The European Journal of Health Economics (2018) 19:1053–1066
https://doi.org/10.1007/s10198-018-0954-6

ORIGINAL PAPER

How well do discrete choice experiments predict health choices? A systematic review and meta-analysis of external validity

Matthew Quaife[1] · Fern Terris-Prestholt[1] · Gian Luca Di Tanna[2] · Peter Vickerman[3]

Received: 28 June 2017 / Accepted: 8 January 2018 / Published online: 29 January 2018

Abstract

Discrete choice experiments (DCEs) are economic tools that elicit the stated preferences of respondents. Because of their increasing importance in informing the design of health products and services, it is critical to understand the extent to which DCEs give reliable predictions outside of the experimental context. We systematically reviewed the literature of published DCE studies comparing predictions to choices made in reality; we extracted individual-level data to estimate a bivariate mixed-effects model of pooled sensitivity and specificity. Eight studies met the inclusion criteria, and six of these gave sufficient data for inclusion in a meta-analysis. Pooled sensitivity and specificity estimates were 88% (95% CI 81, 92%) and 34% (95% CI 23, 46%), respectively, and the area under the SROC curve (AUC) was 0.60 (95% CI 0.55, 0.64). Results indicate that DCEs can produce reasonable predictions of health-related behaviors. There is a great need for future research on the external validity of DCEs, particularly empirical studies assessing predicted and revealed preferences of a representative sample of participants.

Keywords Discrete choice experiment · External validity · Hypothetical bias · Meta-analysis

JEL Classification C590 · I100 · C830

Introduction

Discrete choice experiments (DCEs) ask participants to make choices between hypothetical alternatives, using choice modeling methods to analyze data. They are attractive tools for research and policy as they offer a flexible methodology to estimate which attributes are important in decision-making. Participants are asked to choose their preferred of, generally, between two and five alternatives over a number of choice tasks (usually around ten). Data are analyzed using discrete choice models [1], the results of which can be used to determine the relative importance of different attributes to respondent choices. Results can also be used to predict demand, termed market shares in the marketing literature [2]. DCEs are being increasingly used in health to study patient and/or physician preferences in academic studies, health technology assessments, and regulatory risk–benefit assessment [3, 4]. Particularly useful is the ability to include products or attributes in DCEs which do not exist in reality, and for which no observational or trial data exist [5].

In recent years, DCEs have proven increasingly popular in the health domain, whilst a large (mostly non-health) economic literature has developed around the design, analysis, and application of DCEs [2, 6, 7]. Yet DCEs ask respondents to make hypothetical choices, and it is important to ensure that they measure what researchers think that they do. Disparities between revealed and stated preference data are, in part, due to the hypothetical nature of DCE tasks; this divergence is termed *hypothetical bias*. There has been no widely accepted theory to explain hypothetical bias which, following Beck et al. [8], we define in this paper as discrepancies

Electronic supplementary material The online version of this article (https://doi.org/10.1007/s10198-018-0954-6) contains supplementary material, which is available to authorized users.

✉ Matthew Quaife
matthew.quaife@lshtm.ac.uk

[1] Department of Global Health and Development, London School of Hygiene and Tropical Medicine, London, UK

[2] Centre for Primary Care and Public Health, Queen Mary University of London, London, UK

[3] School of Social and Community Medicine, University of Bristol, Bristol, UK

between preferences exhibited in DCEs to those exhibited in reality. Hypothetical bias may originate when choice tasks do not fully reflect reality in the nature or characteristics of choices, when respondents have incomplete preferences, or if respondents perceive a vested interest in over- or understating the importance of particular attributes.

Where the true attributes of a good or service are known, predicted probability analysis (PPA) uses DCE results to predict utility-maximizing choices. Under the assumption that individuals are rational, DCEs can approximate which choice people would make given the option in reality [9, 10]. PPA is common outside of health and is increasingly being applied by health economists to predict demand for a range of health-related choices, including HIV prevention products [11], contraceptive services [12], vaccination [13], and migraine treatments [9].

In the early stages of introducing health products or services, there is often a great deal of uncertainty around their potential impact and subsequent cost-effectiveness. DCE predictions can be particularly useful for estimating the uptake of new products or services where observational data, from trials or pilot projects, are not available [12, 13]. In the absence of observational data, for example in from a pragmatic trial or demonstration project, "expert opinion" is often used to generate uptake predictions, which are then commonly used to inform impact or cost-effectiveness models [14]. In a previous paper, the authors of this study suggest that DCEs can provide a useful empirical alternative to expert opinion, whist also offering additional benefit through accounting for synergistic relationships between different product and service attributes and use, which are commonly ignored [15].

Background on validity in DCEs

In health, around 60% of studies include internal theoretical validity checks, whilst non-satiation and transitivity tests are applied less frequently [3]. A recent study demonstrated that, of the 112 health DCEs published in 2015, 49% included at least one internal validity check, yet there were substantial differences in how researchers dealt with indications of poor validity, i.e., 46% of studies using a dominance test excluded respondents who failed the test [16].

By contrast, *external validity* is concerned with ensuring the comparability of hypothetical and actual choices. Because respondents are not obliged in reality to make the choices they indicate in a DCE, hypothetical bias may reduce the usefulness of DCE results [17, 18]. There have been some substantive contributions to the methodological literature on the external validity of DCEs (e.g., [19–22]), yet much of the focus in the literature has been on maximizing the internal validity of DCEs. Whilst this is important, there has been very little empirical work assessing whether choices made in DCEs in fact reflect those made in reality, or the circumstances in which they may offer more or less reliable inference [19].

This paper considers variations in external validity in health DCEs attributable to hypothetical bias, and is the first systematic review and meta-analysis assessing the ability of DCEs to predict health behaviors. Proving external validity is important to the practical application and use of DCEs, yet despite their growing popularity, there has been little research on their external validity in the health domain [21, 22].

There are many reasons why hypothetical bias may exist, including that people may be fundamentally rational but inconsistent in utility maximization, e.g., when they are paying more or less attention to a decision context. Indeed, the choice architecture surrounding real-world and hypothetical choices has been shown to affect choices, including healthy eating, physical activity, and alcohol use [23–25]. Furthermore, if DCEs are not incentive compatible, respondents may try and answer strategically, for example to understate their willingness to pay for public services [26, 27]. Behavioral economic research challenges the theory that we are all variants of *homo economicus*, but there has been limited work exploring if behavioral heuristics influence preferences to a greater or lesser extent in stated preference tasks than in reality [28, 29]. Additional reasons why differences may exist in health fields include: difficulties in acquiring revealed preference data due to failures in healthcare markets; the lack of a market analogue for many health decisions; or vested interests from researchers not wanting to reveal the ability (or lack thereof) of DCE models to accurately predict choices and behavior [20, 22, 30, 31].

We note that literature exists, mostly outside of a health context, evaluating the external validity of WTP estimates from stated preference exercises. These exercises are different from DCEs whereby, instead of participants choosing between a set of alternatives, people are presented with open-ended questions, for example how much they would be willing to pay to avoid a certain occurrence. The external validity of WTP estimates has been explored fully and recently in the literature, and we therefore do not include WTP studies in this review. This work explores the external validity of willingness-to-pay estimates from contingent valuations and conjoint analyses, mostly in transport fields [32–34], including three meta-analyses [35–37]. These studies conclude that hypothetical bias can be substantial, with median bias levels ranging from 25 to 300%. Furthermore, efforts to reduce hypothetical bias through a number of methods, such as increasing consequentiality of choices or "cheap-talk" strategies, have also been shown to have mixed results [25, 38–40].

Rationale for review and aim

There has been no synthesis of the predictive abilities of DCEs in health, despite a substantive and recent increase in the number of studies using estimated choice probabilities from DCEs to predict choices, e.g., [9, 41]. These studies implicitly assume that DCEs have sufficient external validity to provide meaningful results. More generally, there has been no systematic review of studies exploring the external validity of any stated preference tasks in health. Existing reviews focus on summarizing DCE applications [3, 7, 42], collating preference research on particular health or disease areas [43, 44], or synthesizing methodological innovations to maximize internal validity [45].

This review aims to systematically review studies comparing stated preference choices, as modeled through predicted probability models resultant from DCE data, to revealed preference choices as gathered through observational or survey means. We report published evidence, describe the quality of included studies using an adapted quality checklist, and quantitatively synthesize the predictive ability of DCEs.

This review is the first to systematically evaluate and synthesize studies which observe participants' stated and revealed preferences through comparing DCE data to real-life health choices. Its findings will enable researchers and policymakers to assess how useful DCEs might be in predicting individual choices.

Methods

Search strategy

We searched the following databases to ensure a comprehensive exploration of the health, economic, and decision science literature: (1) PubMed/Medline; (2) EMBASE; (3) CINAHL; (4) Econlit; (5) Social Policy and Practice; (6) Science Direct. An iterative strategy was employed and the references of identified articles examined by hand for further relevant material. The search included all available years up to August 2015. The following keywords (alongside relevant MeSH terms where databases permitted) were used to build the search strategy:

- Discrete choice experiments ("discrete choice* OR choice experiment* OR stated preference* OR DCE OR conjoint analysis")
  AND
- External validity ("external validity OR predict* OR hypothetical bias* OR market share* OR revealed preference*")

Inclusion and exclusion criteria

We included studies using a discrete choice experiment methodology to predict health-related choices and compared these predictions with observed choices in real life. Studies were not excluded by population or the nature of choice tasks presented to participants. Studies obtaining revealed preference data from lab-based studies were excluded as these may be subject to similar concerns over external validity as DCEs themselves. The term conjoint analysis was explicitly included in the search strategy, and studies incorrectly labeled as conjoint analyses, when they were in fact DCEs, were included in the screening process. We excluded:

1. Studies using a preference elicitation method other than a discrete choice experiment (for example, contingent valuation or conjoint analyses);
2. Studies using "lab-based" experiments to elicit revealed preferences;
3. Letters, general commentaries or perspectives;
4. Studies without English language titles or abstracts.

Screening of studies for eligibility

We imported all identified references into reference management software [46] and removed duplicates. First, titles and abstracts were screened by one researcher and irrelevant articles excluded. Secondly, the full text of selected papers was screened independently by two researchers against the eligibility criteria. Any disagreements were resolved by discussion. Records of studies were kept as per the Preferred Reporting Items for Systematic Reviews and Meta-Analysis (PRISMA) checklist [47]. Data were then extracted into a predefined extraction form. Where information was lacking, attempts were made to contact corresponding authors to obtain the maximum quantity of data.

Assessment of study quality

The characteristics of included studies were assessed by two reviewers against a tool of 17 criteria based on that of Mandeville et al. [42], which is itself an adaptation of the "good-practice" checklist of Lancsar and Louviere [48]. The quality criteria tool used is presented in supplementary material S1. We remove the criteria of using an efficient design, since this may not be an indicator of study quality, and the criteria of ensuring a sufficient response rate due to the subjectivity of what may be considered *sufficient*. In addition, we amend the criteria that attribute and level choice should be "grounded in *qualitative* work with the target population", to "grounded in *piloting* work with the target population". Mandeville et al.'s method only assesses criteria which may substantively affect the quality of included studies, thereby

avoiding common criticisms of quality checklists that they are poorly correlated with study validity and measure the quality of reporting rather than that of the underlying study [49].

The checklist of Lancsar and Louviere does not consider external validity, except through the broad question "Was internal or external validity investigated?". Therefore, we drop this criterion in favor of five further criteria to assess the reliability of external validity assessment. We based four of these criteria on Lancsar and Swait [50] who specify some testable reasons that stated preference models might fail to be externally valid. Finally, we note the potential for selection bias in studies where observational data were gathered on a non-random subset of DCE participants and include an additional criterion to ensure that we account for potential selection bias in comparisons between predicted and actual choices. Both reviewers independently evaluated the quality of included studies by assessing whether the criterion for each study was met or not. If the information available for a criterion was insufficient to evaluate its achievement, we noted this as a separate category.

Statistical analysis

Because DCE predictions are a form of binary classification test, to synthesize the outcomes of included studies, we employ the array of methods used in assessing clinical diagnostic tests. In the context of DCE predictions, high sensitivity (true-positive rate) would indicate reliability in predicting opting-in behaviors. We define an opting-in behavior as a choice, in the DCE or a real-world context, to use a product or service that a respondent does not currently use. High specificity (true-negative rate) would indicate reliability in predicting opting-out behaviors, which we define as a respondent choosing not to use a product or service in the DCE or real-world context.

Synthesizing sensitivity and specificity estimates requires more sophistication than other quantitative syntheses, due to between-study heterogeneity, and the correlation between sensitivity and specificity estimates. When differences between studies are thought to be only due to sampling variation, it would be appropriate to pool estimates though sample size-weighted averages of sensitivity and specificity. However, it is likely that variability beyond chance can be attributed to between-study differences (e.g., study design, method of data collection, context, interviewer, or self-administration). Due to the range of DCE methods and study contexts, we use a random effects model to attempt to account for explainable and unexplainable heterogeneity [51].

There is likely to be interdependence between sensitivity and specificity measures, which requires specific consideration in meta-analytic models [52]. To account for this, we use bivariate mixed-effects logistic regression through the *midas* command in STATA 14, which assumes independent binomial distributions for true positives and negatives conditional on the sensitivity and specificity in each study [53, 54]. By jointly modeling sensitivity and specificity, this method preserves the bivariate data structure of the data and is an improvement on the standard analysis method of applying the DerSimonian and Laird random effects model [51]. The potential for publication bias was assessed through Egger's test [55]. No test for publication bias is without methodological issue [56], yet Egger's test for funnel plot asymmetry is recommended for use by PRISMA guidelines [57].

Finally, we conduct bivariate multivariable meta-regression to explore heterogeneity by regressing sensitivity and specificity estimates on five study-level characteristics: mode of DCE administration (paper/computer), whether the DCE is related to a prevention or treatment choice, the number of DCE choice sets, the number of alternatives within a choice set, and the percentage of DCE respondents for whom revealed preference data were analyzed. The effect of each covariate on sensitivity and specificity is estimated separately [53].

Results

Figure 1 details the flow of papers through the study. In total, 6383 studies were identified through database searching and one additional study identified through its presentation at a health economics conference. The full text of 13 articles was reviewed for eligibility, eight were considered for a qualitative synthesis, and six included sufficient quantitative information for inclusion in a meta-analysis. Figure 2 shows publications identified by year.

Notably, there were very few studies which directly assessed the external validity of DCE predictions. Of the studies that have been published (two were found in pre-publication, conference abstract stage, one of which was published whilst this study was under review [58]), five (63%) were published since 2015, suggesting that the external validity of DCEs is receiving more attention than in the past. Seven studies (88%) sought to predict the choices of patients over a broad range of health choices, from vaccination to sexually transmitted infection testing, to prospective mother's choice of birth location. One study sought to predict the choices of healthcare professionals as they appraised new medicines. Six studies (75%) presented information at the individual level and are included in the meta-analysis.

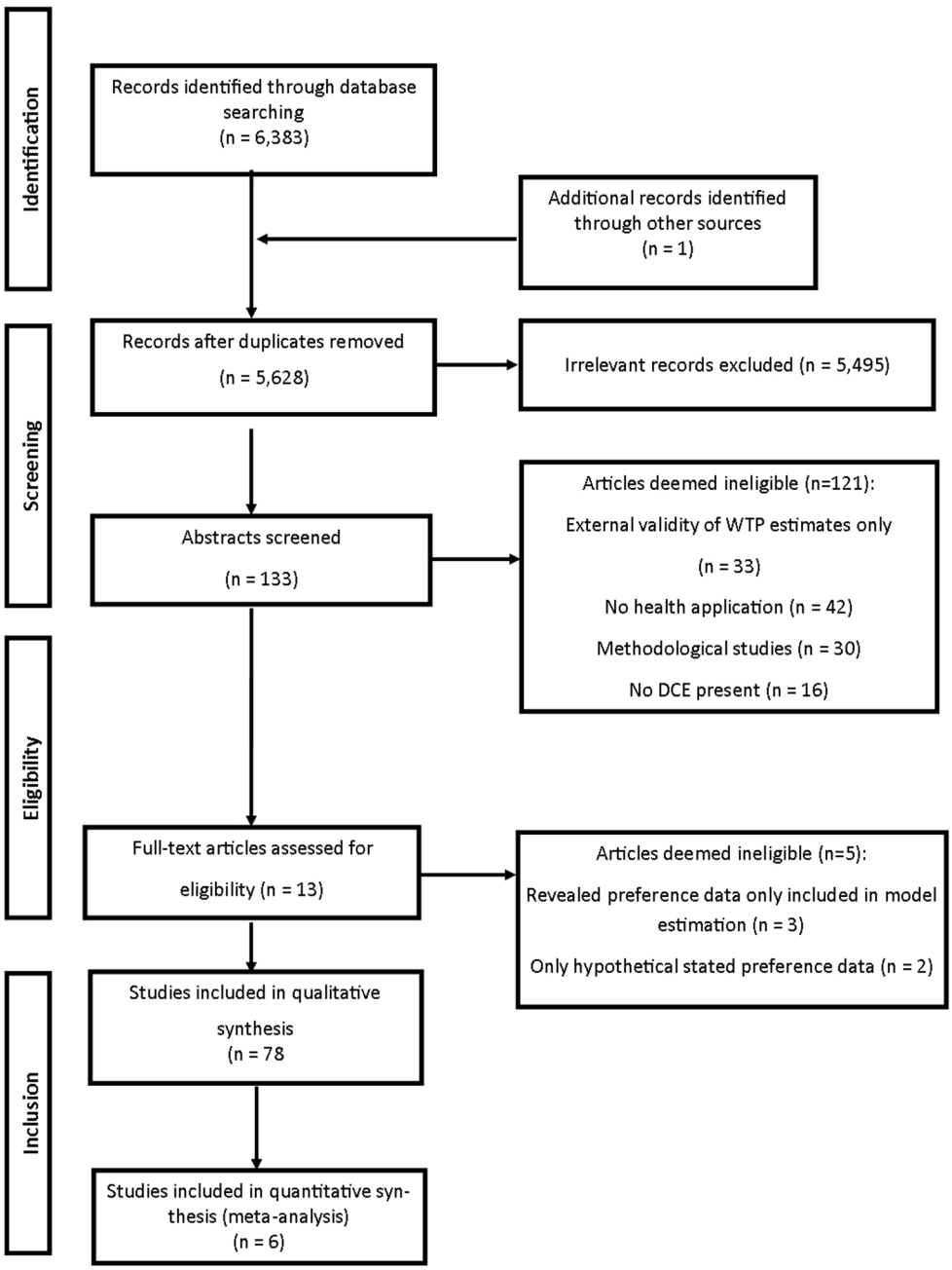

Fig. 1 PRISMA diagram of review process

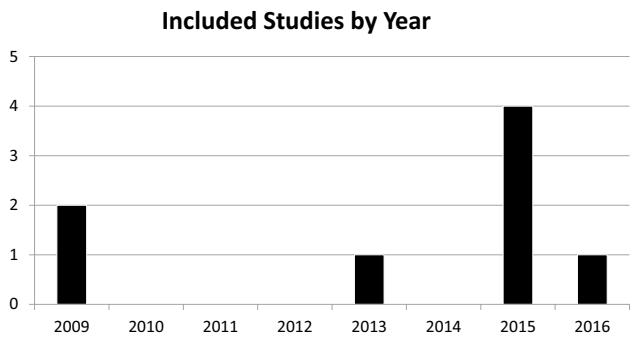

Fig. 2 Summary of publication date of included studies

Assessment of quality of included studies

Results from the quality assessment of included studies are presented in supplementary material S2. Overall, the quality of studies was high; the design and implementation of DCEs were often of good quality. However, the additional criteria we add to explore the robustness of study external validity indicate some notable weaknesses.

When assessed by criteria outside of those exploring external validity, the included studies were of high quality. For example, most DCEs were piloted in a relevant target population and allowed participants to "opt-out" of making

a choice. However, included studies may be subject to selection bias. The response rate of the DCE task was often low, while data on actual behavior was not gathered for the full sample of people who completed the DCE. If participants who did not complete the DCE, or who did and were not followed-up, are non-randomly different from those who were included in final analyses, systematic bias would be introduced into results. If participants were more likely to be included in follow-up when they opted-into a choice, this would overstate the predictive ability of DCEs.

We note that two studies meeting the inclusion criteria, Mohammadi et al. [59] and Chua et al. [60], were identified in conference abstract form. Although the former was published as a full paper during the review process of this study [58], the latter was not, and does not include sufficient information to assess against all quality criteria.

Quantitative synthesis

Table 1 presents information on all included studies, while Table 2 displays the data extracted to assess the predictive ability of DCEs. We consider 844 observations where opt-in or opt-out choices were correctly predicted 75% of the time; 65% of incorrect predictions were false-positives. Figure 3 displays the sensitivity and specificity estimates for each study.[1] These estimates were calculated from the raw data.

In this context, DCEs predict that an individual will either make or not make a particular choice in reality. Therefore, a higher sensitivity would indicate that DCEs are good at predicting when individuals would choose reality, while a higher specificity would indicate that DCEs reliably predict that individuals will not make a particular choice.

We use a bivariate random-effects model to account for substantive heterogeneity between studies, and produce pooled estimates of sensitivity and specificity. Pooled estimation suggests that the sensitivity of DCE predictions was relatively high (0.88, 95% CI 0.81, 0.92), whilst specificity was substantially lower (0.34, 95% CI 0.23, 0.46). These results suggest that DCEs can be moderately informative for predicting future behavior. Specifically, when DCE data suggest that somebody *will* behave in a certain way (for example, opting for a treatment or programme), this is a more reliable statement than when DCEs suggest somebody *will not* behave in a certain way (for example, they will not use a treatment or programme). There is no consistent pattern of the number of false positives outweighing the number of false negatives, however it is possible that imperfect sensitivity may result in DCEs over-predicting demand. For the remainder of this paper, we will use the term "opt-in" to denote those participants who the DCE predicts would use a product or service.

As sensitivity and specificity estimates are pooled through bivariate random-effects modeling, we expect the two estimates to be interdependent. Supplementary material S3 shows a bivariate box plot which describes the extent to which sensitivity and specificity are interdependent with the inner oval representing the median distribution of estimates, and the outer oval the 95% confidence bound. These results indicate that there was a degree of heterogeneity between included studies, as three reside outside of the median distribution while one study tends towards being defined an outlier. There was no strong indication of a skew towards sensitivity or specificity. According to Egger's test, there was no evidence of publication bias ($p = 0.56$), however the capacity for this test to detect publication bias from a limited number of small studies is limited [55].

Visual examination of results, shown in Fig. 3, suggests that there was substantial between-study variation. The quantity I^2 statistic describes the percentage of total variation across studies, which can be attributed to heterogeneity rather than due to chance, where an I^2 of 0% indicates that there was no heterogeneity between studies while an I^2 of above 50% suggests substantial heterogeneity [53, 61]. The I^2 estimates for sensitivity and specificity are 64 and 58%, respectively, indicating that while there was substantial heterogeneity in both measures, estimates of sensitivity were subject to greater variation. Finally, we assessed publication bias through Egger's test, which suggests no evidence of publication bias ($p = 0.56$). However, the capacity for this test to detect publication bias from a limited number of small studies is limited [55].

Under the presence of heterogeneity, summary receiver operating characteristic (SROC) curves can be used to display the results of syntheses where the higher the combined sensitivity and specificity of a test (i.e., the greater true-positive rate), the closer the SROC curve will be to the top left of the SROC space [52]. Figure 4 shows the SROC for included studies, where the curve represents the relationship between the true- and false-positive rates across studies and was fitted to the data through least-squares regression [62]. The area under the SROC curve (AUC) can be a useful summary statistic of predictive ability, and the AUC we present in Fig. 4 (0.60 [95% CI 0.55, 0.64]) provides further evidence that DCEs have a moderate ability to predict choices; although there are no firm limits for "good" AUCs, meta-analyses of diagnostic tests infer a similar conclusions [63, 64].

Finally, the results of univariable meta-regression are presented in Fig. 5 and show that, even among the small number of studies, the specificity of DCE predictions is significantly and positively associated with the SP/RP response rate, alongside the number of alternatives shown in choice tasks. No factor is significantly associated with greater or

[1] One study (Krucien et al.) predicts the uptake of two treatments, and we present each separately in this analysis.

Table 1 Characteristics of included studies

No.	Authors	Publication year	Title and Journal	Place of DCE deployment	Sample size	Survey mode	Study objective
1	Krucien et al. [69]	2015	Empirical testing of the external validity of a discrete choice experiment to determine preferred treatment option: the case of sleep apnea, Health Economics	Patient group in a French hospital's sleep unit	SP: 140, RP: 138 (99% follow-up)	Face-to-face interview with trained nurse	To explore patient preferences for alternative treatments for obstructive sleep apnea syndrome (OSAS)
2	Lambooij et al. [70]	2015	Consistency between stated and revealed preferences: a discrete choice experiment and a behavioral experiment on vaccination behavior compared, BMC Research Methodology	Parents with child < 2 weeks old in the Netherlands	SP: 906, RP: 247 (27% follow-up)	Paper-based questionnaires, medical records	To compare vaccination scenarios against hepatitis B among the parents of newborn children
3	Mohammadi et al. [58]	2015	Testing the external validity of a discrete choice experiment method: an application to latent tuberculosis infection treatment, Value in Health	Patients diagnosed with latent TB infection attending TB clinics in British Columbia, Canada	SP: 214, RP: 204 (95% follow-up)	Not reported in abstract	To explore patient preferences for latent TB treatment, and predict uptake
4	Salampessy et al. [71]	2015	The predictive value of discrete choice experiments in public health: an exploratory application, Patient	Patient group in Utrecht, The Netherlands	SP: 206, RP: 54 (26% follow-up)	Paper-based questionnaires, medical records	To assess the willingness of type 2 diabetes mellitus patients to participate in a combined lifestyle intervention
5	Ryan and Watson [67]	2009	Comparing welfare estimates from payment card contingent valuation and discrete choice experiments, Health Economics	Attendees of a family planning clinic, UK	SP: 142, RP: 111 (78% follow-up)	Paper based questionnaire	To assess the willingness to pay of women to receive Chlamydia screening
6	Chua et al. [60]	2016	External validity of discrete choice experiment: findings from a field experiment. Conference paper.	Attendees of pharmacies in UK	SP: 423 RP: 258 (61% follow-up)	Computer tablet	To explore patient preference for pharmacy-based health checks
7	Kruk et al. [72]	2009	Women's preferences for place of delivery in rural Tanzania: a population-based discrete choice experiment, American Journal of Public Health	General population survey (household), Kasulu District, Tanzania	SP: 1205	Paper-based questionnaires. RP data from census	To evaluate health-system factors that influence women's delivery decisions

Table 1 (continued)

No.	Authors	Publication year	Title and Journal	Place of DCE deployment	Sample size	Survey mode	Study objective
8	Linley and Hughes [73]	2013	Decision-makers' preferences for approving new medicines in Wales: a discrete-choice experiment with assessment of external validity, Pharmacoeconomics	Past and present members of the All Wales Medicines Strategy Group (AWMSG)	SP: 41	Anonymous online questionnaire	To explore the preferences of the AWMSG appraisal committee and appraisal sub-committee for specific new medicines adoption criteria

Table 2 Extracted data for individual-level meta-analysis

No.	Authors	Outcome	True positives	True negatives	False positives	False negatives	Accuracy	Sensitivity	Specificity	PPV	NPV
1	Krucien et al.	CPAP	37	1	2	9	0.78	0.80	0.33	0.95	0.10
2	Krucien et al.	OAs	36	1	1	13	0.73	0.73	0.50	0.97	0.07
3	Lambooij et al.	All outcomes	191	6	33	17	0.80	0.92	0.15	0.85	0.26
4	Mohammadi et al.	All outcomes	147	21	30	6	0.82	0.96	0.41	0.83	0.78
5	Salampessy et al.	All Outcomes	36	4	9	5	0.74	0.88	0.31	0.80	0.44
6	Ryan and Watson	All outcomes	88	2	2	19	0.81	0.82	0.50	0.98	0.10
7	Chua et al.	All outcomes	30	36	58	4	0.52	0.88	0.38	0.34	0.90
8	Kruk et al.	All outcomes	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A	N/A
9	Linley and Hughes	All outcomes	25	12	N/A	N/A	0.64	N/A	N/A	N/A	N/A

Kruk et al. and Linley and Hughes do not contain sufficient data for inclusion in a meta-analysis

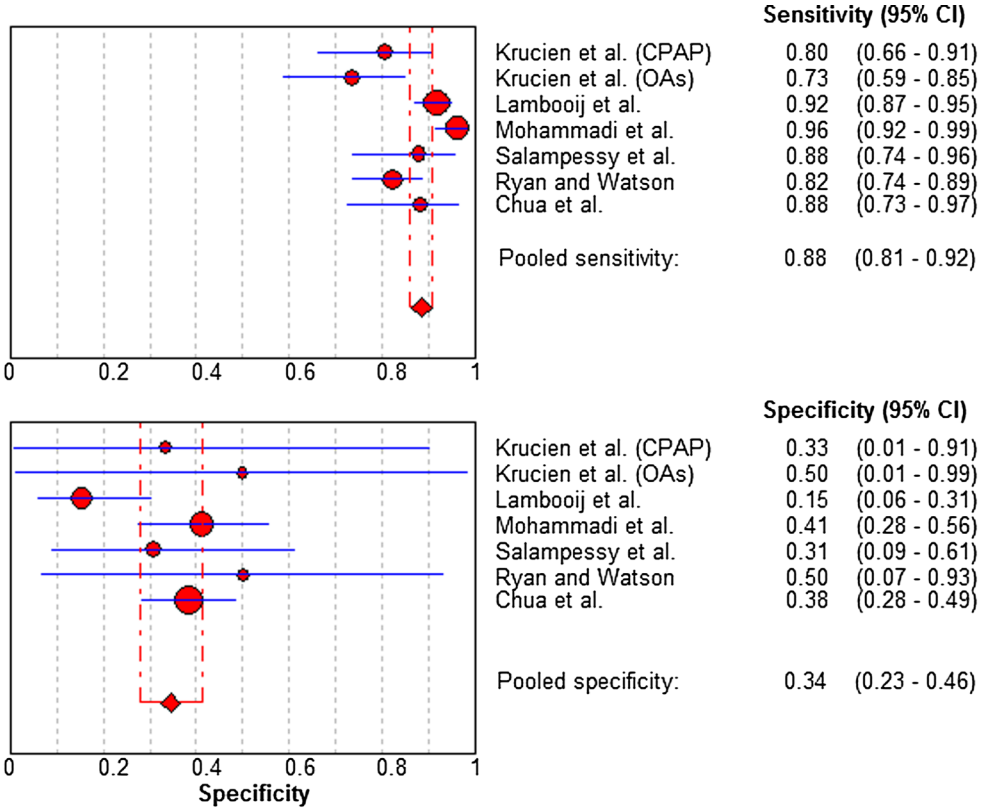

Fig. 3 Synthesis of sensitivity and specificity of DCE predictions

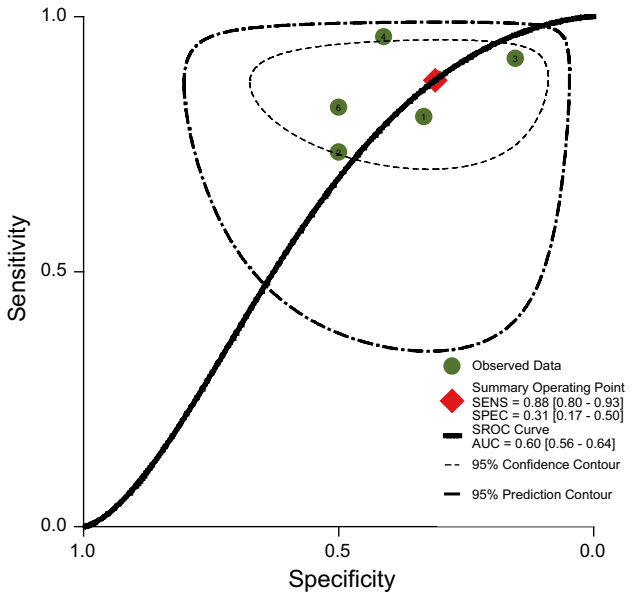

Fig. 4 Summary receiver-operator curve (SROC) of included studies

lower sensitivity. These results will become more precise and allow greater influence as more studies are published assessing the predictive ability of DCEs.

Discussion

This paper reports the results of a systematic literature review and meta-analysis and proposes a method to strengthen predictions made from DCE data. The systematic review identified seven studies as meeting the inclusion criteria. The meta-analysis of six studies with individual-level data found that DCEs have moderate, but not exceptional, accuracy when predicting health-related choices. Pooled sensitivity and specificity estimates were 88% (95% CI 81, 92%) and 34% (95% CI 23, 46%), respectively. All DCEs included in this review were exploring opting-in behaviors, and the mean observed uptake of options across all studies in this review was high at 76% (638 out of 844 observations). Only one study reported a measure of uncertainty around uptake predictions. The sensitivity of predictions was found to be greater than their specificity, suggesting that DCEs are better at predicting who would opt-into a health-related decision rather than who would not. Overall, the review found very few studies comparing DCE predictions to observed choices at an individual level, and this is a key priority for future research.

We explored heterogeneity through use of meta-regression by incorporating study-level characteristics into bivariate mixed-effects models, and found evidence that the RP/SP follow-up rate and the number of alternatives presented to respondents were positively associated with estimates of

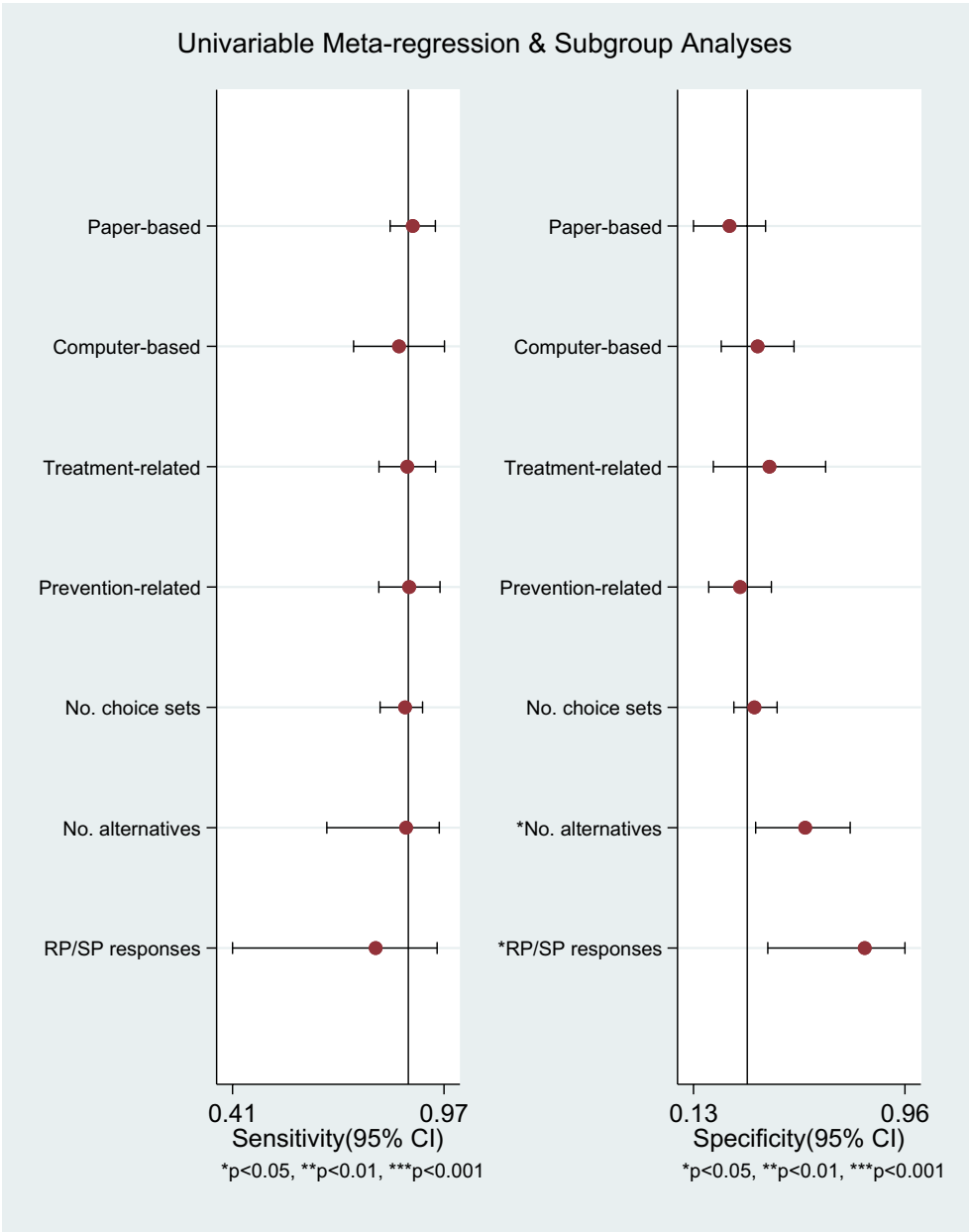

Fig. 5 Meta-regression results

specificity, but not sensitivity. This study, being the first to synthesize evidence on the predictive ability of DCEs, is the first to show the extent of the heterogeneity in specificity when predicting behaviors, and care is needed when interpreting the results of DCE predictions. Due to the low number of studies in the literature, this analysis was underpowered and future research should focus on identifying the determinants of DCE external validity in order that DCEs and prediction methods give predictions with the greatest accuracy.

The finding that opt-in predictions from DCEs are reliable is useful for planning interventions and programmes. This quantification of how well DCEs predict behavior could be used to explicitly account for uncertainty in DCE predictions, for example by using the pooled sensitivity point estimate and confidence intervals from this study to give upper and lower bounds of opting-in behaviors. Accounting for the variation in DCE prediction accuracy in this manner would make for more robust uptake and impact models. Although the pooled specificity estimate suggests that DCEs are not good predictors of opting-out behavior (i.e., should not be trusted when predicting that someone will not uptake a product or service), the pooled sensitivity estimate was relatively high and precise, making it suitable for this application.

When considering how useful DCEs are in predicting behavior, we must consider the alternative data sources available to decision-makers. When predicting the demand for new health products or services, there is likely to be almost no information to base these forecasts on. One option is to run a pilot study or demonstration project; however, even on a small scale, these can be both expensive and time-consuming. Another option would be to canvass expert opinion; however, even experts with the best of intentions can be incorrect or biased in their estimates. In such instances, DCEs can provide a relatively accurate and cost-effective option to predict individual choices. DCEs have been proposed for use to parameterize uptake and use parameters in health economic modeling [15], and if used for this purpose, parameter uncertainty could be partly accounted for using estimates from this review to adjust uptake estimates.

There may be some reasons why observed choices may be different to those predicted by DCEs. Firstly, the information presented in DCE choice tasks is necessarily a simplification of reality. Even in the case of high-quality DCEs, there are likely to be unobserved attributes present in real-life decisions that were not, or poorly, accounted for in the DCE. Where these unobserved attributes influence the decisions of participants, stated and revealed preferences will be based on heterogeneous choice attributes and may diverge. Even high-quality DCEs are unlikely to fully capture all relevant attributes of choice.

Secondly, DCE predictions may suffer from the intention-behavior gap where individuals do not always ultimately behave in ways which they might intend to [65, 66]. For example, when people are a long way ahead of making a choice, they are more likely to commit to a substantial course of action (such as giving up smoking), however as they move closer to the choice situation they are more likely to choose the smaller reward (have a cigarette). This hyperbolic discounting suggests the passage of time changes the perception of the situation and choices, potentially explaining variation between DCE predictions and actual behavior.

Results assume that there is a generalizable and measurable concept of DCE external validity. However, this review was limited by the small number of studies identified which met the inclusion criteria. This meant that we were unable to undertake a meaningful analysis of where DCEs may provide more or less accurate predictions. For example, Ryan and Watson [67] find that a DCE for Chlamydia screening has a high false-negative rate (where more people are screened in reality than predicted in the DCE), whereas Mohammadi et al. [58] find a high false-positive rate of DCEs predicting treatment for tuberculosis (where the DCE over-predicted successful treatment). With a larger number of studies, it would have been interesting to explore whether such divergent results may have been down to study context (treatment requires continued adherence and not just a one-off action), cognitive biases (perhaps social desirability bias to predict successful treatment or not disclose demand for a Chlamydia test outside of a consultation environment), or other reasons.

The meta-analytic tools used in this review are often employed to assess diagnostic tests, with the data normally used to assess these tests gathered in strictly controlled environments. In contrast, the DCEs in this study cover a range of health choices across a broad range of populations, and it is perhaps no surprise that there was substantial heterogeneity between studies. Finally, predicted probabilities are just one interpretation of DCE results. Although probabilities are calculated using the coefficients of DCE models, this review explores just one interpretation of DCE results.

When compared against existing quality assessment tools, the quality of the included studies was high. However, when assessed against the additional external validity criteria, all but two studies were substantially prone to selection bias. As none of these studies gave any detail as to how participants were selected for follow-up, we are unable to fully assess how reliable these estimates might be.

A limitation of this review is the assessment of DCE predictive ability, which is just one facet of external validity. Assessing the external validity of WTP estimates was beyond the scope of this review, whilst the external validity of WTP estimates have been robustly assessed in three systematic reviews and meta-analyses since 2004 [35–37]. In addition, included studies must have been able to define participants' choice sets in the real world, implicitly limiting the scope of this review to predictions of choices within choice sets which can be represented consistently in a stable manner across real-world and hypothetical tasks, e.g., uptake of a test, but not adherence to a treatment over time.

Although the sample size of included studies was incorporated in the standard error around pooled estimates, we are not able to account for potential publication or selection biases. The pooled sensitivity estimates are based on a multi-stage follow-up—participants must initially consent to participate in a DCE, then be successfully followed up to ascertain whether or not they engaged in a predicted behavior. A recent meta-analysis indicates that the response rates to DCE surveys are often relatively low, and vary according to contextual factors such as the number of attributes or population surveyed [68]. This review could not incorporate divergent choices from those who did not respond to DCEs in these samples, nor those who were lost to follow-up. Non-responders or those lost to follow-up may be systematically different from the included sample. Finally, we did not assess the internal validity of included studies. Although DCEs may vary in their predictive

power absolute terms, the choice-modeling literature suggests that their ability to give reliable data on the relative impact of some factors on choices is much more robust.

Conclusions

This study sought to systematically review the external validity of DCEs and is the first to synthesize studies assessing the predictive ability of DCEs in health. Seven studies were identified as meeting the inclusion criteria, and a meta-analysis of six studies with individual-level data found that DCEs have moderate, but not exceptional, accuracy when predicting health-related choices. Pooled sensitivity and specificity estimates were 88% (95% CI 81, 92%) and 34% (95% CI 23, 46%), respectively. This review and meta-analysis suggests that DCEs can be useful in predicting real-world behavior, and provides important estimates of sensitivity and specificity which can be explicitly incorporated into impact and economic models. There is a substantial need for more evidence on how DCE predictions compare to real-world choices.

Author contributions MQ conceived the study and all authors contributed to its design. MQ screened titles, abstracts and full texts, carried out analysis, and wrote the first draft of the manuscript. FTP reviewed the abstracts and full texts. All authors read and approved the final manuscript.

Funding MQ receives a Ph.D. studentship from the Economic and Social Research Council. No funder had a role in the design, analysis or writing of this article.

Compliance with ethical standards

Conflict of interest No conflicts of interest to declare.

References

1. McFadden, D.: Conditional logit analysis of qualitative choice behaviour. In: Zarembka, P. (ed.) Frontiers in Econometrics. Academic Press, New York (1974)
2. Hensher, D., Rose, J., Greene, W.: Applied Choice Analysis, 2nd edn. Cambridge University Press, Cambridge (2015)
3. Clark, M.D., et al.: Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics **32**(9), 883–902 (2014)
4. Muhlbacher, A.C., et al.: Preferences for antiviral therapy of chronic hepatitis C: a discrete choice experiment. Eur. J. Health Econ. **18**(2), 155–165 (2017)
5. Louviere, J.J., Lancsar, E.: Choice experiments in health: the good, the bad, the ugly and toward a brighter future. Health Econ. Policy Law **4**(Pt 4), 527–546 (2009)
6. Ben-Akiva, M.E., Lerman, S.R.: Discrete Choice Analysis: Theory and Application to Travel Demand. MIT Press, Boston (1985)
7. de Bekker-Grob, E.W., Ryan, M., Gerard, K.: Discrete choice experiments in health economics: a review of the literature. Health Econ. **21**(2), 145–172 (2012)
8. Beck, M.J., Fifer, S., Rose, J.M.: 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 (2016)
9. Bingham, M.F., Johnson, F.R., Miller, D.: Modeling choice behavior for new pharmaceutical products. Value Health **4**(1), 32–44 (2001)
10. McFadden, D.L.: Chapter 24 econometric analysis of qualitative response models. In: Zvi, G., Michael, D.I. (eds.) Handbook of Econometrics, pp. 1395–1457. Elsevier, Amsterdam (1984)
11. Quaife, M., et al.: The cost-effectiveness of multipurpose HIV and pregnancy prevention technologies in South Africa. J. Int. AIDS Soc. (2018). https://doi.org/10.1002/jia2.25064
12. Fiebig, D.G., et al.: Preferences for new and existing contraceptive products. Health Econ. **20**(Suppl 1), 35–52 (2011)
13. Hall, J., et al.: Using stated preference discrete choice modelling to evaluate the introduction of varicella vaccination. Health Econ. **11**(5), 457–465 (2002)
14. Terris-Prestholt, F., et al.: How much demand for New HIV prevention technologies can we really expect? Results from a discrete choice experiment in South Africa. PLoS ONE **8**(12), e83193 (2013)
15. Terris-Prestholt, F., Quaife, M., Vickerman, P.: Parameterising user uptake in economic evaluations: the role of discrete choice experiments. Health Econ. **25**, 116–123 (2016)
16. Schmidt-Ott, T., et al.: Rationality tests in discrete choice experiments—the pros and cons of testing dominant alternatives. In: 7th Meeting of the International Academy of Health Preference Research. Glasgow, UK (2017)
17. Mitchell, R.C., Carson, R.T.: Using Surveys to Value Public Goods: The Contingent Valuation Method. Hopkins University Press, Baltimore (1989)
18. Cummings, R.G., Brookshire, S., Schulze, W.D.: Valuing Environmental Goods: A State of the Arts Assessment of the Contingent Valuation Method, Volume I.B of Experimental Methods for Assessing Environmental Benefits. Rowman and Allanheld, Totowa (1986)
19. Lancsar, E., Swait, J.: Reconceptualising the external validity of discrete choice experiments. Pharmacoeconomics **32**, 951–965 (2014)
20. Hensher, D.A.: Hypothetical bias, choice experiments and willingness to pay. Transp. Res. Part B Methodol. **44**(6):735–752 (2010)
21. Ryan, M., Gerard, K.: Using discrete choice experiments to value health care programmes: current practice and future research reflections. Appl Health Econ Health Policy **2**(1), 55–64 (2003)
22. Telser, H., Zweifel, P.: Validity of discrete-choice experiments evidence for health risk reduction. Appl. Econ. **39**(1), 69–78 (2007)
23. Hollands, G.J., et al.: Altering micro-environments to change population health behaviour: towards an evidence base for choice architecture interventions. BMC Public Health **13**(1), 1218 (2013)
24. Barrage, L., Lee, M.S.: A penny for your thoughts: inducing truth-telling in stated preference elicitation. Econ. Lett. **106**(2):140–142 (2010)
25. Bosworth, R., Taylor, L.O.: 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 **12**(1) (2012)
26. Lusk, J.L., Schroeder, T.C.: Are choice experiments incentive compatible? A test with quality differentiated beef steaks. Am. J. Agric. Econ. **86**(2), 467–482 (2004)
27. McCartney, A., Cleland, J.: Choice experiment framing and incentive compatibility: observations from public focus groups. Environmental Economics Research Hub Research Reports from Environmental Economics Research Hub, Crawford School of Public Policy, The Australian National University (2010)
28. Neuman, E.: Reference-dependent preferences for maternity wards: an exploration of two reference points. Health Psychol. Behav. Med. **2**(1), 440–447 (2014)
29. Howard, K., Salkeld, G.: Does attribute framing in discrete choice experiments influence willingness to pay? Results from a discrete choice experiment in screening for colorectal cancer. Value Health **12**(2), 354–363 (2009)
30. Lancsar, E., Swait, J.: Reconceptualising the external validity of discrete choice experiments. **32**(10):951–965 (2014). http://rd.springer.com/journal/40273. http://ovidsp.ovid.com/ovidweb.cgi?T=JS&PAGE=reference&D=emed12&NEWS=N&AN=2015639269
31. Vossler, C.A., Watson, S.B.: Understanding the consequences of consequentiality: testing the validity of stated preferences in the field. J. Econ. Behav. Organ. **86**, 137–147 (2013)
32. Brownstone, D., Small, K.A.: Valuing time and reliability: assessing the evidence from road pricing demonstrations. Transp. Res. Part A Policy Pract. **39**(4), 279–293 (2005)
33. Isacsson, G.: The trade off between time and money: is there a difference between real and hypothetical choices? Tinbergen Institute Discussion Papers 13-123/VIII, Tinbergen Institute (2007) (revised 25 Aug 2013)
34. Loomis, J.: What's to know about hypothetical bias in stated preference valuation studies? J. Econ. Surveys **25**(2):363–370 (2011)
35. Murphy, J.J., et al.: A meta-analysis of hypothetical bias in stated preference valuation. Environ. Resour. Econ. **30**(3):313–325 (2005)
36. List, J., Gallet, C.: What experimental protocol influence disparities between actual and hypothetical stated values? Environ. Resour. Econ. **20**(3), 241–254 (2001)
37. Little, J., Berrens, R.: Explaining disparities between actual and hypothetical stated values: further investigation using meta-analysis. Econ. Bull. **3**(6), 1–13 (2004)
38. 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. **11**, 299–314 (2008)
39. Silva, A., et al.: Can perceived task complexity influence cheap talk's effectiveness in reducing hypothetical bias in stated choice studies? App. Econ. Lett. **19**(17):1711–1714 (2012)
40. Ready, R.C., Champ, P.A., Lawton, J.L.: Using respondent uncertainty to mitigate hypothetical bias in a stated choice experiment. Land Econ. **86**(2):363–381 (2010)
41. Viney, R., Lancsar, E., Louviere, J.: Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Rev. Pharmacoecon. Outcomes Res. **2**(4), 319–326 (2002)
42. Mandeville, K.L., Lagarde, M., Hanson, K.: The use of discrete choice experiments to inform health workforce policy: a systematic review. BMC Health Serv. Res. **14**, 367 (2014)
43. Purnell, T.S., et al.: Patient preferences for noninsulin diabetes medications: a systematic review. Diabetes Care **37**(7), 2055–2062 (2014)
44. Lewis, R.A., et al.: Patients' and healthcare professionals' views of cancer follow-up: systematic review. Br. J. Gen. Pract. **59**(564), e248–e259 (2009)
45. Harrison, M., et al.: Risk as an attribute in discrete choice experiments: a systematic review of the literature. Patient Patient Cent. Outcomes Res. **7**(2), 151–170 (2014)
46. Thompson Reuters, EndNote X7. (2013)
47. Moher, D., et al.: Preferred Reporting Items for Systematic Reviews and Meta-Analyses: the PRISMA statement. PLoS Med. **6**(7), e1000097 (2009)
48. Lancsar, E., Louviere, J.: Conducting discrete choice experiments to inform healthcare decision making: a user's guide. Pharmacoeconomics **26**(8), 661–677 (2008)
49. Juni, P., et al.: The hazards of scoring the quality of clinical trials for meta-analysis. JAMA **282**(11), 1054–1060 (1999)
50. Lancsar, E., Swait, J.: Reconceptualising the external validity of discrete choice experiments. Pharmacoeconomics **32**(10), 951–965 (2014)
51. Hamza, T.H., van Houwelingen, H.C., Stijnen, T.: The binomial distribution of meta-analysis was preferred to model within-study variability. J. Clin. Epidemiol. **61**(1), 41–51 (2008)
52. Harbord, R.M., et al.: A unification of models for meta-analysis of diagnostic accuracy studies. Biostatistics **8**(2), 239–251 (2007)
53. Dwamena, B.: MIDAS: STATA module for meta-analytical integration of diagnostic test accuracy studies. https://econpapers.repec.org/RePEc:boc:bocode:s456880 (2009). Accessed 18 Dec 2017
54. StataCorp, STATA 14. (2014)
55. Egger, M., et al.: Bias in meta-analysis detected by a simple, graphical test. BMJ **315**(7109), 629–634 (1997)
56. Bland, M.: Meta-analysis: heterogenity and publication bias. http://wwwusers.york.ac.uk/~mb55/msc/systrev/week7/hetpub-compact.pdf (2006). Accessed 18 Dec 2017
57. Liberati, A., et al.: The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration. PLoS Med. **6**(7), e1000100 (2009)
58. Mohammadi, T., et al.: Testing the external validity of a discrete choice experiment method: an application to latent tuberculosis infection treatment. Value Health **20**(7), 969–975 (2017)
59. Mohammadi, T.: Exploring the external validity of discrete choice experiment using hierarchical Bayes mixed logit: an application to latent tuberculosis. In: International Health Economics Association Congress. Milan, Italy (2015)
60. Chua, G.N., et al.: External validity of discrete choice experiments: findings from a field experiment. In: Health Economists' Study Group (HESG) Meeting. University of Manchester, Manchester, UK (2016)
61. Higgins, J.P., et al.: Measuring inconsistency in meta-analyses. BMJ **327**(7414), 557–560 (2003)
62. Walter, S.D.: Properties of the summary receiver operating characteristic (SROC) curve for diagnostic test data. Stat. Med. **21**(9), 1237–1256 (2002)
63. Engels, E.A., et al.: Meta-analysis of diagnostic tests for acute sinusitis. J. Clin. Epidemiol. **53**(8), 852–862 (2000)
64. Lee, A., et al.: A systematic review (meta-analysis) of the accuracy of the Mallampati tests to predict the difficult airway. Anesth. Analg. **102**(6), 1867–1878 (2006)
65. Sheeran, P.: Intention—behavior relations: a conceptual and empirical review. Eur. Rev. Soc. Psychol. **12**(1), 1–36 (2002)
66. Ajzen, I.: The theory of planned behavior. Organ. Behav. Hum. Decis. Process. **50**(2), 179–211 (1991)
67. Ryan, M., Watson, V.: Comparing welfare estimates from payment card contingent valuation and discrete choice experiments. Health Econ. **18**(4), 389–401 (2009)
68. Watson, V., Becker, F., de Bekker-Grob, E.: Discrete choice experiment response rates: a meta-analysis. Health Econ. **26**(6), 810–817 (2016)

69. Krucien, N., Gafni, A., Pelletier-Fleury, N.: Empirical testing of the external validity of a discrete choice experiment to determine preferred treatment option: the case of sleep apnea. Health Econ. **24**(8), 951–965 (2015)
70. Lambooij, M.S., et al.: Consistency between stated and revealed preferences: a discrete choice experiment and a behavioural experiment on vaccination behaviour compared. BMC Med. Res. Methodol. **15**, 19 (2015)
71. Salampessy, B.H., et al.: The predictive value of discrete choice experiments in public health: an exploratory application. Patient **8**(6), 521–529 (2015)
72. Kruk, M.E., et al.: Women's preferences for place of delivery in rural Tanzania: a population-based discrete choice experiment. Am. J. Public Health **99**(9), 1666–1672 (2009)
73. Linley, W.G., Hughes, D.A.: Decision-makers' preferences for approving new medicines in Wales: a discrete-choice experiment with assessment of external validity. Pharmacoeconomics **31**(4), 345–355 (2013)