The Patient - Patient-Centered Outcomes Research (2022) 15:329–339
https://doi.org/10.1007/s40271-021-00551-x

ORIGINAL RESEARCH ARTICLE

A Guide to Observable Differences in Stated Preference Evidence

Benjamin Matthew Craig[1] · **Esther W. de Bekker-Grob**[2] · **Juan Marcos González Sepúlveda**[3] · **William H. Greene**[4]

Accepted: 9 September 2021 / Published online: 26 October 2021

Abstract

Background and Objective In health preference research, studies commonly hypothesize differences in parameters (i.e., differential or joint effects on attribute importance) and/or in choice predictions (marginal effects) by observable factors. Discrete choice experiments may be designed and conducted to test and estimate these observable differences. This guide covers how to explore and corroborate various observable differences in health preference evidence.

Methods The analytical process has three steps: analyze the exploratory data, analyze the confirmatory data, and interpret and disseminate the evidence. In this guide, we demonstrate the process using dual samples (where exploratory and confirmatory samples were collected from different sources) on 2020 US COVID-19 vaccination preferences; however, investigators may apply the same approach using split samples (i.e., single source).

Results The confirmatory analysis failed to reject ten of the 17 null hypotheses generated by the exploratory analysis ($p <$ 0.05). Apart from demographic, socioeconomic, and geographic differences, political independents and persons who have never been vaccinated against influenza are among those least likely to be vaccinated (0.838 and 0.872, respectively).

Conclusions For all researchers in health preference research, it is essential to know how to identify and corroborate observable differences. Once mastered, this skill may lead to more complex analyses of latent differences (e.g., latent classes, random parameters). This guide concludes with six questions that researchers may ask themselves when conducting such analyses or reviewing published findings of observable differences.

1 Introduction

Health preference research (HPR) refers to any investigation dedicated to understanding the value of health and health-related alternatives using observational or experimental methods [1]. In HPR, investigators conduct discrete choice experiments (DCEs), randomizing different choice sets to different individuals across multiple tasks to test hypotheses about the value of health and health-related alternatives [2]. Specifically, analyses of stated preference evidence can quantify the effects of the alternatives' attributes on

Key Points for Decision Makers

This guide describes how to explore and corroborate various observable differences in health preference evidence generally. Its worked example identified relevant differences in vaccination preferences against COVID-19.

Demographics and SES may help target vaccination outreach programs, such as engaging school boards and other organizations active in rural communities. Although the historical disparities by race and ethnicity merit recognition, they are not associated with differential effects in either the exploratory or the confirmatory results. Instead, programmatic resources may be directed to address disparities related to SES more generally.

Furthermore, political independents and persons who have never been vaccinated against influenza are among those least willing to be vaccinated against COVID-19. In response, the US CDC might create more educational programs that target groups with a high concentration of registered independents or reduced flu vaccinations.

✉ Benjamin Matthew Craig
bcraig@usf.edu

[1] Department of Economics, University of South Florida, CMC206A, 4202 E. Fowler Avenue, Tampa, FL 33620, USA

[2] Erasmus Choice Modelling Centre and Erasmus School of Health Policy and Management, Erasmus University Rotterdam, Rotterdam, The Netherlands

[3] Duke University, Durham, NC, USA

[4] Department of Economics, Stern School of Business, New York University, New York, NY, USA

preferential choice behaviors (i.e., attribute importance) [3, 4]. Currently, however, a clear guidance is lacking about how to examine differences in attribute importance by observable factors. For example, a study might examine how individuals with different observable characteristics (e.g., age) may make difference choices or how tasks with different observable characteristics (task sequence) may elicit different choices.

The objective of this paper is to provide guidance on how to explore and corroborate various observable differences in health preference evidence generally. To help readers, the guide introduces a worked example, including its decision context, definitions (see bolded terms and Glossary), and an overarching analytical process. For a more extensive coverage of choice modeling, we recommend the textbook *Applied Choice Analysis* by Hensher, Rose, and Greene [5].

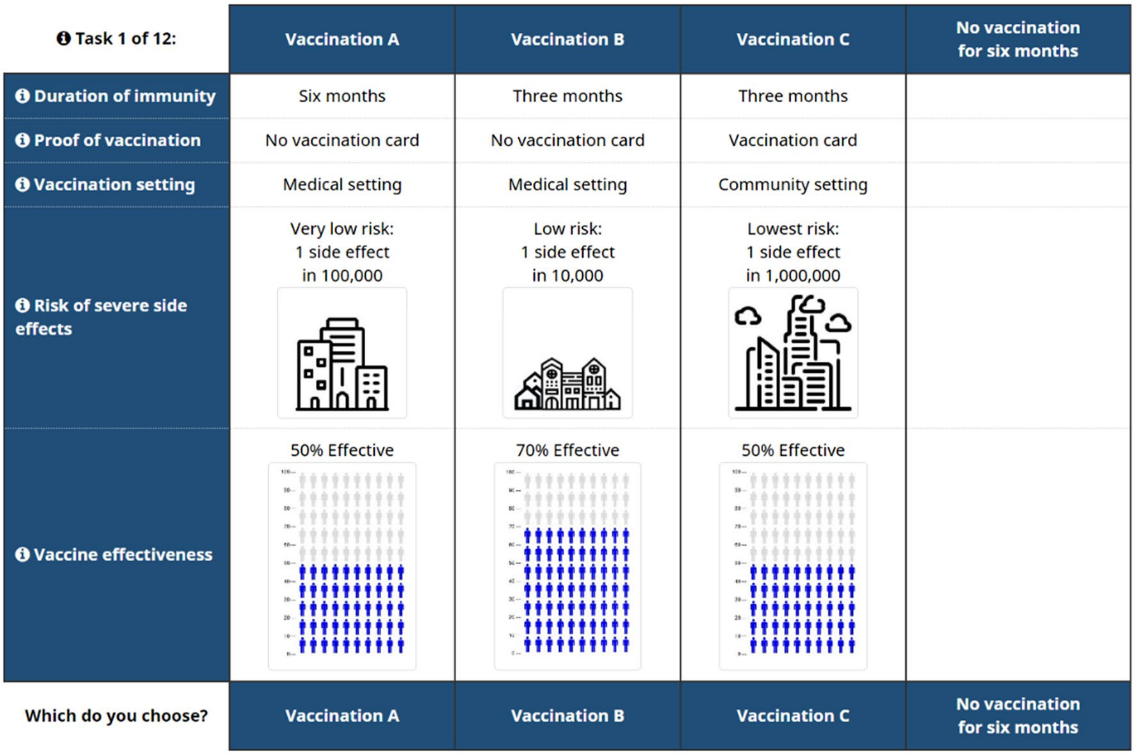
1.1 2020 US COVID-19 Vaccination Preferences (CVP) Study

As a worked example for this guide, we examined secondary data from the 2020 US COVID-19 vaccination preferences (CVP) study [6]. In brief, the 2020 US CVP study included a DCE with eight choice tasks as well as four kaizen tasks. The choice sets in each of the eight choice tasks (Fig. 1) included an opt-out ("no vaccination for six months") and three vaccination alternatives described using five attributes (see Glossary for their definitions):

1. Proof of vaccination (two nominal attribute levels): (1) Vaccination card; (2) No vaccination card;
2. Vaccination setting (two nominal attribute levels): (1) Medical setting; (2) Community setting;
3. Vaccine effectiveness (two ordinal attribute levels): (1) 70%; (2) 50%;
4. Duration of immunity (two ordinal attribute levels): (1) 6 months; (2) 3 months;
5. Risk of severe side effects (four ordinal attribute levels): (1) 1 per 1,000,000; (2) 1 per 100,000, (3) 1 per 10,000, (4) 1 per 1000.

Overall, the 2020 US CVP descriptive system delineated 64 vaccination alternatives ($2^4 \times 4$; i.e. all possible combinations of attribute levels). Based on the values from the initial analysis and for simplicity of presentation [6], each alternative may be expressed as a profile of attribute levels from the best (11111) to the worst (22224) vaccination.

Between the 9th and 12th of November 2020, the 2020 US CVP study recruited an exploratory sample of US adults from a marketing panel (Dynata®; 1153 respondents) and a confirmatory sample via crowdsourcing (Mturk®; 912 respondents). These surveys occurred simultaneously, prior



ℹ Task 1 of 12:	Vaccination A	Vaccination B	Vaccination C	No vaccination for six months
ℹ Duration of immunity	Six months	Three months	Three months	
ℹ Proof of vaccination	No vaccination card	No vaccination card	Vaccination card	
ℹ Vaccination setting	Medical setting	Medical setting	Community setting	
ℹ Risk of severe side effects	Very low risk: 1 side effect in 100,000	Low risk: 1 side effect in 10,000	Lowest risk: 1 side effect in 1,000,000	
ℹ Vaccine effectiveness	50% Effective	70% Effective	50% Effective	
Which do you choose?	Vaccination A	Vaccination B	Vaccination C	No vaccination for six months

Fig. 1 Choice task taken from the 2020 US COVID-19 Vaccination Preferences (CVP) Study

to US approval of any vaccines, but after clinical trial results were announced, promoting vaccination efficacy and safety [7].

Further details on the 2020 US CVP, including its study protocol and experimental design, have been published elsewhere [6, 8]. As a worked example, this guide shows how the value of the COVID-19 vaccinations (described using five attributes) differs systematically by nine observable factors in concordance with random utility theory.

1.2 Random Utility Theory and its Value Specification in DCE

Under random utility theory, each individual $i \in [1, \ldots, N]$ and each alternative $j \in [1, \ldots, J]$ has a utility U_{ij} such that $U_{ij} = V_j + \varepsilon_{ij}$ [9]. In the worked example, we normalize each utility U_{ij} by subtracting the utility of opt-out ("no vaccination for six months") and assume that the random terms ε_{ij} are distributed as type I extreme values [10]. Therefore, the systematic component V_j represents the value of a COVID-19 vaccination relative to no vaccination, and the probability function of vaccination choice j by individual i is a conditional logit, $\Pr\left(y_{ij} = 1\right) = p_{ij} = \frac{\exp\left(V_j\right)}{\sum_{k \in J} \exp(V_k)}$ [5]. Because of the normalization, the utility of the opt-out is zero by construction, which serves as a reference for the alternatives (e.g., negative values imply being worse than "no vaccination for six months").

The value of each vaccination alternative V_j represents the willingness of individual i to be vaccinated with alternative j. In this worked example, the value V_j was approximated by subtracting main-effect coefficients β_k from the value of the best vaccination α (11111). Each of the seven main-effect coefficients β_k represents a loss in value attributed to a worse level (i.e., attribute importance) [3]. The first four attributes have two levels, each representing a loss ($\beta_1, \beta_2, \beta_3$, and β_4), and the fifth attribute has four levels, representing up to three losses (β_5, β_6, and β_7). For example, imagine the worst vaccination in the CVP descriptive system (22224). Its value V_{22224} equals $\alpha - \beta_1 - \beta_2 - \beta_3 - \beta_4 - \beta_5 - \beta_6 - \beta_7$. A more general value specification using alternative specific constants (ASCs) is shown in the Electronic Supplementary Material (ESM), Online Resource 1.

In a DCE, each respondent completes multiple choice tasks $t \in [1, \ldots, T]$ and their error terms may therefore be correlated. In the worked example, the parameters of V (i.e., α, β) were estimated using evidence on preferential choice behaviors y_{ij} by maximum likelihood with respondent-specific clusters [4]. In other words, choice defines value [11]. In econometric notation, the hat symbol on a parameter indicates an estimate, such as $\hat{\alpha}$, as opposed to its true value α, which can be hypothesized but cannot be measured perfectly.

When describing the estimation results, the constant $\hat{\alpha}$ and main-effect coefficients $\hat{\beta}$ are known as fixed effects because each estimate is fixed across respondents and represents a causal relationship between the alternatives and preferential choice behaviors [5]. Each fixed effect may also be expressed by its effect on choice predictions $\hat{p}$ (i.e., marginal effect).

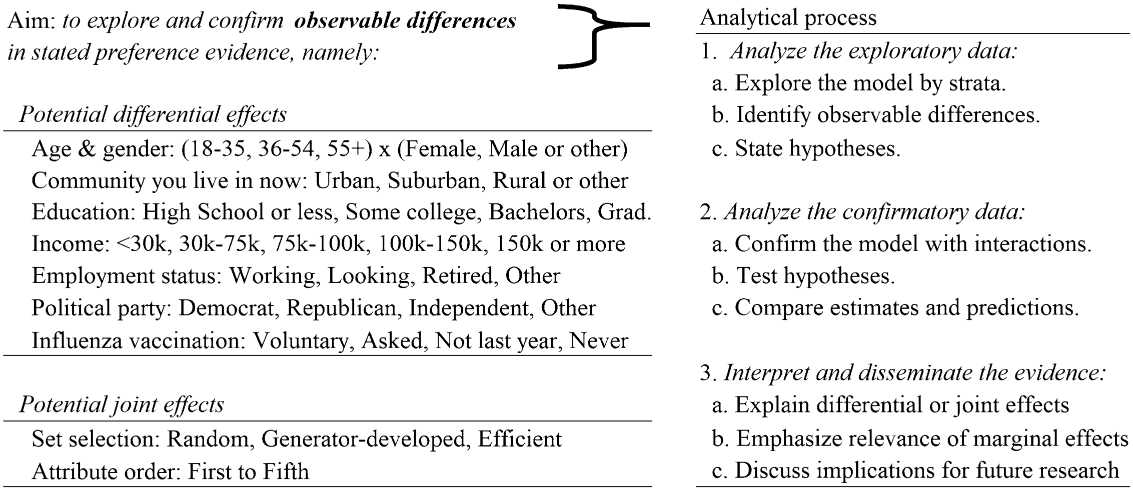
1.3 Observable Difference in Stated Preference Evidence

An observable difference is an estimated relationship between an observable factor Z and a fixed effect (e.g., $\hat{\alpha}, \hat{\beta}$) and represents evidence of preference heterogeneity (i.e., $\alpha, \beta | Z$). For example, does the value of the best COVID-19 vaccination α (11111) differ by age and sex? A future guide may examine observable differences in the proportional magnitude of all attributes (i.e., scale) or the ratio of two fixed effects (e.g., willingness to pay is the ratio of a fixed effect and the fixed effect of out-of-pocket price).

In this guide, each observable factor is categorical and measured explicitly, leaving little ambiguity about the groups. Although the measurement of observable factors is straightforward, the measured relationships between these factors and fixed effects depend in part on the factor distribution (e.g., multicollinearity, micronumerosity). Online Resource 2 describes a known-groups analysis that assesses the relationship between group size and statistical power given the intended effect size.

Latent factors, such as respondent attitudes, are not directly observable or reportable without the use of instruments that approximate their magnitude; therefore, the relationship between a latent factor and a fixed effect cannot be assessed directly. Estimations of latent differences may fail because of a measurement error of the latent factor or a lack of clarity in its definition. For example, a respondent's attitude may extend beyond a positive-negative scale to be multidimensional characterizing affect, behavior, and cognitive aspects [12]. While errors may occur in objective measurement, they are more common in subjective measurement [13]. Compared to observable factors, the models of latent factors, such as risk perception, vary by purpose and context and the groupings may seem vague. Future guides in HPR may cover how to examine differences by latent factors, such as latent classes, random effects, and correlated errors (that are violations of independence from irrelevant alternatives), and test for latent differences [5].

To explore and corroborate observable differences in stated preference evidence (Fig. 2), this guide introduces a three-step analytical process, starting with an exploratory analysis that generates hypotheses to be tested using the confirmatory sample. Based on the study protocol, we intended to recruit 1000 respondents for each sample and



Aim: *to explore and confirm* ***observable differences*** *in stated preference evidence, namely:*

Potential differential effects

- Age & gender: (18-35, 36-54, 55+) x (Female, Male or other)
- Community you live in now: Urban, Suburban, Rural or other
- Education: High School or less, Some college, Bachelors, Grad.
- Income: <30k, 30k-75k, 75k-100k, 100k-150k, 150k or more
- Employment status: Working, Looking, Retired, Other
- Political party: Democrat, Republican, Independent, Other
- Influenza vaccination: Voluntary, Asked, Not last year, Never

Potential joint effects

- Set selection: Random, Generator-developed, Efficient
- Attribute order: First to Fifth

Analytical process

1. *Analyze the exploratory data:*
   a. Explore the model by strata.
   b. Identify observable differences.
   c. State hypotheses.
2. *Analyze the confirmatory data:*
   a. Confirm the model with interactions.
   b. Test hypotheses.
   c. Compare estimates and predictions.
3. *Interpret and disseminate the evidence:*
   a. Explain differential or joint effects
   b. Emphasize relevance of marginal effects
   c. Discuss implications for future research

Fig. 2 Guide to observable differences in stated preference evidence

field identical surveys simultaneously from two sources to avoid temporal and single-source effects. Using nine observable factors taken from the worked example, we demonstrate the exploratory-confirmatory process (Fig. 2) and discuss its merits and limitations for future research.

1.4 Methods

1.4.1 Is a Confirmatory Sample Necessary to Infer Preference Heterogeneity?

A typical analysis of preference heterogeneity explores multiple factors, which confounds the classical interpretation of statistical uncertainty (e.g., *p* values, 95% confidence interval). When multiple models are estimated, the significance of any one parameter is unknown because some spurious relationships may appear significant by chance, and other substantive relationships may be hidden [14]. Presenting significant *p* values under one of the multiple exploratory models is known as cherry picking because such "cherries" can give a false impression of statistical inference [15]. Likewise, not controlling for a relationship in a model because its *p* value is slightly higher than a pre-defined threshold can lead to the omission of a substantive relationship for which the experiment was just not powered.

Instead of picking cherries, an exploratory analysis can generate hypotheses to be tested using a confirmatory sample. In this worked example, two samples were collected simultaneously from different sources so that the hypotheses generated using the exploratory sample could be tested using the confirmatory sample, potentially inferring observable differences (i.e., dual-sample process). As described in the Acknowledgments, the study design and hypotheses of this worked example were distributed to colleagues prior to the confirmatory analysis. Alternatively, a study may register their exploratory results and hypotheses on the Health Preferences Study and Technology Registry (hpstr.org).

Instead of using a dual-sample process, a researcher may split a sample from a single source into two sub-samples (exploratory and confirmatory); however, evidence from a **split-sample process** may be contaminated by the same sampling biases inherent to the single source. Obviously, it is easier to predict choices from the originating source than from an external source. Likewise, the dual-sample process implies that the results may or may not differ by source because of a sampling bias inherent to each source. Regardless of whether the process is completed using two sources or a split sample, it is important to compare the characteristics of the exploratory and confirmatory samples.

Collecting separate exploratory and confirmatory samples alone may not prevent biases in statistical inference. When researchers change their hypotheses based on confirmatory results, this again gives a false impression of statistical uncertainty (i.e., "the tail that wags the dog"). The fact that the model and hypotheses must be stated clearly prior to statistical inference is not specific to HPR [1].

To avoid such biases, the exploratory results (Table 1; Online Resources 3a and 4a) were distributed to various colleagues (see "Acknowledgments") prior to conducting the confirmatory analysis (Online Resource 4b). Alternatively, the exploratory results may be published in a peer-reviewed journal or registry. Simply placing the results online is not sufficient because these may be changed at any time and without notice.

Table 1 Observable differences in attribute importance: hypotheses and predictions using the exploratory data

8 tasks for 1153 respondents	Observable factors (α and β are shown on a log-odds scale by category)						
Hypotheses (H) and predictions (P)	Female			Male or other			*p* value
Age (years) and sex	18–34	35–54	55+	18–34	35–54	55+	
H01: Vaccination, α	2.266	1.880	2.592	2.632	2.157	2.906	0.021
H02: No card vs vaccination card, β_1	0.235	0.474	0.437	0.154	0.034	0.556	<0.001
H03: 50% vs 70% effective, β_3	0.992	1.069	1.424	0.931	0.852	1.576	<0.001
H04 Moderate vs low risk, β_7	0.200	0.597	0.572	0.273	0.230	0.574	0.010
P01: Vaccination, p	0.906	0.868	0.930	0.933	0.896	0.948	
P02: No card vs vaccination card, %	−0.022	−0.064	−0.034	−0.010	−0.003	−0.035	
P03: 50% vs 70% effective, %	−0.125	−0.175	−0.167	−0.087	−0.110	−0.157	
P04: Moderate vs low risk, %	−0.027	−0.112	−0.069	−0.028	−0.033	−0.055	

Community where you live now	Urban	Suburban	Rural or other	*p* value
H05: Vaccination, $\boldsymbol{\alpha}$	2.301	2.489	2.003	0.118
H06: 50% vs 70% effective, $\boldsymbol{\beta}_3$	0.878	1.280	1.155	<0.001
P05: Vaccination, $\boldsymbol{p}$	0.909	0.923	0.881	
P06: 50% vs 70% effective, %	−0.103	−0.153	−0.181	

Educational attainment	< High school	Some college	Bachelors	Graduate	*p* value
H07: Vaccination, $\boldsymbol{\alpha}$	1.636	2.266	2.734	2.575	<0.001
P07: Vaccination, $\boldsymbol{p}$	0.837	0.906	0.939	0.929	

Household income	Under 30k	30k−75k	75k−100k	100k−150k	150k+	*p* value
H08: Vaccination, $\boldsymbol{\alpha}$	1.932	2.207	2.379	2.659	3.096	0.006
P08: Vaccination, $\boldsymbol{p}$	0.874	0.901	0.915	0.935	0.957	

Employment status	Working	Looking	Retired	Other	*p* value
H09: Vaccination, $\boldsymbol{\alpha}$	2.403	2.032	2.764	1.783	0.004
H10: Community vs medical, $\boldsymbol{\beta}_2$	0.231	0.230	0.487	0.567	<0.001
H11: 50% vs 70% effective, $\boldsymbol{\beta}_3$	0.969	0.972	1.512	1.182	<0.001
P09: Vaccination, $\boldsymbol{p}$	0.917	0.884	0.941	0.856	
P10: Community vs medical, %	−0.019	−0.026	−0.034	−0.085	
P11: 50% vs 70% effective, %	−0.109	−0.141	−0.163	−0.210	

Political party affiliation	Democrat	Republican	Independent	Other	*p* value
H12: Vaccination, $\boldsymbol{\alpha}$	2.814	2.443	1.938	1.117	<0.001
P12: Vaccination, $\boldsymbol{p}$	0.943	0.920	0.874	0.753	

Influenza vaccination	Voluntary	Asked	Not last year	Never	*p* value
H13: Vaccination, $\boldsymbol{\alpha}$	3.179	3.084	2.327	1.468	<0.001
H14: 50% vs 70% effective, $\boldsymbol{\beta}_3$	1.499	0.879	1.221	0.806	<0.001
H15: Moderate vs low risk, $\boldsymbol{\beta}_7$	0.743	0.121	0.336	0.211	<0.001
P13: Vaccination, $\boldsymbol{p}$	0.960	0.956	0.911	0.813	
P14: 50% vs 70% effective, %	−0.117	−0.056	−0.160	−0.153	
P15: Moderate vs low risk, %	−0.062	−0.008	−0.045	−0.044	

Set selection	Random	Generator	Efficient	*p* value
H16: Moderate vs low risk, β_7	0.594	0.265	0.313	0.005
P16: Moderate vs low risk, %	−0.086	−0.035	−0.041	

Attribute order	1st	2nd	3rd	4th	5th	*p* value
H17: 50% vs 70% effective, β_3	1.000	1.097	0.962	1.185	1.332	0.028
P17: 50% vs 70% effective, %	−0.121	−0.137	−0.114	−0.153	−0.181	

1.4.2 Step 1. Analyze the Exploratory Sample

As described in the analytical process (Fig. 2), we first explore the results by strata starting with a known-groups analysis, separating the sample by known groups and assessing the relationship between group size and statistical power given the *p* value (0.05) [Online Resource 2]. From its findings, we inferred that 84 respondents per group is sufficient to identify observable differences when present for this specific model. Assuming that each group (or stratum) is of sufficient size, the exploratory analysis proceeds by estimating the fixed effects by each stratum of the observable factor, $\widehat{\alpha}, \widehat{\beta}|Z$. Like a grid search, stratification (1a) "casts a wide net" to systematically explore potential differences by each observable factor.

In this exploratory analysis, the stratified results were simplified using two identification thresholds (1b) followed by a joint Wald test. To identify potential differences in the fixed effects, we conducted a Wald test for each parameter and assessed whether its *p* value is less than 0.05. Unlike the likelihood ratio or Lagrange multiplier tests, Wald tests account for individual-specific clusters within the panel data [5].

To further assess the magnitude of the observable differences, we calculated the marginal effects (i.e., effect of attributes on choice predictions) by strata and assessed whether their range (i.e., the maximum effect minus the minimum effect among the strata) is greater than 0.05. Do the marginal effects differ substantively? Some observable differences may be statistically significant but have little influence on the choice predictions, and it is prudent to focus on the meaningful differences.

In the worked example, evidence that an observable difference passed these two identification thresholds ($p < 0.05$, range > 0.05) generates parameter-specific hypotheses (1c) for the confirmatory analysis (Table 1; Online Resource 4a). The selection of these two identification thresholds (or any alternative other threshold) was arbitrary and useful. A well-performed exploratory analysis can aid in informing the efficient allocation of scarce scientific resources by identifying potentially significant and meaningful relationships for further investigation.

Apart from these hypothesized relationships, a researcher may conduct a joint Wald test, testing whether all parameters are identical across strata simultaneously ($p < 0.05$). Unlike the parameter-specific tests and their two identification criteria, a significant *p* value on a joint test does not generate a hypothesis regarding an observable factor. However, an insignificant *p* value may generate a hypothesis of no differences by the observable factor (1c), which is demonstrated in this worked example.

1.4.3 Step 2. Analyze the Confirmatory Sample

The confirmatory analysis begins by comparing the exploratory and confirmatory samples (Online Resource 5) and estimating the differences in fixed effects using interactions (2a), instead of stratification (Table 2b, Online Resource 4b). As part of the confirmatory interaction analyses, we conducted a Wald test for each hypothesis, potentially corroborating an observable difference (2b). In addition to hypothesis testing, we compared the exploratory and confirmatory estimates (2c) to aid their interpretation. Next, we conducted the stratified analyses using the confirmatory sample (Online Resource 3b) and tested the hypotheses of no differences (Online Resource 2b). The stratified analyses may corroborate the absence of any observable differences or generate new hypotheses for further study.

1.4.4 Step 3. Interpret and Disseminate the Evidence

Once corroborated, each observable difference (or its absence) was interpreted and disseminated. A differential effect is an observable difference that is associated with an observable factor, such as respondents' age. A joint effect is an observable difference caused by an interaction of two or more randomized factors, such as task sequence. Differential effects may imply preference heterogeneity (e.g., differences in preference between groups), and joint effects may indicate a loss in internal validity, motivating improvements in experimental methods.

In a DCE, experimental factors (unrelated to the alternatives and decision context) may be randomly assigned to respondents and interacted with the indicators of specific alternatives or attribute levels to estimate joint effects. In the worked example, respondents were randomly assigned to experimental designs (random, generator developed, efficient), task sequences (first to eighth), object positions (left-middle-right), and attribute orders (first to fifth). When such experimental factors influence choices, the observable differences are unrelated to preference heterogeneity.

Even when corroborated, the relevance of an observable difference depends largely on the range of marginal effects, namely how much the factor influences the choice predictions (also known as effect size or magnitude). In the exploratory analysis, the second identification criterion is based on the range of marginal effects. Now that the observable effect is corroborated, these marginal effects have more practical implications. To better understand their relevance, the observable differences were ranked by the range of marginal effects and summarized for their broader implications in future research.

2 Results

2.1 Exploratory Results

In concordance with Fig. 2, we first conducted stratified analyses using the exploratory sample for all observable

Table 2 Observable differences in attribute importance: hypothesis and predictions using the confirmatory data

8 tasks for 912 respondents	Observable factors (α and β are shown on a log-odds scale by category)						
Hypotheses (H) and predictions (P)	Female			Male or other			p value
Age (years) and sex	18–34	35–54	55+	18–34	35–54	55+	
H01: Vaccination, α	3.246	1.824	2.789	2.761	2.463	3.124	0.002
H02: No card vs vaccination card, β_1	0.395	0.427	0.367	0.371	0.394	0.644	0.395
H03: 50% vs 70% effective, β_3	1.211	1.407	1.721	1.291	1.518	1.509	0.083
H04 Moderate vs low risk, β_7	0.827	0.507	0.605	0.475	0.457	1.063	0.029
P01: Vaccination, p	0.963	0.861	0.942	0.941	0.922	0.958	
P02: No card vs vaccination card, %	−0.017	−0.059	−0.024	−0.024	−0.034	−0.035	
P03: 50% vs 70% effective, %	−0.078	−0.258	−0.198	−0.127	−0.201	−0.124	
P04: Moderate vs low risk, %	−0.074	−0.101	−0.070	−0.054	−0.063	−0.116	
Community where you live now	Urban	Suburban	Rural or other				p value
H05: Vaccination, α	2.967	2.584	2.064				0.005
H06: 50% vs 70% effective, β_3	1.211	1.672	1.418				<0.001
P05: Vaccination, p	0.951	0.930	0.887				
P06: 50% vs 70% effective, %	−0.098	−0.216	−0.231				
Educational attainment	< High school	Some college	Bachelors	Graduate			p value
H07: Vaccination, α	1.885	2.198	2.686	3.939			<0.001
P07: Vaccination, p	0.868	0.900	0.936	0.981			
Household income	Under 30k	30k–75k	75k–100k	100k–150k	150k+		p value
H08: Vaccination, α	2.119	2.666	3.137	2.833	3.079		0.010
P08: Vaccination, p	0.893	0.935	0.958	0.944	0.956		
Employment status	Working	Looking	Retired	Other			p value
H09: Vaccination, α	2.594	2.986	3.062	2.102			0.212
H10: Community vs medical, β_2	0.214	0.077	0.260	0.354			0.366
H11: 50% vs 70% effective, β_3	1.420	1.099	1.726	1.593			0.061
P09: Vaccination, p	0.930	0.952	0.955	0.891			
P10: Community vs medical, %	−0.015	−0.004	−0.012	−0.039			
P11: 50% vs 70% effective, %	−0.166	−0.084	−0.163	−0.266			
Political party affiliation	Democrat	Republican	Independent	Other			p value
H12: Vaccination, α	3.471	2.711	1.641	0.582			<0.001
P12: Vaccination, p	0.970	0.938	0.838	0.642			
Influenza vaccination	Voluntary	Asked	Not last year	Never			p value
H13: Vaccination, α	3.728	3.808	2.366	1.923			<0.001
H14: 50% vs 70% effective, β_3	2.179	1.301	1.737	0.967			<0.001
H15: Moderate vs low risk, β_7	1.185	0.460	0.932	0.225			<0.001
P13: Vaccination, p	0.977	0.978	0.914	0.872			
P14: 50% vs 70% effective, %	−0.152	−0.054	−0.262	−0.150			
P15: Moderate vs low risk, %	−0.084	−0.022	−0.158	−0.040			
Set selection	Random	Generator	Efficient				p value
H16: Moderate vs low risk, β_7	0.705	0.470	0.595				0.227
P16: Moderate vs low risk, %	−0.096	−0.059	−0.078				
Attribute order	1st	2nd	3rd	4th	5th		p value
H17: 50% vs 70% effective, β_3	1.343	1.380	1.587	1.366	1.439		0.461
P17: 50% vs 70% effective, %	−0.153	−0.159	−0.198	−0.157	−0.170		

factors included in the 2020 US CVP survey instrument (1a). The full exploratory results of the stratified analyses are provided in Online Resource 3a. Only nine of the 14 analyses generated parameter-specific hypotheses based on the two identification thresholds (1b). Specifically, the analyses of nine observable factors (Fig. 2) generated 17 parameter-specific hypotheses (H01–H17) using the exploratory sample.

Next, we conducted nine interaction analyses using the exploratory sample (1b), one for each of the nine observable factors in Fig. 2. Its full results are provided in Online Resource 4a; however, Table 1 shows just the estimates of observable differences $\widehat{\alpha}$, $\widehat{\beta}|Z$ and the predictions (P01–P17) related to the 17 hypotheses (H01–H17). Among these hypotheses (1c), seven describe a relationship between the value of the COVID-19 vaccination α and the observable factor Z. The other ten hypotheses describe a relationship between the main-effects coefficients β and the observable factor Z. Table 1 also shows $\widehat{\alpha}$ as a choice prediction $\widehat{p}$ and each $\widehat{\beta}$ as a marginal-effect percentage.

Furthermore, the joint test results of the stratified analyses (Online Resource 3a) generated two hypotheses (H18–H19; 1c): no differences by US census region (Northeast, Midwest, South, West); no differences by marital status (Married or separated, Never married, Divorced, Other). The stratified results of the remaining three factors did not generate any hypotheses (i.e., race and ethnicity, task sequence, and object position) but may motivate further exploration. For example, if an observable factor is related to differences in scale (i.e., heteroskedasticity), the Wald tests of specific parameters may be insignificant, but the joint test may be significant.

2.2 Confirmatory Results

We first compared the exploratory and confirmatory samples (Online Resource 5) then conducted the confirmatory analysis. Table 2 shows the confirmatory results (2a), replicating the exploratory interaction analyses (Table 1). The full results of interaction and stratified analyses using the confirmatory sample (2b) are included in Online Resources 4b and 3b, respectively. Next, we highlight three key findings (2c) that were hypothesized by the exploratory analysis and corroborated by the confirmatory analysis.

First, vaccination uptake α is associated with respondent demographics and socioeconomic status (SES). Predicted uptake is significantly lower for persons of age 35–54 years, who reside in rural communities, with only a high school degree or less, and/or lower household income. The fact that age, sex, and SES are associated with lower uptake may not be surprising to some; however, it is noteworthy that these associations were corroborated, and the associations with race and ethnicity were not. The relationship with employment status was not corroborated, but this may be because of greater homogeneity in the confirmatory sample owing to its recruitment through crowd sourcing, instead of a marketing panel.

Second, vaccination uptake α is strongly associated with self-reported respondent behaviors, namely influenza vaccination and being unaffiliated with either political party. The associations between uptake and observable behaviors may be derived from a common source, for example, some persons who are immune to the influence of political or public health authorities (i.e., naysayers) may be reluctant, regardless of the vaccination's attributes. We did not control for demographics or SES in the estimation of these behavioral associations, which may diminish after taking them into account.

Third, the confirmatory analysis found little evidence that corroborates heterogeneity in any of the main-effect coefficients β. Effectiveness β_3 is lower among persons who reside in urban areas compared with other areas. Influenza vaccination is associated with the effects of both safety and efficacy (β_7 and β_3), such that persons who were asked to be vaccinated care less about its merits than others. For example, healthcare professionals (and others asked to be vaccinated against COVID-19 by their employers) might care less about its merits. The rest of the hypotheses on main effect coefficients β were not confirmed, but worth further investigation.

In this worked example, only ten of the 17 hypothesized differences (H1–H17) were corroborated ($p < 0.05$) and each represented a differential effect. This analysis did not confirm any joint effects that would suggest a lack of internal validity. We also did not find differences by US census region (H18: $p = 0.18$), but we found differences by marital status (H19: $p < 0.001$), which may be tested in a future study. The stratified analyses generated other new hypotheses: based on the two identification criteria, main-effects coefficients for safety and effectiveness (β_7 and β_3) may be associated with each of the five respondent characteristics as well as the two behavioral factors (influenza vaccination and political party affiliation).

3 Discussion

3.1 Interpretation of the Evidence

Overall, the worked example demonstrated three key findings about the heterogeneity in US COVID-19 vaccination preferences. The first result on demographics and SES may help target outreach programs, for example, engaging school boards and other organizations active in rural communities (3a). Although the historical disparities by race and ethnicity merit recognition, they are not associated with differential effects in either the exploratory or the confirmatory results. Instead, programmatic resources may be directed to address disparities related to SES more generally.

In the worked example, political independents and persons who have never been vaccinated against influenza are among those least likely to be vaccinated (0.838 and 0.872, respectively; 3b). In response, the authors believe that the Centers for Disease Control and Prevention might create more educational programs that target groups with a high concentration of registered independents or reduced flu vaccinations (e.g., college campuses, US states like Alaska and Maine). This targeting may be particularly relevant in preparation for the 2021–2022 influenza season.

3.2 Limitations

How useful is such a 2020 study when preferences on COVID-19 vaccination will likely change over time [7]? Temporal confounding motivated the simultaneous collection of the exploratory and confirmatory data in the worked example; however, it also implies that the evidence may not be generalizable to 2021 because vaccination preferences could have shifted as the context of the pandemic has evolved and people have more exposure to outcomes of COVID-19. If temporal confounding was not present, the confirmatory study would have been designed to test the hypotheses generated during the analysis of the exploratory sample. Each study team must assess its own temporal confounding as well as the wisdom of allocating scarce resources toward a simultaneous or subsequent confirmatory study.

The interpretation of observable differences, like these, may seem transparent, but they can also be overly simplistic and misleading. For example, heterogeneity in main-effect coefficients is not the same as heterogeneity in marginal effects because a marginal effect summarizes both the coefficient and the constant. The observable differences in main effects are usually found to be less meaningful than differences in the marginal effects. Likewise, an analyst may care about relative effects (i.e., ratios of attribute importance), such as willingness to pay or maximum acceptable risk. In some cases, attribute importance estimates may vary by an observable factor, but their ratio does not.

Furthermore, interactions imply an independence of the observable factors; however, these factors are likely correlated (e.g., SES and naysayer behaviors) because of a latent process. More advanced analysts in HPR may skip the estimation of observable differences and proceed directly to more complex methods that account for preference heterogeneity, such as random parameters or latent classes. For example, the proposed analytical process (Fig. 2) does not attempt to separate taste and scale heterogeneity [16].

When conducting exploratory and confirmatory studies in HPR or any other field, trust and order matters. You must trust that the study team followed its protocol, particularly the order of analyses. If we were to have used switched the order of the samples (i.e., recruited the exploratory sample via crowd sourcing and the confirmatory sample using the marketing panel), the conclusions would have been different. However, such a switch based on the results is not appropriate (i.e., tail that wags the dog).

Overall, the primary limitation of the worked example is its sampling frames. The exploratory sampling frame was a marketing panel and the confirmatory sampling frame was a crowd-sourcing vendor. Both tend to list more educated respondents who have means to participate in online surveys, which is not generalizable to the US general population. Any inference on observable differences must account for this sampling frame bias in its interpretation of the preference evidence. Although we could have re-weighted the results to imply gains in representativeness, this subterfuge may exacerbate existing biases. It is better to recognize the limitations of crowd sourcing, which may not be the best source to confirm results from a marketing panel.

4 Conclusions

This guide describes how to identify and corroborate observable differences using dual samples, which may or may not be affordable in other investigations. Unless underpowered (see Online Resource 2), every health preference study can split its sample and follow the analytical process described by the guide. Although advanced analyses of latent classes and random parameters are welcome, this analytical process (Fig. 2) provides a more principled approach to examining preference heterogeneity based on observable factors. When conducting such analyses or reviewing published findings of observable differences, researchers may consider the following questions:

1. How were the observable differences specified (e.g., statistical power)?
2. How were the observable differences estimated (e.g., a single interaction)?
3. How were the observable differences corroborated (e.g., dual or split samples)?
4. To aid interpretation:
    a. Was the observable factor randomized (differential vs joint effects)?
    b. Were the changes in the choice predictions meaningful (marginal effects)?
    c. What are the implications of these findings for future research?

Glossary

Alternative-specific constant (ASC) A parameter representing the value of a specific object.

Attribute importance A parameter representing the value of an object's attribute (dummy) or difference in attribute level (incremental) [3].

Cherry picking The presentation of significant p values under one of the multiple exploratory specification which can give a false impression of statistical inference [15].

Choice defines value The parameters of a value function V are estimated using empirical evidence on preferential choice behaviors y_{ij} [11].

Differential effect An observable difference that is associated with an observable factor.

Discrete choice experiments (DCEs) An experiment that randomly assigns different choice sets to different individuals to test hypotheses.

Dual-sample process The use of samples from two different sources for exploratory and confirmatory analyses.

Fixed effect A fixed parameter representing a causal relationship between alternatives and the preferential choice behaviors y_{ij}.

Health preference research (HPR) Any investigation dedicated to understanding the value of health and health-related alternatives using observational or experimental methods [1].

Interaction The product of two or more independent variables.

Joint effect An observable difference caused by an interaction of two or more randomized factors.

Known-groups analysis An analysis that separates a sample into groups known to have observable differences. A known-groups analysis is often conducted to assess whether pre-determined differences are observed under a variety of constraints (e.g., sample sizes). Likewise, an unknown-groups analysis separates a sample into groups without known differences to assess whether pre-determined differences are absent under a variety of constraints. Each analysis may identify potential causes of spurious results (e.g., a lack of statistical power).

Latent difference A relationship between a latent factor and a fixed effect that represents a specific form of preference heterogeneity (e.g., ASC by risk perception class).

Latent factor A categorical variable that is not directly observable or reportable without the use of instruments that approximate their magnitude; therefore, the relationship between a latent factor and a fixed effect cannot be assessed directly.

Marginal effect An observable difference in choice prediction.

Observable difference A relationship between an observable factor and a fixed effect that represents a specific form of preference heterogeneity (e.g., ASC by age group).

Observable factor A categorical variable that is measured explicitly, leaving little ambiguity about the groups; therefore, the relationship between an observable factor and a fixed effect may be assessed directly.

Statistical power The probability that a test will correctly reject a false null hypothesis. For a known-group analysis (Online Resource 2), 21 respondents per block was sufficient to identify the observable differences ($p < 0.05$) in over 80% of the bootstrap iterations.

Preferential choice behaviors A behavior y_{ij} that resolves ambiguity in preferences between objects in a set (i.e., choice set) [4]

Random utility theory Each individual $i \in N$ and each alternative $j \in J$ has a utility U_{ij} such that $U_{ij} = V_j + \varepsilon_{ij}$, where V_j are the alternatives' values and ε_{ij} are errors clustered by individual [9].

Relative attribute importance A ratio of two fixed effects where each represents attribute importance (i.e., the importance of one attribute relative to another).

Scale heterogeneity
A relationship between an observable or latent factor and the scale parameter, representing the proportional magnitude of all fixed effects.

Split-sample process
The separation of a sample from a single source into two sub-samples for exploratory and confirmatory analyses.

Stratification
The interaction between all variables with the same observable factor simultaneously, inherently separating the sample into groups (i.e., strata).

The tail that wags the dog
The practice of changing hypotheses based on confirmatory results that can give a false impression of statistical inference [15].

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40271-021-00551-x.

Acknowledgements We thank Catharina (Karin) G.M. Groothuis-Oudshoorn, Deborah Marshall, Semra Ozdemir, and Joffre D. Swait, who can attest that the authors sent their hypotheses and predictions to them on 22 March, 2021, declaring them openly on this date in anticipation of a future confirmatory analysis.

Declarations

Funding Benjamin M. Craig provided all financial support for the project.

Conflicts of interest/competing interests The authors have no conflicts to disclose.

Ethics approval Not applicable.

Consent to participate Not applicable.

Consent for publication Not applicable.

Availability of data and material The datasets generated and analyzed during the current study are available from the corresponding author on reasonable request.

Code availability The code generated used in the analysis of the study datasets is available from the corresponding author on reasonable request.

Author contributions As part of the original study, BMC developed the research questions, methodology, and design of the study; executed the protocol; and published the primary paper. For this extension, BMC, EWBG, JMGS, and WG conducted the secondary data analysis, interpreted the results, and wrote the final manuscript.

References

1. Craig BM, et al. Health preference research: an overview. Patient. 2017;10(4):507–10.
2. Soekhai V, et al. Discrete choice experiments in health economics: past, present and future. Pharmacoeconomics. 2019;37(2):201–26.
3. Gonzalez JM. A guide to measuring and interpreting attribute importance. Patient. 2019;12(3):287–95.
4. Coombs CH. A theory of data. New York: Wiley; 1964.
5. Hensher DA, Rose JM, Greene WH. Applied choice analysis: a primer, vol. XXiV. Cambridge: Cambridge University Press; 2005. p. 717.
6. Craig BM. United States COVID-19 vaccination preferences (CVP): 2020 hindsight. Patient. 2021;14(3):309–18.
7. Craig BM, González Sepúlveda JM, Johnson FR, et al. COVID-19 health preference research: four lessons learned. ISPOR Value Outcomes Spotlight. 2020;6(5):1–2.
8. Poteet S, Craig BM. QALYs for COVID-19: a comparison of US EQ-5D-5L value sets. Patient. 2021;14(3):339–45.
9. Luce RD. A theory of individual choice behavior. New York: Columbia University; 1957.
10. McFadden D. Conditional logit analysis of qualitative choice behavior. In: Zarembka P, editor. Frontiers in econometrics. New York: Academic Press; 1974. p. 105–42.
11. Jakubczyk M, et al. Choice defines value: a predictive modeling competition in health preference research. Value Health. 2018;21(2):229–38.
12. van Harreveld F, Nohlen HU, Schneider IK. Chapter Five: the ABC of ambivalence: affective, behavioral, and cognitive consequences of attitudinal conflict. In: Olson JM, Zanna MP, editors. Advances in experimental social psychology. New York: Academic Press; 2015. p. 285–324.
13. Nunnally JC, Bernstein IH. Psychometric theory. In: McGraw-Hill series in psychology, vol. XXiV. 3rd ed. New York: McGraw-Hill; 1994. p. 752.
14. Greene WH. Econometric analysis, 8th edition. Upper Saddle River: Pearson/Prentice Hall; 2018. p. 1178.
15. Greenland S, et al. Statistical tests, P values, confidence intervals, and power: a guide to misinterpretations. Eur J Epidemiol. 2016;31(4):337–50.
16. Groothuis-Oudshoorn CG, et al. Key issues and potential solutions for understanding healthcare preference heterogeneity free from patient-level scale confounds. Patient. 2018;11(5):463–6.