Lett.* 2012;19:1711–4.
58. Ready RC, Champ PA, Lawton JL. Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. *Land Econ.* 2010;86:363–81.
59. Hess S, Beharry-Borg N. Accounting for latent attitudes in willingness-to-pay studies: the case of coastal water quality improvements in Tobago. *Environ Resour Econ.* 2012;52:109–31.
60. Vass CM, Rigby D, Payne K. The role of qualitative research methods in discrete choice experiments: a systematic review and survey of authors. *Med Decis Mak.* 2017;37:298–313.
61. Ryan M, Watson V, Entwistle V. Rationalising the “irrational”: a think aloud study of a discrete choice experiment responses. *Health Econ.* 2009;18:321–36.