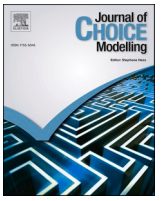
Journal of Choice Modelling 59 (2026) 100609

Contents lists available at ScienceDirect

Journal of Choice Modelling

journal homepage: www.elsevier.com/locate/jocm



"It should be relevant, reliable and feasible": Introducing FACE, an instrument for assessing the face validity of choice experiments

Annamarie de Ruijter [a,*], Job van Exel [b,c], Niek Mouter [a,d]

[a] Delft University of Technology, Faculty of Technology, Policy and Management, Transport and Logistics Group, Jaffalaan 5, 2628 BX, Delft, the Netherlands
[b] Erasmus University Rotterdam, Erasmus School of Health Policy & Management, Burgemeester Oudelaan 50, 3062 PA, Rotterdam, the Netherlands
[c] Erasmus University Rotterdam, Erasmus Centre for Health Economics Rotterdam (EsCHER), Burgemeester Oudelaan 50, 3062 PA, Rotterdam, the Netherlands
[d] Populytics, Strawinskylaan 339, 1077 XX, Amsterdam, the Netherlands

ARTICLE INFO

Keywords:
Choice experiments
Survey design
Face validity
FACE instrument

ABSTRACT

Face validity indicates to what extent participants are engaged in making choices; and understand and interpret the presented information as intended by its designer. It is an important but often overlooked aspect of the overall validity of choice experiments and no comprehensive instruments for assessing it are available. Improving its design potentially improves the quality of participants' responses and the study itself, which increases the relevance and usability for policy and practice. In this study we developed and tested an instrument to assess the FAce validity of Choice Experiments (FACE) in a uniform, systematic manner. The instrument is based on 9 components identified in the literature: clarity, completeness, decision certainty, familiarity, feasibility, legibility, relevance, sensitivity, and transparency. FACE covers these components in 13 statements with 5-point Likert scales.

1 020 participants completed the instrument following a discrete choice experiment on COVID-19 pandemic preparedness measures in the Netherlands. This first application of FACE showed that the face validity of a choice experiment was determined by whether participants considered its study design to be relevant, reliable and feasible. Moreover, we found that relevance and reliability were most strongly related to characteristics of the survey design, while feasibility was most strongly related to participants' socio-demographic characteristics. Face validity was rated high(er) by participants who were younger, male, lower educated, vaccinated against COVID-19 and sufficiently engaged in the experiment. FACE should be regarded as a first-version instrument that can be refined and further validated. We provide recommendations on how to improve FACE in future research.

1. Introduction

Choice experiments, like Discrete choice experiments (DCEs), Best-Worst Scaling (BWS) and Participatory Value Evaluation (PVE) are frequently used to identify people's preferences for policies, technologies, products, and services. Choice experiments are especially popular in the domains of health, environment, transport, and marketing (Soekhai et al., 2019; Janssen et al., 2017; Hoyos, 2010;

* Corresponding author. Laat 58, 1811 EK, Alkmaar, the Netherlands.
E-mail address: a.m.deruijter@tudelft.nl (A. de Ruijter).

https://doi.org/10.1016/j.jocm.2026.100609
Received 30 January 2025; Received in revised form 29 January 2026; Accepted 12 April 2026
Available online 21 April 2026

Hensher, 2010; McDonnell Feit et al., 2010; Schuster et al., 2024; Boxebeld, 2024). For the validity of choice experiments, it is vital that participants are willing to engage in and understand the choices that they are asked to make (Pearce et al., 2021). The methods used should accurately measure what the study intends to measure, especially because true values of the policy, good or service must be approximated as they cannot be observed in real life (Gaber and Gaber, 2010; Holden, 2010; Mariel et al., 2021). The validity of a choice experiment concerns both its scientific quality and the relevance and usefulness of its recommendations for decision-makers (Soekhai et al., 2019). Hence, it is important that choice experiments are validated to verify that the resulting data is of sufficient quality (Que et al., 2017; Anastasi and Urbina, 2007; Rowe and Frewer, 2004).

DCEs are probably the most used choice experiments and, therefore, most evidence about validity testing is related to this specific type of choice experiments. The focus in validity testing is often on the internal or external validity of the results. Internal validity examines if participants' choice behaviour aligns with the rational choice theory assumptions in which DCEs are grounded, such as continuity, monotonicity, transitivity, and stability. The methods used to assess this are widely used and accepted (Rakotonarivo et al., 2016). External validity has also been studied extensively. It evaluates if DCEs can accurately predict choices and preferences outside the study context (Janssen et al., 2017). For example, the comparison between stated and revealed preferences (de Bekker-Grob et al., 2020; Yan et al., 2019; Engstrom and Forsell, 2018) or between preferences during and after the experiment (Liebe et al., 2012; Gamper et al., 2018; Morkbak and Olsen, 2015).

Besides verifying goodness of fit and prediction success, broader and more indirect aspects of validity should also be assessed in choice experiments. Bishop and Boyle (2019), for example, introduced a framework in which they outline content validity, construct validity and convergent validity for DCEs. Content validity verifies whether the DCE and its various components cover all dimensions of what it intends to measure. This gives an indication of whether respondents are induced to make choices that align with their true preferences (Mariel et al., 2021). Construct validity relates to the researcher's prior expectations of attribute values and how these relate to one another. Mariel et al. (2021) advise to consider construct validity less strictly than internal validity, as no economic theory assumptions are violated here but deviations do warrant further explanation. Convergent validity examines whether measuring the same construct with a similar method yields in comparable results (Higgins and Straub, 2006).

An important aspect of validity that has received less attention so far is whether participants understand and are sufficiently engaged with the choices they are asked to make. If participants are not sufficiently engaged, they may misinterpret or neglect (part of) the information and instructions provided and are more prone to choose randomly or with greater error (Norman et al., 2016). In this study, therefore, we focus on the face validity of choice experiments. The literature of psychometrics defines face validity as the extent to which a measure appears to measure what it claims to measure (Anastasi and Urbina, 1997; Nunnally and Bernstein, 1994). The choice experiment and preference elicitation literature defines face validity as to what extent participants correctly interpret and understand the information presented to them in an experiment. It addresses participants' impressions on whether the survey design facilitates them in making the required choices (Taherdoost, 2016). In this context, the survey design consists of the way questions are formulated, instructions and contextual information is provided, and scenarios are presented (Viney et al., 2002). Face validity is thus not about the experimental design (i.e., the combination of attributes and levels to create the different choice scenarios), as respondents do not directly observe this. We use the definition from the choice experiment and stated preference literature in this study.

Face validity thus tests if the study materials are interpreted and understood by participants as intended by the designer of the experiment. According to Pearce et al. (2021), the concept of participants' understanding can be distinguished into grasping the choice scenarios and the act of making the choice itself. Hence, for the face validity of a choice experiment it is important that the instructions and scenarios can be easily understood and correctly interpreted by participants. In the study by Janssen et al. (2017), a panel of experts concluded that ensuring respondents' understanding and interpretation of the study materials was the most desirable and most actionable feature of a high-quality preference experiment. Although solely assessing face validity is insufficient to properly describe and present the decision problem in a DCE, these experts argued that more studies should consider face validity alongside more commonly measured types of validity. Soekhai et al. (2019) added that qualitative methods seemed most suitable to verify face validity, due to its subjective nature, and can complement the often-quantitative nature of choice experiment (validity) analysis.

Pearce et al. (2021) argued that face validity should not only be incorporated into standardized validity assessment but also researched in more detail than is currently often done. Insights into the different components that comprise participants' understanding and interpretation of choice experiments, how to measure these components and their impact on a choice experiment's validity seem invaluable but are largely missing in the literature. Solely questioning self-reported choice task difficulty does not cover the full construct, a more comprehensive method is necessary. In addition, Gaber and Gaber (2010) argued that face validity assessment should be done in a systematic way, so that conclusions about the face validity of different studies are drawn on similar grounds and the validity of choice experiments can be compared. More recently, a systematic review by Nouwens et al. (2025) concluded that only 2% of the 1 279 studied DCEs tried to understand participants' responses. The authors call on researchers to report more details on their designs and often-overlooked validity tests.

Several reasons are mentioned in the literature for why face validity remains underexplored in the field of preference elicitation. Firstly, due to long-standing debates on the subjective and intuitive nature of face validity, no formal and agreed-upon measurement approach yet exists (Haynes et al., 1995; Anastasi and Urbina, 1997; Bannigan and Watson, 2009; Patel and Desai, 2020). In the psychometric literature, face validity is often evaluated via subjective ratings or informal assessments by laypersons or test takers (Boateng et al., 2018). Less often used and not originally intended for face validity, the content validity index (CVI) can be adapted to collect ratings from laypersons instead of experts. In a CVI, experts quantify test items on standardized scales (Polit and Beck, 2006). Mokkink et al. (2010) assessed face validity by asking patients and professionals whether test items were relevant and understandable using the COSMIN checklist. Interviews with small groups of respondents to evaluate how test items are perceived and interpreted are more common (Willis, 2004). In the choice experiment and preference elicitation literature, researchers often base their judgement

about face validity on response rates, completion times, missing values, or self-reported choice task difficulty (de Bekker-Grob et al., 2010, 2019; Whitty and Gonçalves, 2018). In these studies, lower face validity is associated with lower response rates, more frequent opt-out selection, attribute non-attendance, higher drop-out rates and higher completion times.

Secondly, often a pilot test is conducted to check for any problems with face validity. However, the results are usually reported only briefly and in general terms (Pearce et al., 2021). For example, that no issues were raised or that no survey design alternations were necessary following the pilot (de Bekker-Grob et al., 2010, 2021; de Freitas et al., 2019; Hauber et al., 2016; Naik-Panvelkar et al., 2012). However, details on how the pilot test was conducted (e.g. think aloud studies, focus groups, questionnaires), which specific tests were performed to assess face validity, and how the results were interpreted, are generally missing (Fifer et al., 2018; Mansfield et al., 2017; Marshall et al., 2018; Muhlbacher et al., 2015; Tada et al., 2019). Considering the relevance of face validity for choice experiments and the current lack of methods for assessing this, our study aims to explore how the face validity of choice experiments can be assessed in a systematic and comprehensive manner.

The contribution of this study is threefold. First, we reviewed the literature on face validity to identify its components and operationalize the construct. Specifically, we inspected the literature to identify which approaches are currently used to assess the face validity of choice experiments. Second, we developed an instrument to assess the face validity of choice experiments that will facilitate systematic assessment of face validity in a comprehensive manner. Third, we tested this instrument to a DCE case study with the aim of

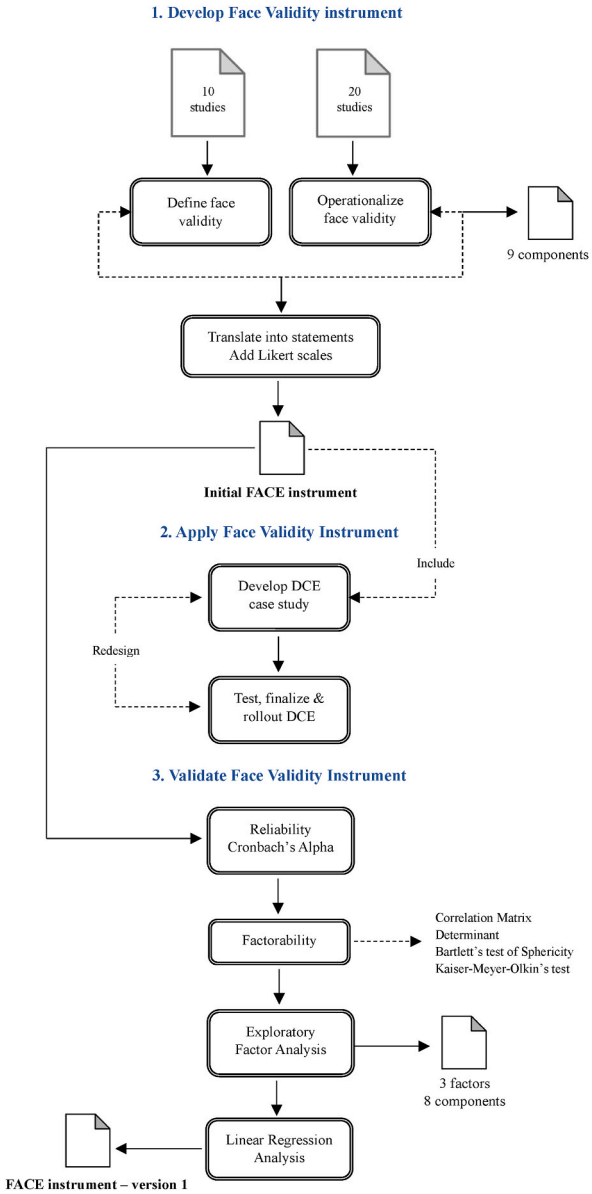


Fig. 1. Flowchart methodological study process.

examining which components of face validity mentioned in the literature contribute most to assessing this construct empirically and provide recommendations on how to improve the face validity (assessment) of choice experiments in the future.

The remainder of the paper is organized as follows: in section 2 we outline the methodology that is used to develop the assessment instrument. It includes a description of how we developed the DCE and of how we applied the instrument in our case study. Section 3 presents the results of the study and section 4 a discussion.

2. Methodology

Fig. 1 shows an overview of the methodological process of this study. The process involved four key steps: 1) developing the face validity assessment instrument based on an exploratory literature review; 2) applying the instrument in a DCE case study; 3) validating the instrument; and 4) advising how the instrument can be used to systematically measure the face validity of choice experiments. Sections 2 and 3 of this paper are structured according to the first three stages. The fourth stage will be addressed in the discussion section. To further reflect on the validity of the instrument beyond quantitative analyses, we complement this study with qualitative insights discussed later in the paper.

2.1. Developing the face validity assessment instrument

To operationalize face validity, we first searched for studies that defined and explained this construct using the Google Scholar and Science Direct databases (see Fig. 2). We searched using the terms 'validity', 'face validity,' 'theoretical validity', 'respondent understanding', 'respondent interpretation', 'respondent comprehension' and 'choice comprehension'. We identified 23 studies but disregarded those that solely mentioned face validity but did not define or explain it. As a result, we found 10 studies that examined face validity theoretically and used these to define the construct (Desai and Patel, 2020; Royal, 2016; Taherdoost, 2016; Moores et al., 2012; Gaber and Gaber, 2010; Lidwell et al., 2010; Broder et al., 2007; Hardesty and Bearden, 2004; Nevo, 1985; Turner, 1979).

Next, we combined each of the aforementioned search terms with 'discrete choice experiment', 'discrete choice study', 'choice experiment', 'preference elicitation study', 'preference elicitation experiment', 'research instrument' and 'survey' to find applied studies on face validity. For example, we also searched for 'face validity discrete choice experiment'. We identified 31 studies but disregarded those that focused on a different definition of face validity than on participants' understanding and interpretation of the study design. As a result, we selected 22 additional studies to operationalize face validity (Juschten and Omann, 2023; Jiang et al., 2023; Pearce et al., 2021; de Bekker-Grob et al., 2008, 2010, 2019, 2021; Merlo et al., 2020; Soekhai et al., 2019; Mansfield et al., 2019; Whitty and Gonçalves, 2018; Janssen et al., 2017; Que et al., 2017; Rakotonarivo et al., 2016; Norman et al., 2016; Janssen and Bridges, 2016; Bolarinwa, 2015; Clark et al., 2014; Vista et al., 2009; Roberts and Priest, 2006; Viney et al., 2002; Ryan et al., 2001). These concerned DCE, BWS and PVE studies grounded in the environmental, health and social siences.

From the 10 theoretical studies, we inferred that the construct of face validity consisted of different aspects and each study mentioned one or a few of these. We made a list of these components until saturation was reached. After reading the last 3 papers, no

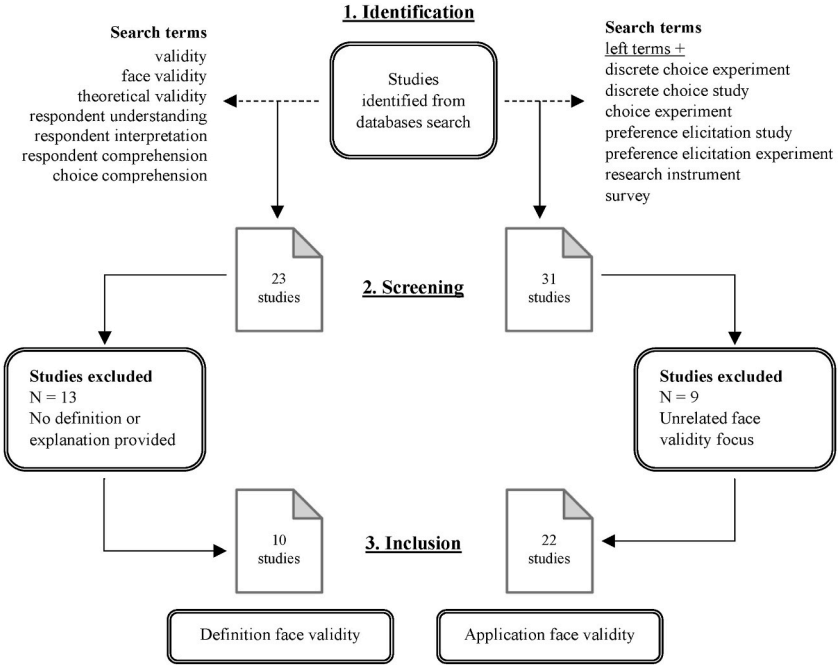


Fig. 2. Flowchart exploratory literature review face validity.

new components were added to the list. We identified 9 components highlighting different aspects of face validity based on the literature we reviewed (in alphabetical order): clarity, completeness, decision certainty, familiarity, feasibility, legibility, relevance, sensitivity and transparency.

We followed the recommendations from Gaber and Gaber (2010), Janssen et al. (2017), Soekhai et al. (2019) and Pearce et al. (2021) to develop a qualitative and systematic instrument. From the 21 applied studies, we retrieved the approaches used to assess the components. These approaches varied from quantitative to pilot testing and managing completion times. However, the approaches were not sufficiently comprehensive to justify the multifaceted construct of face validity. In 2 out of the 10 theoretical studies a qualitative approach was introduced. Some, but not all, components were presented as statements for experts on a yes/no format and for respondents on a 5-point Likert scale.

Desai and Patel (2020) presented statements on clarity, decision certainty, feasibility, legibility and transparency. Moores et al. (2012) presented statements on completeness, decision certainty, feasibility, relevance and sensitivity. We took these statements as a starting point and included additional statements such that the meaning of each component in the literature could be fully questioned. Subsequently, we combined these statements with Likert scales with 5 answer categories (i.e., strongly agree, agree, neutral, disagree and strongly disagree) on which participants can indicate their agree of agreement. The statements included in FACE from the first version of the instrument, which we expect to refine based on additional cognitive testing and further application across domains.

2.2. Applying the face validity assessment instrument

We tested the instrument to assess the **FA**ce validity of **C**hoice **E**xperiments (henceforth: FACE) in a DCE on COVID-19 pandemic preparedness in the Netherlands. In the DCE, respondents had to repeatedly choose between 2 packages of measures the government could implement to respond to a potential virus outbreak. The scenarios were defined by the outbreak context, measure duration, restrictions on work, education and activities, and long-term effects. We provided respondents with information on the study's purpose and context, instructions, and privacy. After informed consent was given, each respondent was randomly assigned to complete 16 of these scenarios. After these choice tasks, respondents received the face validity statements. The order of these statements was randomized for each respondent. Furthermore, some statements were formulated positively while other statements were formulated negatively to identify straight lining.

Fig. 3 shows an example of a choice task in the DCE. For more details on the design, see Appendix A: Description DCE on COVID-19 preparedness measures. Each respondent received an additional choice task that repeated the third scenario and 2 control questions

	Situation A	**Situation B**
Duration of the measures	2 weeks	6 weeks
Face masks	Yes, but only in public transport	Yes, in all public locations
Education (schools and universities)	Online	At location
Work (this does not apply to professions that are necessary to keep society running such as the policy, fire brigade, rubbish transport, healthcare, education, childcare, and public transport)	Work from home	No restrictions
Activities (non-essential shops, markets, bars and restaurants, cinemas, gyms, and sport clubs)	Closed	Accessible for everyone
Number of Dutch citizens per day with **long-term complaints** after infection	1,000	5,000
Number of Dutch citizens per day who **pass away** after infection	400	200

Which situation do you prefer?

- ☐ Situation A
- ☐ Situation B

If the government introduced these measures, would you comply with them?

- ☐ No, certainly.
- ☐ No, probably.
- ☐ Sometimes yes, sometimes no.
- ☐ Yes, probably.
- ☐ Yes, certainly.

Fig. 3. Example of a DCE choice task.

about information in the tasks to evaluate their choice consistency and engagement. The focus of this paper is however not on the results of the DCE but on the assessment of its face validity. The results of the DCE will be reported in a separate paper. This study was approved by research ethics review committee of the Erasmus School of Health Policy & Management, Erasmus University Rotterdam. The DCE was performed among adult citizens (18-75 years) of the Netherlands, between September and October 2023. In the weeks prior, the DCE was pilot tested by 100 participants to optimize the initial attributes and levels. We recruited the final sample from the panel bureau Norstat with the aim to be representative of the Dutch adult population on age, gender, and educational level.

Table 1 reports the socio-demographic characteristics of the final sample. The Chi-squared tests show that this sample is representative in terms of gender ($X^2 = 0.734$, p-value = 0.115) but not for age and education ($X^2 = 1.000$, p-value = 0.000). The sample had a higher education level, and the age distribution was slightly different, the oldest age category was underrepresented. Since here we are interested in testing the face validity instrument in different population groups, representativity is less relevant than for the other DCE results. During the data cleaning process, we checked if participants failed our control questions or completed the choice tasks too quickly (i.e., in less than 10 min). This did not exclude anyone.

2.3. Validating the face validity assessment instrument

To evaluate and validate FACE, we used a four-step approach. First, we evaluated the internal consistency of the instrument using Cronbach's Alpha. This statistic indicates whether the components in the instrument measure the same construct. The stronger the correlations among the components, the higher the internal consistency and the more reliable the instrument's measurements.

Second, we checked how suitable the data was for applying factor analysis. We evaluated the correlation matrix and its determinant, Barlett's test of Sphericity and Kaiser-Meyer-Olkin's (KMO) test. Correlations between face validity components need to be sufficiently strong to establish a relationship with the resulting factors. Third, we performed exploratory factor analysis to investigate how the components included in the instruments are, using principal component analysis with promax oblique rotations to account for expected correlation among factors. We used the Eigenvalue rule of $\lambda \geq 1.0$ to retain factors. Lambda denotes the amount of explained variance in one component by a particular factor. And we used the factor loadings rule of ≥ 0.4 to retain components. As the correlation of each component with a given factor needed to be sufficiently strong as well.

Finally, we converted the Likert scale answer categories into scores of 1 (fully disagree) to 5 (fully agree). Using these scores, we computed factor scores and factor-based scales for each participant, indicating how they related to each factor. A factor score is a linear combination of *all* components that load on a particular factor. We standardized the scores across all components, weighted these by score coefficients and summed them based on the regression method (see Appendix E: Scoring coefficients for the factor scores). In contrast to a factor score, a factor-based scale only considers the mean scores of the components that are *included* in a factor (see Appendix F: Factor-based scales).

Since we are interested in the groups of participants that relate to a higher and lower face validity assessment, we included a dummy dependent variable for in our linear regression analyses. The dummy takes the value of 1 if the factor-based scale has a value below 25% of the distribution (i.e., lower face validity) and takes the value of 0 otherwise. We related it to respondents' socio-demographic characteristics, experiment performance and attitude towards the presentation of the decision problem. In addition to the quantitative validation steps described above, we conducted a small qualitative think-aloud study to explore whether respondents interpreted the FACE statements as intended; the findings are presented in the discussion section.

3. Results

This section presents the results of developing, applying, and validating the instrument.

Table 1
Socio-demographic characteristics of the sample and target population.

Characteristics	Sample	Population	Chi-Squared test
Male	48.2%	49.3%	Statistic: 0.734
Female	51.8%	50.7%	P-value: 0.115
18 - 24 years	9.8%	10.9%	Statistic: 1.000
25 – 34 years	17.2%	15.8%	P-value: 0.000
35 – 44 years	19.9%	14.8%	
45 – 54 years	18.9%	18.0%	
55 – 64 years	24.5%	16.7%	
65 – 74 years	9.7%	13.7%	
Lower education[a]	21.3%	28.5%	Statistic: 1.000
Middle education	33.8%	36.8%	P-value: 0.000
Higher education	44.9%	34.6%	

Note.
[a] Education is grouped according to the CBS scale of lower (no education, primary education, LBO, VMBO/MAVO), middle (HAVO/VWO, MBO) and high (HBO, University).

3.1. Developing the face validity assessment instrument

As discussed before, we identified 9 components of face validity in the literature and retrieved the ways that each component was assessed in previous related experiments. Below we discuss more in detail how we identified the 9 components in these studies and how we transferred them into statements.

3.1.1. Clarity

We found that de Bekker-Grob et al. (2008), Que et al. (2017) and Mansfield et al. (2019) were among the few studies that examined how respondents perceived the clarity of the questions and instructions in their DCE. All 3 studies stressed that easy-to-understand choice tasks are a key component within face validity assessment. The authors argued that if the survey design provides unclear information that is difficult to understand or can be interpreted in multiple ways, it becomes more difficult for respondents to make well-considered choices. Subsequently, modelling the data is challenging as the relation between the choice and the information it is based on, is perhaps less apparent. Increased random variability will be reflected in the results and decrease accuracy (Viney et al., 2002). Hence, the statement on clarity is as follows: 'I found the choices I had to make clear'.

3.1.2. Completeness

Besides verifying whether participants included all information presented to them into their choices, it is also important to check whether participants felt that the information provided in the survey design is complete. If they missed information relevant to them, it could be that participants based their choices (also) on other information than was shown in the survey design. Omitted variable bias can complicate the modelling and analysis process (Viney et al., 2002). Moores et al. (2012) add that the DCE design should include information relevant to both the decision problem context and participants' potential concerns about it. Thus, the statements to verify whether respondents perceived the provided information to be complete are as follows: 'I have carefully read and included all information in my choices' and 'I missed relevant information to make choices'.

3.1.3. Feasibility and relevancy

A choice experiment assumes that it is feasible for participants to consider and trade-off all attributes of alternatives presented to them in their choices. However, quantitative feasibility tests are often limited to checking whether a participant always makes choices according to the best level of a given attribute. This disregards the additional information that is provided in the DCE survey design (Vista et al., 2009). Ryan et al. (2001) argued that qualitative assessment instead is preferable to examine whether participants focussed on and considered all information while making choices and did not use any simplifying strategies to circumvent complexity.

Currently, feasibility is often assessed by asking participants to indicate how easy they perceived the choice task to be on a 5-point Likert scale (e.g. Moores et al., 2012; Que et al., 2017; Jiang et al., 2023). Hence, the statements on feasibility are formulated as follows: 'I found the choices I had to make easy' and 'I could focus well on the choices I had to make'. Ryan et al. (2001) also stated that participants potentially find it easier to pay attention and make an effort to understand the information if they find the choice task and the decision problem relevant and interesting. Hence, the statements on relevancy are the following: 'I thought it was an important subject to give my opinion on' and 'I found the choices I had to make interesting'.

3.1.4. Transparency

Holden (2010, p. 637-638) argued that face validity assessment should touch upon the concept of transparency. Because face validity is established from participants' judgement its assessment can be more susceptible to framing, both consciously and unconsciously. Hence, we accounted for both the feeling of trustworthiness and steering of choices in our assessment instrument with the following statements: 'I found this research trustworthy' and 'I thought that this research steered my choices in a certain direction'.

3.1.5. Legibility

Broder et al. (2007) assessed the readability of information presented to participants in their survey design. They used the Flesh-Kinkaid readability score to determine the appropriate difficulty level and if wording needed revision. Choice experiments that are conducted in the Netherlands on behalf of the government, for example, must comply with the B1 language level qualification. The government advises to check readability and appropriateness of the difficulty level, either by the researchers themselves or by an external party such as a dedicated language agency (Dienst Publiek en Communicatie, Ministerie van Algemene Zaken, 2025). However, it is important that this finding is verified beyond the respondents who participate in the pilot study. Hence, we argue that face validity assessment should also verify whether participants found the information easy to read in the final data collection, using this statement: 'I found all information easy to read'.

3.1.6. Decision certainty

Whitty and Gonçalves (2018) argued that the extent to which participants are certain about the best option strongly relates to their engagement and understanding. Decision certainty indicated if the DCE survey design facilitates participants in interpreting the decision problem and making the required considerations. Whitty and Gonçalves (2018) used the standard errors of their choice model to make a statement about decision certainty. They associated a lower variance with better choice consistency and more precise preference estimates. Jiang et al. (2023) proxied decision certainty by asking respondents how difficult they found it to decide on their answers and the extent to which the survey design facilitated them to distinguish between alternatives.

Moores et al. (2012) argued that the DCE survey design should also appear valuable to respondents. Suitability referred to the

extent to which respondents felt that the design appeared to match the purpose of the survey, not to the appropriateness of experimental design or other technical details. Hence, we included both choice certainty and design suitability in our assessment instrument, using the following statements: 'I was always sure about the best choice' and 'I thought that this was a good method to give advice to the government as a citizen'.

3.1.7. Sensitivity and familiarity

Que et al. (2017) stated that respondents could find it more challenging to understand and stay engaged in the experiment if they are unfamiliar with the policy topic. Choice experiments on environmental and health topics often include attributes on probabilities and risk, medical terminolgy and difficult-to-imagine scenarios about the long-term future. Consequently, the presentation of these choices could become sensitive and emotionally intense for participants to see and process (Whitty and Gonçalves, 2018; Pearce et al., 2021). Norman et al. (2016) and Moores et al. (2012) examined sensitivity by asking whether participants found the choice tasks that they had to see upsetting or annoying. Therefore, we included both choice sensitivity and topic familiarity in the face validity assessment instrument. The accompanied statements are as follows: 'I found the choices I had to make unpleasant' and 'I was already familiar with the subject before I participated in this research'. Table 2 shows the 9 components and the 14 statements that are included in the instrument.

3.2. Applying the face validity assessment instrument

A total of 1 020 participants completed the 14 face validity statements. Table 3 reports the corresponding descriptive statistics. It shows that respondents found the choice tasks clearly presented (3.9 out of 5.0), decision problem important (4.0 out of 5.0), information easy to read (4.0 out of 5.0) and could focus on considering all information (4.1 out of 5.0). However, they indicated that the choice tasks were not always as easy (3.2 out of 5.0). Furthermore, they were not always as certain of the best choice among the two choice options (3.3 out of 5.0) and whether they received sufficient information to make a well-considered choice (2.7 out of 5.0).

3.3. Validating the face validity assessment instrument

A Cronbach's Alpha of $\alpha = 0.831$ indicated that the FACE instrument was sufficiently reliable (see Appendix B: Internal consistency face validity assessment instrument). All the components except 'completeness' added to the reliability, as removing these decreased alpha. Reliability only marginally improved from $\alpha = 0.831$ to $\alpha = 0.833$ if the component on choice unpleasantness was removed. We observed no weakly ($r \leq 0.3$) or highly ($r \geq 0.8$) correlated components and the determinant confirmed that the correlation matrix contained no collinearity or singularity (see Appendix C: Correlation matrix face validity components). Consequently, we retained all 14 components that we identified from the face validity literature to question in the DCE. Both the Barlett's and the KMO's test indicated that these 14 components shared sufficient variance that can be explained by a set of factors. Therefore, we concluded that our data was factorable.

Table 4 presents the main results of the exploratory factor analysis. The pattern matrix shows the relationships between the 14 components and the 3 identified factors. The factor loadings show that all components except for subject familiarity significantly loaded on any of the 3 extracted factors. The greater (smaller) the magnitude of the loading, the stronger (weaker) the relation between that specific component and the factor. Thus, the component on choice attributes (0.789) has a stronger relation with factor 1 than the component on choice clarity (0.605) and the component on choice difficulty (0.770) has a stronger relation with factor 3 than the component on choice unpleasantness (−0.668), as holds both for positive and negative relationships.

The uniqueness values in the most right column of the pattern matrix denote the amount of variance for each component that could not be explained by the factors. Hence, a lower uniqueness value implied a better explanatory power of the factors. The components on choice difficulty (0.341) and study trustworthiness (0.391) were explained best by the 3 factors. The component on subject familiarity (0.873) was most difficult to explain and did not load on any factor. As a result, this component was removed from the instrument.

Table 2
Face validity of Choice Experiments (FACE) assessment instrument.

Component	Item	Statement
Clarity	Choice clarity	"I found the choices I had to make clear."
Completeness	Information completeness	"I missed relevant information to make choices."
Feasibility	Choice difficulty	"I found the choices I had to make easy."
	Choice focus	"I could focus well on the choices I had to make."
	Choice attributes	"I have carefully read and included all information in my choices."
Relevancy	Subject importance	"I thought it was an important subject to give my opinion on."
	Choice interest	"I found the choices I had to make interesting."
Transparency	Study trustworthiness	"I found this research trustworthy."
	Choice steering	"I thought that this research steered my choices in a certain direction."
Legibility	Information legibility	"I found all the information easy to read."
Sensitivity	Choice unpleasantness	"I found the choices I had to make unpleasant."
Decision certainty	Choice certainty	"I was always sure about the best choice."
	Design suitability	"I thought this was a good method to give advice to the government as a citizen."
Familiarity	Subject familiarity	"I was already familiar with the subject before I participated in this research."

Table 3
Descriptive statistics of the face validity (FACE) assessment instrument.

Item	Fully disagree	Disagree	Neutral	Agree	Fully agree	Mean	95% CI
Choice clarity	1.8%	4.6%	15.5%	58.0%	20.1%	3.901	0.831
Information completeness	10.6%	34.9%	32.0%	18.5%	4.0%	2.705	1.016
Choice difficulty	4.7%	20.6%	30.8%	34.0%	9.9%	3.238	1.037
Choice focus	0.6%	4.9%	13.7%	57.3%	23.5%	3.982	0.789
Choice attributes	0.6%	3.2%	11.9%	58.8%	25.5%	4.054	0.743
Subject importance	2.4%	3.7%	13.7%	49.9%	30.3%	4.021	0.895
Choice interest	3.7%	8.6%	28.2%	46.4%	13.1%	3.566	0.952
Study trustworhtiness	4.6%	7.9%	37.8%	39.2%	10.5%	3.430	0.944
Choice steering	11.5%	30.8%	31.4%	19.6%	6.7%	2.790	1.091
Information legibility	1.4%	5.2%	11.7%	55.5%	26.2%	3.999	0.843
Choice unpleasantness	13.1%	26.5%	25.9%	27.4%	7.2%	2.889	1.157
Choice certainty	3.3%	21.0%	28.2%	34.8%	12.7%	3.325	1.045
Design suitability	5.9%	11.8%	24.1%	43.5%	14.7%	3.494	1.065
Subject familiarity	8.2%	11.7%	21.6%	42.9%	15.6%	3.460	1.135

Table 4
Pattern matrix with rotated factor loadings and unique variances.

Rotated factor loadings				
Component	Factor1[a]	Factor2[b]	Factor3[c]	Uniqueness
Choice clarity	0.605			0.428
Information completeness		0.675		0.455
Choice difficulty			0.770	0.341
Subject importance	0.690			0.437
Study trustworthiness		−0.583		0.391
Choice steering		0.798		0.405
Decision certainty			0.730	0.417
Design suitability		−0.500		0.456
Information legibility	0.707			0.417
Subject familiarity				0.873
Choice unpleasantness			−0.668	0.417
Choice interest	0.561			0.437
Choice focus	0.760			0.400
Choice attributes	0.798			0.407

Note: N = 1 020; LR test independent vs. saturated $X^2(df = 91) = 4\ 251.25$ with prob > $X^2 = 0.0000$.
[a] Eigenvalue = 3.9131 with prop = 0.3010.
[b] Eigenvalue = 2.7415 with prop = 0.2109.
[c] Eigenvalue = 2.3741 with prop = 0.1826.

Subsequently, we retained 13 components and 3 factors in the instrument. The correlation among these factors of |0.19| - |0.32| justified our choice of extraction and rotation methods (see Appendix D: Correlation matrix rotated factors).

The loadings on Factor1 show that face validity improves if participants consider the choice tasks to be clear (0.605), interesting (0.561) and important (0.690), the information is easy to read (0.707), and participants are engaged in making the choices (0.760; 0.789). We labelled this factor 'relevant'. The loadings on Factor2 show that face validity improves if participants consider the study to be trustworthy (−0.583; 0.789; −0.500) and complete in terms of the provided information (0.675). We labelled this factor 'reliable'. The loadings on Factor3 show that face validity improves if participants can finish the choice tasks (0.770; 0.730) and the choice tasks seem pleasant (−0.668). We labelled this factor 'feasible'. Overall, this implies that the face validity of a choice experiment is determined by whether participants consider the survey design to be relevant (Factor1), reliable (Factor2) and feasible (Factor3).

After we identified these main elements of face validity, we evaluated how participants related to each of them. We regressed the factor scores and factor-based scales on only the socio-demographic variables in partial model 1, the experiment-related variables in partial model 2, the decision problem-specific variables in partial model 3 and on all these variables combined in the full model. Table 5 presents these factor score results related to feasibility. The factor score results for relevance and reliability; and the factor-based scale results can be found in Appendix G: Linear regression analysis factor scores and in Appendix H: Linear regression analysis factor-based scales.

The first partial model shows that participants who are younger, male, and lower educated assessed the study's face validity higher. The second partial model concludes that participants who were sufficiently engaged in the experiment, determined by whether they correctly answered the repeated choice task and knowledge questions, also judged the study's face validity more positively. The third partial model shows that participants who were vaccinated against COVID-19 multiple times and agreed on the presentation and formulation of the policy problem also rated face validity higher. These partial model findings are confirmed in the factor-based scale regressions.

For the full models, a clear pattern emerged between the three main elements of face validity and the type of independent variables

Table 5
Linear regression analysis of the factor scores on feasibility (Factor3).

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic characteristics				
Age	−0.008***			−0.006**
Gender (female)	−0.222***			−0.216***
Education	−0.120**			−0.089**
Experiment characteristics				
Repeated choice task		0.332***		0.333***
Duration question		−0.030		0.018
Symptoms question		−0.312***		−0.174*
Completion time choice tasks		−0.001		−0.001
Decision-problem characteristics				
Perceived risk new wave			−0.050	−0.029
Support for policy			−0.066**	−0.057*
Vaccinated once			0.028	−0.006
twice			−0.190*	−0.183*
trice+			−0.291***	−0.244**
Constant	0.845***	0.041	0.556***	1.022***
$\mathbf{R^2}$	0.03	0.03	0.03	0.07

Note: Dependent variable has value 1 if respondent rated 'feasibility' as lower (i.e., 25% lowest responses), 0 otherwise. Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. Education is grouped according to the CBS scale of low (no education, primary education, LBO, VMBO/MAVO), middle (HAVO/VWO, MBO) and high (HBO, University). N = 1 020.

in the regression analyses. Relevance (Factor1) and reliability (Factor2) were most strongly connected to characteristics of the study design and the decision problem, while feasibility (Factor3) was most strongly related to participants' socio-demographic characteristics.

4. Discussion

In this study, we developed and conducted an initial test of an instrument (FACE) to systematically assess the face validity of choice experiments. This instrument can assist researchers to signal, assess and improve the face validity of their survey design and, consequently, the data quality of their experiments. FACE includes nine components of face validity discussed in the literature, which are covered in 13 statements. The first test on FACE showed that the face validity of a choice experiment was determined by whether participants considered its study design to be relevant, reliable and feasible. Moreover, relevance and reliability were most strongly related to characteristics of the survey design and the decision problem, while feasibility was most strongly related to participants' socio-demographic characteristics. It is important to note that FACE should be regarded as a first-version instrument. This study provides an initial empirical test. Further refinement and validation will be needed before a final version can be established.

As a first step in the further refinement of FACE, we conducted a supplementary think-aloud validation study (n = 16) to examine how respondents interpreted the instrument's statements after completing a choice experiment, which replicated the design of Boxebeld et al. (2025). Participants were recruited through forward snowball sampling, starting from the researchers' direct networks. Interviews were conducted both in person and online. Prior to participation, all participants provided informed consent for both participation and recording. During the sessions, they verbalized their thoughts while completing the choice task and commented on the set of face validity statements. The statements were questioned in a randomized order for each participant. The data were analysed using a deductive thematic approach focusing on the dimensions identified from the literature on face validity.

Overall, the findings support the face validity of FACE. For most statements, a clear majority of participants (approximately 60-100%) interpreted the items in line with their intended meaning (see Appendix I: Think-aloud interview analysis results). In particular, some statements, such as the one on focus ("I could focus well on the choices I had to make"), were interpreted as intended by all participants. However, several consistent interpretation differences emerged that warrant attention. First, statements related to choice clarity and choice difficulty were interpreted differently. Participants frequently distinguished between procedural clarity (knowing how to operate the choice task mechanism) and substantive understanding (interpreting the meaning of attributes and trade-offs). For example, one participant noted that "clarity is not only about knowing what you can do with the sliders, but also about understanding what the effects of those choices actually mean". Similarly, difficulty was often interpreted as applying to the interface rather than to the decision itself: "the sliders were easy to use, but the choices themselves are not easy at all". In addition, several participants appeared uncertain about what exactly was meant by 'the choice task', sometimes interpreting this as the interface and sometimes as the decision problem itself.

Second, the statement on information use revealed that participants did not interpret 'all information' literally, but rather as all information they considered relevant. Selective reading was widely framed as a rational strategy rather than as limited engagement, as one participant remarked: "I did not read everything, but I did read everything I needed to make my choices". Third, the statement on trustworthiness was interpreted in multiple ways, referring to the institutional credibility of the researchers, the procedural integrity of

the study, or the reliability of the underlying data and model. Finally, responses to the statement on choice certainty showed that several participants did not interpret it as referring to the existence of an objectively best option, but rather as referring to whether they could stand behind their decision. As one participant put it: "there is no single best choice, only well-considered choices that you can stand behind".

These findings suggest that while the instrument performs well overall, several statements were not always interpreted as intended, suggesting that further specification of the wording is warranted to reduce interpretive ambiguity in future applications. For example, the statement on clarity could be reformulated to focus on whether it was clear what participants could choose rather than on 'clear choices' in general. The statement on difficulty could explicitly refer to the technical feasibility of the choice task rather than to the difficulty of the choices themselves. Similarly, the attribute statement could be specified to refer to information participants considered relevant rather than to 'all' information. Finally, the item on choice certainty could be reformulated to capture confidence in making well-considered rather than 'best' choices. Based on these observations, future research could test alternative formulations of these items, such as "I found the questions in this research clear", "I found the choice task technically easy to use", "I have read and included all information that I considered relevant in my choices" and "I felt confident that I made well-considered choices". Such work could examine whether these formulations further improve the face validity of a revised version of the instrument.

While the main study used a comprehensive set of statements based on an exploratory literature review and resulted in a promising instrument, it has various limitations. First, we did not use FACE to assess and improve the face validity of our survey design but only applied it during the final data collection and analysis to evaluate and validate the instrument itself. In this study, we developed FACE and showed that it is able to identify and measure the three main elements of face validity. An important next step for further research involves further validating the instrument through studying the face and content validity of the instrument itself.

Although we conducted an initial study to examine the instrument's face validity, further qualitative work could involve using think-aloud techniques while respondents participate in a choice experiment and complete the instrument. Preferably in an individual interview setting. This allows the researcher to observe respondents while completing the experiment as well as the instrument and to compare these observations with the self-reported data from the instrument on whether they actually understood and included the information as intended. Another possibility to test whether respondents understand the FACE statements as intended would be to provide descriptions of the meaning of the statements to respondents and to ask them to rate the clarity of the statements and recommend improvements. However, a balance between increased comprehensibility and complexity should be considered as adding too much text can decrease feasibility and legibility. Determining the optimal balance is in our view an interesting avenue for further research. Also, of importance is studying the content validity of the instrument, or how well the statements cover the dimensions of face validity. For instance, for further research projects, researchers with published expertise in psychometric validity can check the content of each statement and can potentially add components that have been missed.

To test the content validity of the instrument, we think that in further research it is particularly desirable to further deepen our understanding of how respondents interpret the FACE statements and, if needed, to test whether alternative formulations could help respondents to better understand the statements as intended. Such research projects would improve the face validity and content validity of the instrument. For example, it is useful to test whether respondents who disagree with the statement 'I found the choices I had to make easy' mainly refer to the nature of the decision problem or to the complexity of the presentation of the choice tasks. Or, for the statement 'I found the choices I had t make unpleasant', to test whether participants' responses refer to the decision problem itself, the way the questions are formulated or another characteristic of the experiment. Another example is to verify whether respondents who disagree with the statement 'I found the choices I had to make clear' base their answer on whether the questions were phrased in such a way that they were easy to understand or on their own cognitive state and skills.

Second, although we developed FACE to be applicable to choice experiments in general, we applied the instrument to one case study and one country, which impacts the generalizability of our findings. The case study was based on a health-related DCE, but COVID-19 had broad societal effects beyond just health. On the one hand, the topic of the DCE was probably familiar for most participants. COVID-19 had been relevant to their personal lives for a significant period, which may have affected the feasibility of the DCE positively. On the other hand, in most countries there is still considerable debate about the effects of COVID-19 and vaccinations, and the desirability of government intervention, which may have affected the reliability of the DCE for participants.

We considered COVID-19 to be an interesting and broad case study for testing FACE, as there was potential for generalizing the results beyond health. However, the extent to which the instrument is applicable to different types of choice experiments, such as best-worst scaling or participatory value evaluation, and in other policy domains, such as transport, energy, environment, and marketing, should be verified in future research (Cheung et al., 2016; Parvin et al., 2016; Juschten and Omann, 2023; Boxebeld et al., 2023). While we do not anticipate significant changes in the 3 main face validity elements, the factor loadings and relative importance of the components within these elements could differ across domains and methods. The potential relationship between the face validity elements and the survey design characteristics also presents an interesting area for further research.

Third, we acknowledge that using 13 statements increases FACE's accuracy but may decrease its applicability, as available space in surveys is usually limited. Further research could focus on the psychometric properties of the instrument and attempt to create a shorter version. This would provide researchers with a choice to focus on comprehensiveness and accuracy by using the full version or efficiency and practicability by using the short version. Fourth, while the literature indicates decision problem familiarity as a component of face validity, it did not load on any of the 3 factors (see Table 4). The absence of this relationship could be because there was insufficient variation in familiarity with the decision problem, because familiarity and respondent understanding are not sufficiently related, or because the formulation of the statement does not capture this meaning adequately. Future research should further examine the relevance and content validity of this component for assessing face validity of choice experiments.

Finally, another limitation of this study is that we are not yet able to determine cut-off values for face validity. Currently, the

instrument can give an indication of how a survey design scores on the 3 face validity elements and their components, and whether changes to the survey design improve the face validity of the experiment but cannot be used to determine when a survey design is sufficiently face valid. This also presents an interesting avenue for future research. Especially if quantitative face validity assessment using FACE can be complemented with qualitative assessments (e.g., in focus groups, think aloud or eye-tracking studies), and can be related to established tests of internal and external validity, more insight can be provided on potential cut-off values.

When the FACE instrument is further improved and validated, it can be applied in different phases of the research process. Applying FACE is probably most valuable during the pilot phase of a choice experiment, as the survey design can still be adjusted and optimized at this stage. We argue that it is also important to apply the instrument in the final data collection and analysis. FACE can provide a signal to researchers that face validity aspects in their experiment require attention. For example, the instrument can be used during the data screening process to identify participants who might have misinterpreted or misunderstood the information provided to them. This requires researchers to be cautious on choosing the appropriate model and interpreting the results, because the underlying data may (partly) deviate from rational choice or utility maximization assumptions. However, this does not necessarily imply that these participants should be removed from the sample as a low face validity is not inherently negative. For instance, it is conceivable that respondents' awareness and familiarity of a new decision problem may be lower compared to a known decision problem.

Face validity assessment can also provide valuable complementary information about the quality of responses next to standard measures of consistency and completion time often used in choice experiments. It can give researchers insight into potential reasons why participants speed through (certain) choice tasks and FACE statements. We argue that in subsequent studies, completion times for the choice experiment and the FACE questions should be reported, compared and substantiated. Furthermore, FACE can help researchers to perform sub-group analyses for the participants who differed in how relevant, reliable or feasible they found the experiment. Sensitivity analyses can be done with and without participants who assessed the face validity, or a specific part of it, as low to see the impact on the results. For example, latent class analysis can be done to identify participant groups that assess the face validity as low, which can be used as input for further improving choice experiments. Also, the statements can be related to clusters in the experiment data and test whether participants who do not have clear preferences have assessed (aspects of) face validity as low. For example, participants who understood the choice tasks but found them incomplete and based their choice (also) on information that was not presented.

Concluding, this study contributes to the literature by presenting an instrument to systematically assess the face validity of choice experiments, a relevant but still often overlooked aspect of the overall validity of choice experiments. We argue that researchers should give more attention to face validity and its uniform measurement, as this allows for comparisons across studies and provides insights that can help improve the design of choice experiments. This facilitates participants in understanding and interpreting the study materials in the way that it is intended by the designer, improving the quality of their response and the experiment itself. The same rationale may apply to the acceptance of the results of choice experiments by policymakers and other stakeholders; and to the inclusiveness of research to accommodate subgroups in the population that experience low face validity.

CRediT authorship contribution statement

Annamarie de Ruijter: Writing – review & editing, Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Job van Exel:** Writing – review & editing, Methodology, Conceptualization. **Niek Mouter:** Writing – review & editing, Conceptualization.

Funding

This work was supported by a grant from the National Institute for Public Health and the Environment (RIVM) in the Netherlands.

Declaration of competing interest

The authors have nothing to declare.

Appendix A. Description DCE on COVID-19 preparedness measures

We tested the face validity assessment instrument in a DCE on governmental measures to prevent the spread of a potential future wave of COVID-19, in the Netherlands. We presented participants with scenarios that described a package of measures that responded to a possible virus outbreak. The attributes and levels were selected based on a review of previous studies on COVID-19 prevention and control measures (Li et al., 2021; Sicsic et al., 2023; Veldwijk et al., 2023) and feedback from epidemiologic experts. The attributes included the duration of the measure, wearing of facemasks in public, restrictions on education, work and activities; and the number of cases of long-term complaints, and deaths after infection. Based on the initial attributes and levels, we constructed a D-efficient design with 64 scenarios grouped into four blocks of 16 choice tasks. Participants were randomly assigned to one of these blocks.

On the introductory screen, we provided information on the study's purpose, contextual situation, instructions and data privacy. We also asked participants for their informed consent. After they gave their consent, they were presented with the 16 choice scenarios. For each scenario, they were asked if they preferred situation A or situation B. We also asked them to indicate the likelihood of complying with the measures from their preferred scenario in the first five choice tasks. Each participant received an additional choice task that repeated the third scenario and two test questions about information in the tasks to evaluate their engagement and consistency. We

asked follow-up questions to learn more about participants' motivation for supporting or opposing the measures. We also collected information on their socio-demographic backgrounds, health and future pandemic risk perceptions.

Appendix B. Internal consistency face validity assessment instrument

Table B.1
Cronbach's Alpha Face validity assessment instrument (FACE)

Item	Sign	Cronbach's Alpha (if item deleted)			
		FACE Instrument	**Factor 1: Relevance**	**Factor 2: Reliability**	**Factor 3: Feasibility**
Choice clarity	+	0.811	0.789		
Information completeness	-	0.824		0.668	
Choice difficulty	+	0.823			0.486
Subject importance	+	0.822	0.801		
Study trustworthiness	+	0.808		0.591	
Choice steering	-	0.831		0.678	
Choice certainty	+	0.826			0.560
Design suitability	+	0.809		0.625	
Information legibility	+	0.813	0.779		
Choice unpleasantness	-	0.833			0.641
Choice interest	+	0.815	0.804		
Choice focus	+	0.816	0.779		
Choice attributes	+	0.820	0.785		
Overall		**0.831**	**0.818**	**0.704**	**0.658**

Note: The negatively worded statements have been inverted, which is indicated by a minus (−) sign. The alpha coefficients denote the values if that specific item is removed from the instrument or factor.

Appendix C. Correlation matrix face validity statements

Table C.1
Pearson correlation matrix 14 face validity statements

Item	r												
Choice clarity	1.000												
Information completeness	−0.272	1.000											
Choice difficulty	0.396	−0.224	1.000										
Subject importance	0.326	−0.103	0.090	1.000									
Study trustworthiness	0.424	−0.346	0.262	0.377	1.000								
Choice steering	−0.190	0.378	−0.037	−0.201	−0.351	1.000							
Choice certainty	0.304	−0.173	0.472	0.090	0.250	0.012	1.000						
Design suitability	0.435	−0.311	0.258	0.402	0.589	−0.289	0.285	1.000					
Information legibility	0.569	−0.226	0.320	0.343	0.346	−0.168	0.225	0.364	1.000				
Choice unpleasure	−0.165	0.326	−0.391	−0.013	−0.204	0.220	−0.322	−0.206	−0.143	1.000			
Subject familiarity	0.092	0.045	0.105	0.115	0.045	−0.039	0.041	0.051	0.117	0.003	1.000		
Choice interesting	0.369	−0.209	0.183	0.488	0.530	−0.277	0.148	0.456	0.375	−0.080	0.063	1.000	
Choice focusing	0.481	−0.138	0.275	0.376	0.319	−0.166	0.267	0.285	0.544	−0.130	0.149	0.400	1.000
Choice attributes	0.430	−0.128	0.205	0.455	0.294	−0.125	0.195	0.310	0.501	−0.065	0.118	0.341	0.539

Note: Determinant correlation matrix = 0.009, Bartlett's test of Sphericity: $X^2 = 4\,802.875$; df = 105; P-value = 0.000, Kaiser-Meyer-Olkin Measure of Sampling Adequacy = 0.873.

Appendix D. Correlation matrix rotated factors

Table D.1
Pearson correlation matrix rotated promax oblique factors

	Factor 1: Relevance	Factor 2: Reliability	Factor 3: Feasibility
Factor 1: Relevance	1.000		
Factor 2: Reliability	−0.316	1.000	
Factor 3: Feasibility	0.236	−0.189	1.000

Appendix E. Scoring coefficients for the factor scores

Table E.1
Scoring coefficients factor scores

Item	Factor 1: Relevance	Factor 2: Reliability	Factor 3: Feasibility
Choice clarity	0.178	−0.005	0.129
Information completeness	−0.087	0.336	0.122
Choice difficulty	0.028	−0.064	0.405
Subject importance	0.216	0.074	−0.179
Study trustworthiness	0.076	0.274	−0.005
Choice steering	−0.053	0.411	−0.101
Choice certainty	0.023	−0.085	0.386
Design suitability	0.087	0.229	0.019
Information legibility	0.219	−0.057	0.076
Choice unpleasantness	−0.116	0.159	0.347
Choice interest	0.162	0.174	−0.118
Choice focus	0.241	−0.099	0.053
Choice attributes	0.257	−0.100	−0.018

Note: To obtain the scoring coefficients, the regression method was used.

Appendix F. Factor-based scales

Table F.1
Descriptive statistics factor-based scales

	Mean	SD	Min	25%*	Median	75%	Max
FBS 1: Relevance	3.92	0.61	1.50	3.58	4.00	4.33	5.00
FBS 2: Reliability	3.36	0.75	1.00	3.00	3.50	4.00	5.00
FBS 3: Feasibility	3.22	0.83	1.00	2.67	3.33	4.00	5.00

Note: N = 1,020, *Q1/25% is used as the dependent variable in the linear regression analyses for the factor-based scales.

Table F.2
Descriptive statistics dependent dummy variable factor-based scale regressions

	FBS 1: Relevance	FBS 1: Relevance	FBS 1: Relevance
N Dummy < Q1 = 0	765	783	702
N Dummy < Q1 = 1	255	237	318

Note: We generated a dummy variable that takes the value of 1 if the factor scale has a value below Q1 (25%, i.e. lower face validity) and takes the value of 0 otherwise. This dummy is the dependent variable in the linear regression analyses for the factor-based scales.

Appendix G. Linear regression analysis factor scores

Table G.1
Linear regression analysis of the factor scores on relevance (Factor1)

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic variables				
Age	−0.002			−0.005**
Gender (female)	−0.010			−0.046
Education	0.007			−0.022
Experiment variables				
Repeated choice task		0.238***		0.256***
Duration question		0.316***		0.247***
Symptoms question		0.200*		0.145
Completion time choice sets		−0.001*		−0.001*
Decision-problem variables				
Perceived risk new wave			−0.156	−0.006

(*continued on next page*)

Table G.1 (*continued*)

	Partial model 1	Partial model 2	Partial model 3	Full model
Support for policy			0.145***	0.132***
Vaccinated once			0.118	0.082
Twice			−0.154	−0.041
trice+			0.152	0.153
Constant	0.116	−0.595***	−0.567***	−0.737***
R^2	0.02	0.03	0.04	0.06

Note: Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. N = 1 020.

Table G.2
Linear regression analysis of the factor scores on reliability (Factor2)

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic variables				
Age	−0.004**			−0.007***
Gender (female)	−0.071			−0.109*
Education	0.027			−0.010
Experiment variables				
Repeated choice task		0.158*		0.202**
Duration question		0.350***		0.232***
Symptoms question		0.114		0.027
Completion time choice tasks		0.001*		0.001*
Decision-problem variables				
Perceived risk new wave			0.045	0.062
Support for policy			0.213***	0.204***
Vaccinated once			0.304**	0.239*
Twice			0.254**	0.214**
trice+			0.250**	0.258**
Constant	0.275*	−0.532***	−1.101***	−1.007***
R^2	0.01	0.03	0.09	0.11

Note: Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. N = 1 020.

Appendix H. Linear regression analysis factor-based scales

Table H.1
Linear regression analysis of the factor-based scales on relevance (Factor1)

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic variables				
Age	0.001			0.001
Gender (female)	−0.022			−0.007
Education	−0.031			−0.017
Experiment variables				
Repeated choice task		−0.099***		−0.102***
Duration question		−0.134***		−0.106***
Symptoms question		−0.115**		−0.085*
Completion time choice sets		0.000		0.000
Decision-problem variables				
Perceived risk new wave			0.020	0.016
Support for policy			−0.054***	−0.048***
Vaccinated once			−0.054	−0.039
Twice			0.013	0.028
trice+			−0.077*	−0.068
Constant	0.304***	0.534***	0.438***	0.623***
R^2	0.01	0.03	0.03	0.06

Note: Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. N = 1 020.

Table H.2
Linear regression analysis of the factor-based scales on reliability (Factor2)

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic variables				
Age	0.001			0.002*
Gender (female)	0.006			0.015
Education	0.016			0.029
Experiment variables				
Repeated choice task		−0.051		−0.064*
Duration question		−0.082**		−0.048
Symptoms question		−0.041		−0.012
Completion time choice sets		−0.000		−0.000
Decision-problem variables				
Perceived risk new wave			−0.022	−0.024
Support for policy			−0.061***	−0.059***
Vaccinated once			−0.061	−0.052
Twice			−0.085*	−0.085*
trice+			−0.087**	−0.098**
Constant	0.14**	0.340***	0.578***	0.555***
R^2	0.00	0.01	0.05	0.06

Note: Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. N = 1 020.

Table H.3
Linear regression analysis of the factor-based scales on feasibility (Factor3)

	Partial model 1	Partial model 2	Partial model 3	Full model
Socio-demographic variables				
Age	0.003***			0.003***
Gender (female)	0.120***			0.114***
Education	0.051***			0.046**
Experiment variables				
Repeated choice task		−0.115***		−0.123***
Duration question		0.054		0.052
Symptoms question		0.096**		0.055
Completion time choice sets		0.000		0.000
Decision-problem variables				
Perceived risk new wave			−0.001	−0.011
Support for policy			0.019	0.013
Vaccinated once			−0.085	−0.074
Twice			−0.022	−0.030
trice+			0.001	−0.028
Constant	−0.079	0.269***	0.257***	−0.053
R^2	0.03	0.02	0.01	0.05

Note: Significance levels of 1%, 5% and 10% are indicated by ***, ** and *, respectively. N = 1 020.

Appendix I. Think-aloud interview analysis results

Table I.1
Interview analysis results of participants' interpretation of the FACE statements

Item	Intended construct	% alignment[a]	Main issue (if any)
Choice clarity	Clarity of the task and instructions	70%	Procedural vs. substantive clarity
Information completeness	Completeness of information	35%	Different notions of 'relevant'
Choice difficulty	Manageability of the choice task	50%	Interface vs. decision difficulty
Subject importance	Importance of the policy topic	90%	-
Study trustworthiness	Trustworthiness of the research and its design	60%	Trust in design vs. institution vs. impact
Choice steering	Directional framing by the choice task design	55%	Constraints vs. directional framing

(*continued on next page*)

Table I.1 (*continued*)

Item	Intended construct	% alignment[a]	Main issue (if any)
Choice certainty	Confidence when making choices	30%	'Best' vs. 'defensible' choices
Design suitability	Suitability of the design for expressing preferences and advising policymakers	85%	-
Information legibility	Readability of information	80%	Amount vs. language
Choice unpleasantness	Experienced emotional discomfort or aversion during the choice task	50%	Emotionally vs. cognitively unpleasant
Choice interest	Intrinsic interest in the choice task and attributes	85%	Minor framing differences
Choice focus	Ability to concentrate on the task and process information attentively	100%	-
Choice attributes	Active use of provided information in choices	25%	'All' vs. 'all I need"
Subject familiarity	Prior familiarity with the policy topic	95%	-

Note.

[a] Alignment percentages are indicative and based on qualitative coding of whether participants' primary interpretation of each statement aligned with the intended construct (full = 1, partial = 0.5, and mismatch = 0). Items with lower alignment scores were not necessarily misunderstood, but often interpreted in a broader of different way than intended.

Data availability

Data will be made available on request.

References

Anastasi, A., Urbina, S., 1997. Psychological Testing, 7Th Ed. Prentice Hall/Pearson Education.

Anastasi, A., Urbina, S., 2007. Psychological Testing (2Nd Impression). Prentice-Hall: Pearson, NJ.

Bannigan, K., Watson, R., 2009. Reliability and validity in a nutshell. J. Clin. Nurs. 18 (23), 3237–3243. https://doi.org/10.1111/j.1365-2702.2009.02939.x.

Bishop, R.C., Boyle, K.J., 2019. Reliability and validity in nonmarket valuation. Environ. Resour. Econ. 72, 559–582. https://doi.org/10.1007/s10640-017-0215-7.

Boateng, G.O., Neilands, T.B., Frongillo, E.A., Melgar-Quinonez, H.R., Young, S.L., 2018. Best practices for developing and validating scales for health, social and behavioral research: a primer. Front. Public Health 6 (149), 1–18. https://doi.org/10.3389/fpubh.2018.00149.

Bolarinwa, O.A., 2015. Principles and methods of validity and reliability testing of questionnaires used in social and health science researches. Niger. Postgrad. Med. J. 22 (4), 195–201. https://doi.org/10.4103/1117-1936.173959.

Boxebeld, S., Mouter, N., van Exel, J., 2025. Trade-offs in long-term care for older people in an ageing society: a constrained portfolio choice experiment. The Journal of the Economics of Ageing 32, 100599. https://doi.org/10.1016/j.jeoa.2025.100599.

Boxebeld, S., 2024. Ordering effects in discrete choice experiments: a systematic literature review across domains. Journal of Choice Modelling 51, 100489. https://doi.org/10.1016/j.jocm.2024.100489.

Boxebeld, S., Mouter, N., van Exel, J., 2023. Participatory value evaluation (PVE): a new preference-elicitation method for decision making in healthcare. Appl. Health Econ. Health Pol. 22 (2), 145–154. https://doi.org/10.1007/s40258-023-00859-9.

Broder, H.L., McGrath, C., Cisneros, G.J., 2007. Questionnaire development: face validity and item impact testing of the child oral health impact profile. Community Dent. Oral Epidemiol. 35 (s1), 8–19. https://doi.org/10.1111/j.1600-0528.2007.00401.x.

Cheung, K.L., Wijnen, B.F., Hollin, I.L., Janssen, E.M., Bridges, J.F., Evers, S.M., Hiligsmann, M., 2016. Using best-worst scaling to investigate preferences in health care. PharmaEconomics 34, 1195–1209. https://doi.org/10.1007/s40273-016-0429-5.

Clark, M.D., Determann, D., Petrou, S., Moro, D., de Bekker-Grob, E.W., 2014. Discrete choice experiments in health economics: a review of the literature. Pharmacoeconomics 32, 883–902. https://doi.org/10.1007/s40273-014-0170-x.

de Bekker-Grob, E.W., Donkers, B., Bliemer, M.C., Veldwijk, J., Swait, J., 2020. Can healthcare choice be predicted using stated preference data? Soc. Sci. Med. 246, 1–13. https://doi.org/10.1016/j.socscimed.2019.112736.

de Bekker-Grob, E.W., Donkers, B., Veldwijk, J., Jonker, M.F., Buis, S., Huisman, J., Bindels, P., 2021. What factors infuence non-participation Most in colorectal cancer screening? A discrete choice experiment. The Patient - Patient-Centered Outcomes Research 14, 269–281. https://doi.org/10.1007/s40271-020-00477-w.

de Bekker-Grob, E.W., Essink-Bot, M.L., Meerding, W.J., Pols, H.A., Koes, B.W., Steyerberg, E.W., 2008. Patients' preferences for osteoporosis drug treatment: a discrete choice experiment. Osteoporos. Int. 19, 1029–1037. https://doi.org/10.1007/s00198-007-0535-5.

de Bekker-Grob, E.W., Hol, L., Donkers, B., van Dam, L., Habbema, J.F., van Leerdam, M.E., Steyerberg, E.W., 2010. Labeled versus unlabeled discrete choice experiments in health economics: an application to colorectal cancer screening. Value Health 13 (2), 315–323. https://doi.org/10.1111/j.1524-4733.2009.00670.x.

de Bekker-Grob, E.W., Swait, J.D., Kassahun, H.T., Bliemer, M.C., Jonker, M.F., Veldwijk, J., Donkers, B., 2019. Are healthcare choices predictable? The impact of discrete choice experiment designs and models. Value Health 22 (9), 1050–1062. https://doi.org/10.1016/j.jval.2019.04.1924.

de Freitas, H.M., Ito, T., Hadi, M., Al-Jassar, G., Szatkowski, M.H., Nafees, B., Lloyd, A.J., 2019. Patient preferences for metastatic hormone-sensitive prostate cancer treatments: a discrete choice experiment among men in three European countries. Adv. Ther. 36, 318–332. https://doi.org/10.1007/s12325-018-0861-3.

Desai, S., Patel, N., 2020. ABC of face validity of questionnaire. Int. J. Pharmaceut. Sci. Rev. Res. 65 (1), 164–168. https://doi.org/10.47583/ijpsrr.2020.v65i01.025.

Dienst Publiek en Communicatie, Ministerie van Algemene Zaken, 2025. Taalniveau B1. https://www.communicatierijk.nl/vakkennis/rijkswebsites/aanbevolen-richtlijnen/taalniveau-b1.

Engstrom, P., Forsell, E., 2018. Demand effects of consumers' stated and revealed preferences. J. Econ. Behav. Organ. 150, 43–61. https://doi.org/10.1016/j.jebo.2018.04.009.

Fifer, S., Rose, J., Hamrosi, K.K., Swain, D., 2018. Valuing injection frequency and other attributes of type 2 diabetes treatments in Australia: a discrete choice experiment. BMC Health Serv. Res. 18 (675), 1–11. https://doi.org/10.1186/s12913-018-3484-0.

Gaber, J., Gaber, S.L., 2010. Using face validity to recognize empirical community observations. Eval. Progr. Plann. 33 (2), 138–146. https://doi.org/10.1016/j.evalprogplan.2009.08.001.

Gamper, E.M., Holzner, B., King, M.T., Norman, R., Viney, R., Pharmd, V.N., Kemmler, G., 2018. Test-retest reliability of discrete choice experiment for valuations of QLU-C10D health states. Value Health 21 (8), 958–966. https://doi.org/10.1016/j.jval.2017.11.012.

Hardesty, D.M., Bearden, W.O., 2004. The use of expert judges in scale development: implications for improving face validity of measures of unobservable constructs. J. Bus. Res. 57 (2), 98–107. https://doi.org/10.1016/S0148-2963(01)00295-8.

Hauber, A.B., Nguyen, H., Posner, J., Kalsekar, I., Ruggles, J., 2016. A discrete-choice experiment to quantify patient preferences for frequency of glucagon-like peptide-1 receptor agonist injections in the treatment of type 2 diabetes. Curr. Med. Res. Opin. 32 (2), 251–262. https://doi.org/10.1185/03007995.2015.1117433.
Haynes, S.N., Richard, D.C., Kubany, E.S., 1995. Content validity in psychological assessment: a functional approach to concepts and methods. Psychol. Assess. 7 (3), 238–247. https://psycnet.apa.org/doi/10.1037/1040-3590.7.3.238.
Hensher, D.A., 2010. Hypothetical bias, choice experiments and willingness to pay. Transp. Res. Part B Methodol. 44 (6), 735–752. https://doi.org/10.1016/j.trb.2009.12.012.
Higgins, P.A., Straub, A.J., 2006. Understanding the error of our ways: mapping the concepts of validity and reliability. Nurs. Outlook 54 (1), 23–29. https://doi.org/10.1016/j.outlook.2004.12.004.
Holden, R.R., 2010. The Corsini Encyclopedia of Psychology - Face Validity. John Wiley & Sons, NJ: Hoboken. https://doi.org/10.1002/9780470479216.corpsy0341.
Hoyos, D., 2010. The state of the art of environmental valuation with discrete choice experiments. Ecol. Econ. 69 (8), 1595–1603. https://doi.org/10.1016/j.ecolecon.2010.04.011.
Janssen, E.M., Bridges, J.F., 2016. Nine tests for assessing validity of A discrete-choice experiment. Value Health 19 (7), A352. https://doi.org/10.1016/j.jval.2016.09.034.
Janssen, E.M., Marshall, D.A., Hauber, A.B., Bridges, J.F., 2017. Improving the quality of discrete-choice experiments in health: how can we assess validity and reliability? Expert Rev. Pharmacoecon. Outcomes Res. 17 (6), 531–542. https://doi.org/10.1080/14737167.2017.1389648.
Jiang, R., Pullenayegum, E., Shaw, J.W., Muhlbacher, A., Lee, T.A., Walton, S., Pickard, A.S., 2023. Comparison of preferences and data quality between discrete choice experiments conducted in online and face-to-face respondents. Med. Decis. Mak. 43 (6), 667–679. https://doi.org/10.1177/0272989X231171912.
Juschten, M., Omann, I., 2023. Evaluating the relevance, credibility and legitimacy of a novel participatory online tool. Environ. Sci. Pol. 146, 90–100. https://doi.org/10.1016/j.envsci.2023.05.001.
Li, L., Long, D., Rad, M.R., Sloggy, M.R., 2021. Stay-at-home orders and the willingness to stay home during the COVID-19 pandemic: a stated-preference discrete choice experiment. PLoS One 16 (7), e0253910. https://doi.org/10.1371/journal.pone.0253910, 1-20.
Lidwell, W., Holden, K., Butler, J., 2010. Universal Principles of Design, Revised and Updated: 125 Ways to Enhance Usability, Influence Perception, Increase Appeal, Make Better Design Decisions; and Teach Through Design. Rockport Publishers Inc, Beverly, MA.
Liebe, U., Meyerhoff, J., Hartje, V., 2012. Test–retest reliability of choice experiments in environmental valuation. Environ. Resour. Econ. 53, 389–407. https://doi.org/10.1007/s10640-012-9567-1.
Mansfield, C., Poulos, C., Boeri, M., Hauber, B., 2019. Pmu 129 performance of a comprehension question in discrete-choice experiment surveys (DCE). Value Health 22 (3), S730–S731.
Mansfield, C., Sikirica, M.V., Pugh, A., Poulos, C.M., Unmuessig, V., Morano, P., Martin, A.A., 2017. Patient preferences for attributes of type 2 diabetes mellitus medications in Germany and Spain: an online discrete-choice experiment survey. Diabetes Therapy 8, 1365–1378. https://doi.org/10.1007/s13300-017-0326-8.
Mariel, P., Hoyos, D., Meyerhoff, J., Thiene, M., 2021. Chapter 8: validity and reliability. In: Environmental Valuation with Discrete Choice Experiments. SpringerBriefs in Economics, pp. 111–123. https://doi.org/10.1007/978-3-030-62669-3_8.
Marshall, D.A., Deal, K., Conner-Spady, B., Bohm, E., Hawker II, G., Loucks, L., Noseworthy, T., 2018. How do patients trade-off surgeon choice and waiting times for total joint replacement: a discrete choice experiment. Osteoarthr. Cartil. 26 (4), 522–530. https://doi.org/10.1016/j.joca.2018.01.008.
McDonnell Feit, E., Beltramo, M.A., Feinberg, F.M., 2010. Reality check: combining choice experiments with market data to estimate the importance of product attributes. Manag. Sci. 56 (5), 785–800. https://doi.org/10.1287/mnsc.1090.1136.
Merlo, G., van Driel, M., Hall, L., 2020. Systematic review and validity assessment of methods used in discrete choice experiments of primary healthcare professionals. Health Economics Review 1–9. https://doi.org/10.1186/s13561-020-00295-8.
Mokkink, L.B., Terwee, C.B., Knol, D.L., Stratford, P.W., Alonso, J., Patrick, D.L., de Vet, H.C., 2010. The COSMIN checklist for evaluating the methodological quality of studies on measurement properties: a clarification of its content. BCM Medical Research Methodology 10 (22), 1–8. https://doi.org/10.1186/1471-2288-10-22.
Moores, K.L., Jones, G.L., Radley, S.C., 2012. Development of an instrument to measure face validity, feasibility and utility of patient questionnaire use during health care: the QQ10. Int. J. Qual. Health Care 24 (5), 517–524. https://doi.org/10.1093/intqhc/mzs051.
Morkbak, M.R., Olsen, S.B., 2015. A within-sample investigation of test–retest reliability in choice experiment surveys with real economic incentives. Agricultural and Resource Economics 59 (3), 375–392. https://doi.org/10.1111/1467-8489.12067.
Muhlbacher, A.C., Junker, U., Juhnke, C., Stemmler, E., Kohlmann, T., Leverkus, F., Nubling, M., 2015. Chronic pain patients' treatment preferences: a discrete-choice experiment preferences: a discrete-choice experiment. Eur. J. Health Econ. 16, 613–628. https://doi.org/10.1007/s10198-014-0614-4.
Naik-Panvelkar, P., Armour, C., Rose, J., Saini, B., 2012. Patients' value of asthma services in Australian pharmacies: the way ahead for asthma care. J. Asthma 49 (3), 310–316. https://doi.org/10.3109/02770903.2012.658130.
Nevo, B., 1985. Face validity revisited. J. Educ. Meas. 22 (4), 287–293. https://doi.org/10.1111/j.1745-3984.1985.tb01065.x.
Norman, R., Viney, R., Aaronson, N.K., Brazier, J.E., Cella, D., Costa, D.S., King, M.T., 2016. Using a discrete choice experiment to value the QLU-C10D: feasibility and sensitivity to presentation format. Qual. Life Res. 25, 637–649. https://doi.org/10.1007/s11136-015-1115-3.
Nouwens, S.P., Marceta, S.M., Bui, M., van Dijk, D.M., Groothuis-Oudshoorn, C.G., Veldwijk, J., de Bekker-Grob, E.W., 2025. The evolving landscape of discrete choice experiments in health economics: a systematic review. PharmaEconomics 1–58. https://doi.org/10.1007/s40273-025-01495-y.
Nunnally, J., Bernstein, I., 1994. Psychometric Theory, 3Rd Ed. McGraw-Hill, New York. https://doi.org/10.1177/014662169501900308.
Parvin, S., Wang, P., Uddin, J., 2016. Using best-worst scaling method to examine consumers' value preferences: a multidimensional perspective. Cogent Bus. Manag. 3 (1), 1199110. https://doi.org/10.1080/23311975.2016.1199110.
Pearce, A., Harrison, M., Watson, V., Street, D.J., Howard, K., Bansback, N., Bryan, S., 2021. Respondent understanding in discrete choice experiments: a scoping review. The Patient - Patient-Centered Outcomes Research 14, 17–53. https://doi.org/10.1007/s40271-020-00467-y.
Polit, D.F., Beck, C.T., 2006. The content validity index: are you sure you know what's being reported? Critique and recommendations. Res. Nurs. Health 29 (5), 489–497. https://doi.org/10.1002/nur.20147.
Que, S., Awuah-Offei, K., Weidner, N., Wang, Y., 2017. Discrete choice experiment validation: a resource project case study. Journal of Choice Modelling 22, 39–50. https://doi.org/10.1016/j.jocm.2017.01.006.
Rakotonarivo, O.S., Schaafsma, M., Hockley, N., 2016. A systematic review of the reliability and validity of discrete choice experiments in valuing non-market environmental goods. J. Environ. Manag. 183, 98–109. https://doi.org/10.1016/j.jenvman.2016.08.032.
Roberts, P., Priest, H., 2006. Reliability and validity in research. Nurs. Stand. 20 (44), 41–45.
Rowe, G., Frewer, L.J., 2004. Evaluating public-participation exercises: a research agenda. Sci. Technol. Hum. Val. 29 (4), 512–556. https://doi.org/10.1177/0162243903259197.
Royal, K., 2016. ‘‘Face validity’’ is not a legitimate type of validity evidence. Am. J. Surg. 212 (5), 1026–1027. https://doi.org/10.1016/j.amjsurg.2016.02.018.
Ryan, M., Bate, A., Eastmond, C.J., Ludbrook, A., 2001. Use of discrete choice experiments to elicit preferences. Quality in Health Care 10 (Suppl. I), i55–i60. https://doi.org/10.1136/qhc.0100055.
Schuster, A.L., Schuster, A.L., Crossnohere, N.L., Campoamor, N.B., Hollin, I.L., Bridges, J.F., 2024. The rise of best-worst scaling for prioritization: a transdisciplinary literature review. Journal of Choice Modelling 50, 100466. https://doi.org/10.1016/j.jocm.2023.100466.
Sicsic, J., Blondel, S., Chyderiotis, S., Langot, F., Mueller, J.E., 2023. Preferences for COVID-19 epidemic control measures among French adults: a discrete choice experiment. Eur. J. Health Econ. 24, 81–98. https://doi.org/10.1007/s10198-022-01454-w.
Soekhai, V., de Bekker-Grob, E.W., Ellis, A.R., Vass, C.M., 2019. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics 37, 201–226. https://doi.org/10.1007/s40273-018-0734-2.
Tada, Y., Ishii, J., Kimura, J., Hanada, K., 2019. Patient preference for biologic treatments of psoriasis in Japan. J. Dermatol. 46 (6), 466–477. https://doi.org/10.1111/1346-8138.14870.

Taherdoost, H., 2016. Validity and reliability of the research instrument; how to test the validation of a questionnaire/survey in a research. Int. J. Acad. Res. Manag. 5 (3), 28–36. https://doi.org/10.2139/ssrn.3205040.
Turner, S.P., 1979. The concept of face validity. Qual. Quantity 13, 85–90. https://doi.org/10.1007/BF00222826.
Veldwijk, J., van Exel, J., de Bekker-Grob, E.W., Mouter, N., 2023. Public preferences for introducing a COVID-19 certifcate: a discrete choice experiment in the Netherlands. Appl. Health Econ. Health Pol. 21, 603–614. https://doi.org/10.1007/s40258-023-00808-6.
Viney, R., Lancsar, E., Louviere, J., 2002. Discrete choice experiments to measure consumer preferences for health and healthcare. Expert Rev. Pharmacoecon. Outcomes Res. 2 (4), 89–96. https://doi.org/10.1586/14737167.2.4.319.
Vista, A.B., Rosenberger, R.S., Collins, A.R., 2009. If you provide It, will they read it? Response time effects in a choice experiment. Can. J. Agric. Econ. 57 (3), 365–377. https://doi.org/10.1111/j.1744-7976.2009.01156.x.
Whitty, J.A., Gonçalves, A.S., 2018. A systematic review comparing the acceptability, validity and concordance of discrete choice experiments and best–worst scaling for eliciting preferences in healthcare. Patient 11, 301–317. https://doi.org/10.1007/s40271-017-0288-y.
Willis, G.B., 2004. Cognitive Interviewing: a Tool for Improving Questionnaire Design. SAGE Publications. Thousant Oaks, CAL.
Yan, X., Levine, J., Zhao, X., 2019. Integrating ridesourcing services with public transit: an evaluation of traveler responses combining revealed and stated preference data. Transport. Res. C Emerg. Technol. 105, 683–696. https://doi.org/10.1016/j.trc.2018.07.029.