AMERICAN MARKETING ASSOCIATION

Article

Journal of Marketing Research
2022, Vol. 59(5) 963-982
© American Marketing Association 2022
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/00222437211073581
journals.sagepub.com/home/mrj
SAGE

The More You Ask, the Less You Get: When Additional Questions Hurt External Validity

Ye Li, Antonia Krefeld-Schwalb, Daniel G. Wall, Eric J. Johnson, Olivier Toubia, and Daniel M. Bartels

Abstract

Researchers and practitioners in marketing, economics, and public policy often use preference elicitation tasks to forecast real-world behaviors. These tasks typically ask a series of similarly structured questions. The authors posit that every time a respondent answers an additional elicitation question, two things happen: (1) they provide information about some parameter(s) of interest, such as their time preference or the partworth for a product attribute, and (2) the respondent increasingly "adapts" to the task—that is, using task-specific decision processes specialized for this task that may or may not apply to other tasks. Importantly, adaptation comes at the cost of potential mismatch between the task-specific decision process and real-world processes that generate the target behaviors, such that asking more questions can reduce external validity. The authors used mouse and eye tracking to trace decision processes in time preference measurement and conjoint choice tasks. Respondents increasingly relied on task-specific decision processes as more questions were asked, leading to reduced external validity for both related tasks and real-world behaviors. Importantly, the external validity of measured preferences peaked after as few as seven questions in both types of tasks. When measuring preferences, less can be more.

Keywords

preference elicitation, measurement, external validity, time preference, conjoint analysis

Online supplement: https://doi.org/10.1177/00222437211073581

Managers, policy makers, and researchers often elicit people's preferences in surveys to predict their behaviors in the field (Freeman, Herriges, and Kling 2014; Gustafsson, Herrmann, and Huber 2013; Netzer et al. 2008). From consumer surveys to conjoint analysis in marketing, and from measuring time and risk preferences in economics to contingent valuation in public policy, eliciting preferences to predict behaviors is important. Yet how many questions should we pose to respondents to maximize an elicitation task's external validity—that is, the ability to use the preferences measured in an elicitation task to make predictions about behaviors in other settings (Pearl and Bareinboim 2014)?

The typical goal of improving measurement precision suggests that more questions are better (Broomell and Bhatia 2014). Every time a survey respondent answers an elicitation question, we obtain additional information about some parameter(s) of interest, such as their temporal discount rate or partworths for product attributes. Although it may be tempting to assume that more data are always better—a stance that follows from information theory (Shannon 1948)—the "more is better" assumption only holds if the underlying data-generating process does not change (Ly et al. 2017). In practice, this may not be the case. Survey respondents' choices often violate the independence and stationarity assumptions of information theory (e.g., Birnbaum 2013).

We instead posit that the underlying decision processes respondents use to answer a series of elicitation questions may change, especially when those questions use a similar, repetitive format. Indeed, studies using eye tracking to trace how respondents process information in decision tasks have

Ye Li is Assistant Professor, Management and Marketing, School of Business, University of California, Riverside, USA (email: ye.li@ucr.edu). Antonia Krefeld-Schwalb is Assistant Professor, Marketing, Rotterdam School of Management, Erasmus University, Netherlands (email: krefeldschwalb@rsm.nl). Daniel G. Wall is Postdoctoral Fellow, University of Pennsylvania, USA (email: danwall@sas.upenn.edu). Eric J. Johnson is Norman Eig Professor of Business, Marketing, Graduate School of Business, Columbia University, USA (email: ejj3@gsb.columbia.edu). Olivier Toubia is Glaubinger Professor of Business, Marketing, Graduate School of Business, Columbia University, USA (email: ot2107@gsb.columbia.edu). Daniel M. Bartels is Professor, Marketing, Booth School of Business, University of Chicago, USA (email: bartels@uchicago.edu).

found that they tend to process less and less of the presented information with additional questions (Toubia et al. 2012; Yang, Toubia, and De Jong 2015, 2018). This reduction in information acquisition may happen because respondents increasingly rely on task-specific decision processes as they answer more questions, which we term "adaptation"—that is, respondents may change their information processing and decision making in ways that are specific to the task. For example, respondents may process less information, learn to weigh certain attributes more heavily, or adopt simplifying heuristics for combining attributes.

Importantly, the tasks that researchers and practitioners use for eliciting preferences are usually more repetitive, more structured, and substantially different from the real-world behaviors they are trying to predict using the elicited preferences. This means that respondents' adapted decision processes may mismatch the decision processes they use in the real-world behaviors that researchers are trying to predict.

For example, a conjoint choice task measuring new car preferences may display information on car price, fuel economy, safety ratings, warranty, country of manufacture side by side for easy comparison across several options. These features may initially receive similar weights in decisions, but respondents may learn to respond more efficiently as they answer additional questions by more heavily weighing one or two distinguishing features (e.g., manufactured domestically, five-star safety ratings). However, consumers at a car dealership may initially pay more attention to different features. For instance, they may focus on prominently displayed features such as price and fuel economy at first and eventually read the fine print to get a more complete picture. Moreover, consumers rarely make new car choices repeatedly over a short period of time and may compare only a few options, whereas respondents in a choice task typically make numerous repeated choices, each over many options.

Thus, asking additional questions in this example could decrease the external validity of the preferences measured in the conjoint choice task: as respondents adapt to the task and increasingly rely on task-specific decision processes across a series of similarly structured choices, their decision process increasingly mismatches the real-world decision processes for less repetitive decisions. In this article, we examine consequences of the trade-off between increasing measurement precision with more questions, on the one hand, and increasing mismatch in decision processes, on the other.

This trade-off between precision and mismatch is relevant for at least two related applications: (1) designing elicitation tasks and (2) testing theories. Marketers, psychologists, economists, and policy makers have expended considerable effort to understand how to best measure choices in elicitation tasks to predict real-world choices. In conjoint analysis, there has been concern regarding how many questions to ask, and practitioners have shown that how people weigh product attributes can change across questions within a single conjoint task. For example, brand becomes less important than price when the number of questions increases (Johnson and Orne 1996). A similar concern has led to the development of adaptive procedures to increase measurement validity (e.g., Green, Krieger, and Agarwal 1991; Toubia et al. 2003).

The second application is more general: to test theories, researchers in marketing and psychology often ask respondents to make many decisions because (1) complex theoretical models require more observations as the number of parameters increases; (2) techniques for measuring biobehavioral data (e.g., neural data, pupil dilation) demand many trials, often in the hundreds, to overcome physiological noise; and (3) there is an increasing interest in individual parameter estimation, which requires each respondent to make more choices. In both choice modeling and model testing, we are concerned that task-specific adaptation might place an unexpectedly low limit on how many questions we can ask respondents before encountering flat or even negative returns in external validity.

This is a general question, and in this article, we aim to understand the trade-off between more precise parameter estimates and respondents adapting to the specific task. In what follows, we first formalize how adaptation may influence the external validity of elicited preferences. Then, in four studies, we find that respondents adapt to preference elicitation tasks as they answer more questions, and that this adaptation can decrease the external validity of the measured preferences. We conclude by examining the implications for preference elicitation more generally and highlighting the importance of maximizing the match in underlying decision processes between the elicitation task and real-world behaviors.

Theoretical Development

Respondents' decision making may change in multiple ways as they adapt to an elicitation task. They may learn about the potential ranges of each attribute offered and where on the screen each attribute is displayed and thus improve the speed and efficiency of their information search (Brucks 1985; Johnson, Bellman, and Lohse 2003). They may learn which attributes they care more about (Dzyabura and Hauser 2019). They may adopt a heuristic—a lower-effort decision process that produces satisfactory responses (Gigerenzer and Gaissmaier 2011; Meißner, Musalem, and Huber 2016; Payne, Bettman, and Johnson 1988; Shah and Oppenheimer 2008)—for example, by considering less information or using simplified mathematical operations for comparing the options (Yang, Toubia, and De Jong 2015). Finally, they may become bored, demotivated, or fatigued with the task and cope by responding randomly (e.g., Howell, Ebbes, and Liechty 2021) or seeking variety (i.e., switching options for the sake of it; Inman 2001). For the purposes of our discussion, we remain agnostic to the exact form of adaptation; indeed, different respondents can adapt to the same elicitation task in different ways. The critical hypothesis is that respondents' decision processes change over time in ways that may affect the underlying preferences estimated for their choices.

Can adaptation affect the external validity of elicited preferences? That is, might changes in decision processes during an elicitation task lead to parameter estimates that are less able to predict behaviors in other settings? We argue that there are conditions in which the external validity of preferences estimated from an elicitation task could peak and subsequently decrease as respondents answer more elicitation questions. We focus on two countervailing forces that change as respondents answer additional questions: more questions increase the precision of parameter estimates (i.e., estimates converge toward some value), but they may also increasingly lead respondents to adapt their decision processes to the task. Because these adaptations are task specific, respondents' choices using the adapted process may be less reflective of their preferences in real-world behaviors. In other words, as respondents answer more questions, parameter estimates converge but toward values that reflect task-specific adapted decision processes that are potentially mismatched with the decision-making processes driving the behaviors researchers want to predict.[1] (For a stylized conceptual model that formalizes this discussion, see Web Appendix A.)

This reasoning generalizes to many types of elicitation tasks, but for the sake of illustration, we describe an example of measuring individual time preferences with the goal of predicting real-world intertemporal choices such as saving, smoking, or exercising. For example, a financial services firm may be interested in assessing time preferences to help predict who will repay their credit card debt on time. The standard economic analysis specifies that an individual's likelihood of repaying their credit card debt is determined at least in part by their temporal discount rate, d, the rate at which future outcomes are discounted relative to present outcomes. In this setting, an elicitation task would typically consist of a series of binary choices between smaller amounts of money available sooner and larger amounts available later. These choices can then be fit with a choice model to identify individual discount rates.

Consider, for example, Kable and Glimcher's (2007) study, in which each respondent was offered 144 choices between $20 now and delayed options ranging from $20.25 to $110 at delays of 6 hours to 180 days. Initially, respondents might evaluate all four pieces of information and calculate the rate of return. However, because respondents always face the same amount ($20) and delay (none) for the sooner option, they may adapt by learning to simply calculate the ratio of the later amount to its time delay (e.g., "I get $10 dollars a day for waiting")—a heuristic akin to that proposed by recent work (Marzilli Ericson et al. 2015; Scholten and Read 2010; Scholten, Read, and Sanborn 2014). This adapted decision process may efficiently produce reliable choices in this task but is unlikely to reflect how people make most real-world intertemporal choices, which involve sooner options that vary in amount and are not always immediately available.

Although we can estimate respondents' time preferences from their choices in the elicitation task, increasing task-specific adaptation with more questions means that choices are generated by decision processes that are increasingly mismatched with those used in the real-world behaviors we want to predict. For example, respondents' task-specific neglect of some of the presented information may not generalize to decision making in the real world. The benefit of increased precision in the parameters that comes with more questions might be diminished or even overwhelmed by increasing reliance on adapted decision processes.

Depending on the task format, we expect different dynamics in decision processes and preferences to emerge. While some attributes may gain importance in one task format, the same attributes may lose importance in another. For example, if delays in an intertemporal choice task are more prominent than amounts, respondents might adapt by increasingly comparing delays while neglecting amounts. However, if amounts are more prominent, respondents may adapt by increasingly comparing only amounts instead.

We thus hypothesize,

> **H_1:** Respondents adapt their decision processes as they answer more elicitation questions.
>
> **H_2:** Adaptation is task specific, reflecting idiosyncrasies of the elicitation task format.

Because H_1 and H_2 are likely to be true in sufficiently long tasks (Meißner, Musalem, and Huber 2016), we incorporated them into a stylized model (see Web Appendix A) that formalizes the two countervailing forces of increasing adaptation and increasing measurement precision with more questions asked. The model shows how these dynamics can impact the external validity of preference elicitation tasks and describes conditions under which we expect a peak in external validity, after which additional questions decrease validity. Thus,

> **H_3:** Adaptation in decision processes changes how respondents make choices and therefore impacts the preferences estimated from those choices.
>
> **H_4:** If the adapted decision process mismatches the decision process used in the predicted behavior, the external validity of elicitation tasks can peak and then decrease with more questions asked.

Overview of Studies

Although our hypotheses apply to any preference elicitation task, we focus on two important test cases: time preference measurement and conjoint analysis. We include the measurement of temporal discount rates in time preference elicitation tasks for several reasons: (1) They are among the most important and widely studied individual differences in the social sciences,

[1] Note that our claim is that elicitation tasks are *often* mismatched with real-world behaviors, but this is not always the case. Tasks can be designed to be high-fidelity simulations of the decisions people face in real-world situations, such as full-motion flight simulators. The reverse could also be true, such that routinized real-world decisions (e.g., grocery shopping, Netflix watching) might not be captured well by one-off decisions in an elicitation task, so that adaptation may actually increase match to the real world.

relating to behavior across a broad set of domains (e.g., health and financial decisions; Chabris et al. 2008; Reimers et al. 2009). (2) Time preferences have a large and growing literature of descriptive models (e.g., Frederick, Loewenstein, and O'Donoghue 2002; Marzilli Ericson et al. 2015; Scholten and Read 2010) and measurement methods (Cohen et al. 2020). (3) Time preferences have become increasingly important in marketing, in areas as diverse as consumer finance and food choice (Atlas, Johnson, and Payne 2017; Story et al. 2014).

As a second test case, we study conjoint analysis, for similar reasons: (1) Conjoint analysis is among the most important techniques in academic marketing and applied marketing research alike (Green and Srinivasan 1978, 1990). (2) There has been substantial literature on optimizing the statistical efficiency of conjoint choice tasks, and recent work has started using process tracing to better understand how respondents make choices in these tasks (Johnson, Meyer, and Ghose 1989; Meißner, Musalem, and Huber 2016; Yang, Toubia, and De Jong 2015, 2018). (3) Conjoint studies typically include a holdout sample that serves as a convenient measure of validity.

These two domains are complementary in that they span a broad range from simple to complex choices, and we are interested in how adaptation occurs in both. Time preference tasks often offer choices in which only four pieces of information are presented, a smaller outcome available sooner (e.g., $50 today) and a larger outcome available later (e.g., $60 in one month). Conjoint tasks, in contrast, typically employ complex displays of three or more choice options, each of which typically vary on up to ten attributes (e.g., Toubia et al. 2003).

We examine the existence of adaptation and its effects on external validity in these two domains across four studies. Study 1 tests H_1–H_3 by collecting process data to demonstrate task-specific adaptation in an intertemporal choice task. Studies 2a and 2b test H_4 by searching for peaks and subsequent decreases in the external validity of time preferences in two existing data sets. Study 2a examines the correlation of time preferences with an index of real-world behaviors and Study 2b does the same for predicting consumer credit scores. Study 3 tests H_1, H_3, and H_4 in a conjoint choice task by examining how decision processes change and how these changes affect external validity. Table 1 gives an overview of the studies and empirical evidence we observed for our hypotheses.

Study 1: Task-Specific Adaptation in Time Preference Elicitation

We designed Study 1 to test that adaptation occurs (H_1), that it is task specific (H_2), and that changes in decision processes relate to changes in preferences (H_3). To show that respondents adapt their decision processes as they answer more questions, we observed their information search processes by tracking mouse movements using MouselabWEB (Willemsen and Johnson 2010). This process-tracing technique is widely used in marketing, economics, and psychology (e.g., Costa-Gomes, Crawford, and Broseta 2001; Goldstein et al. 2014; Pachur et al. 2018; Schulte-Mecklenbeck et al. 2013). For choice tasks with relatively few options and attributes, such as intertemporal choice, MouselabWEB has minimal impact on the choice process and provides data that are analogous to eye-tracking methods (Lohse and Johnson 1996).

To test whether adaptation is task specific, we manipulated the task format, presenting delays as either days or the equivalent number of hours (e.g., 2 days vs. 48 hours). Because the trade-offs are identical across delay formats, any differences in decision processes or choices we find should be due to differences in task-specific adaptation. To strengthen this comparison, we also manipulated delay format within-participants, giving participants a second set of essentially identical choices (with slight jitter to prevent them from simply recalling their past responses) in a second consecutive session that used either the same or different delay formats.

This design enables us to examine three main predictions. First, respondents' decision processes will adapt as they answer more questions (H_1). As an indicator of respondents' adaptation, we expect search to become more comparative (i.e., comparing options within attribute) with an increase in partial neglect of some of the information presented. Comparative search has been associated with decision strategies aimed at reducing effort in known environments by comparing options on only the most relevant attributes (Payne 1976; Perkovic, Bown, and Kaptan 2018; Reisen, Hoffrage, and Mast 2008), and recent work suggests that it is prevalent in intertemporal choice tasks (Marzilli Ericson et al. 2015; Reeck, Wall, and Johnson 2017). Second, adaptation, and thus information search, will differ across the two delay formats, reflecting task-specific adaptation (H_2). In particular, we expected participants to compare the delay information more when it was presented as hours than as days, because the larger, less frequently encountered hour quantities would be more prominent (Coulter and Coulter 2005). Third, we expect these adaptations to be associated with changes in preferences (H_3), depending on which attribute is increasingly compared.

Methods

We recruited 353 participants from Amazon Mechanical Turk (MTurk; 47.3% female; age range = 18 to 74 years, M_{age} = 34.9 years). Because our analyses require complete data, we excluded 53 participants with incomplete data due to a programming error that caused participants to skip questions when clicking too quickly, leaving 300 participants for analysis.

For the intertemporal choice task, we constructed a set of 16 choices by crossing four sooner amounts ($21, $22, $24, and $26) at four delays (now, 1 day, 3 days, 7 days) with four larger amounts ($27, $29, $33, and $41) at four longer delays (11, 23, 34, and 45 days) in a partial factorial design (for full task details, see Web Appendix B1). Participants saw two back-to-back sessions of these 16 questions in a 2 (first session: day vs. hour) × 2 (second session: day vs. hour) between-subjects design that manipulated the format(s) in

Table 1. Study designs and hypotheses tested.

Study	Domain	Design	Manipulation	Process Measure	External Validity Measure	H_1	H_2	H_3	H_4
1	Time preference	Static	Time units	Mouse tracking	None	✓	✓	✓	
2a	Time preference	Adaptive	None	None	(1) Another time preference task (BLB) (2) Self-reported behavior with time–value trade-offs				✓
2b	Time preference	Adaptive	None	None	Credit scores				✓
3	Conjoint	Static	Position of external validity task	Eye tracking	Choice task with additional options	✓		✓	✓

Notes: BLB = set of 12 static intertemporal choices designed by Bartels, Li, and Bharti (2021).

which delays were presented. Delays were presented in either the day format or the equivalent number of hours, although we used "now" in both formats because "0 days" and "0 hours" have different connotations.

We randomized the order in which each participant saw the 16 questions in the first session and used the same order in the second session to hold any order and carryover effects constant. Maintaining the same question order allows us to test how choice consistency changes with question number. To disguise the equivalence of these choices, we "jittered" the dollar amounts by randomly adding or subtracting a small percentage (-2%, $-1\%, +1\%$, or $+2\%$) to each of the amounts for each choice. To make the task incentive aligned, we paid 1 in every 100 participants a bonus payment based on one of their choices selected at a random.

Results

To test our hypotheses, we examine changes in participants' decision processes by investigating the mouse-tracking data over time (H_1) and between task format conditions (H_2). We start by investigating trends in global search patterns (comparative vs. integrative search) and then focus on attribute-specific search changes in order to identify task-specific adaptation. Finally, we investigate whether preferences changed accordingly (H_3).

Decision process dynamics across questions (H_1) and formats (H_2). We first examined whether participants' decision processes changed across the 32 total questions in the two sessions of the elicitation task and whether adaptation differed between the format conditions. Although we cannot directly observe participants' decision processes per se, researchers commonly use participants' information search patterns as a proxy (e.g., Reeck, Wall, and Johnson 2017; Schulte-Mecklenbeck et al. 2017; Stillman, Shen, and Ferguson 2018).

Participants' information search decreased with more questions: the number of acquisitions (i.e., opening an information box) decreased from an average of 8.9 on the first question to 4.9 on the last question. Although the decrease is suggestive of adaptation, it could also reflect increasing familiarity with the task.

We thus turn to a more informative way to summarize information search in binary choice: the Payne Index, which is a commonly used measure of the relative amount of integrative versus comparative search (Payne, Bettman, and Johnson 1988). The Payne Index, which ranges from -1 to $+1$, is defined as the number of integrative transitions (i.e., moving between attributes within an option) minus the number of comparative transitions (i.e., moving between options on an attribute), divided by the total number of these transitions. Figure 1 plots the average Payne Index by question, session, and condition. The session 1 plot shows that Payne Index decreased with more questions asked, consistent with information search becoming more comparative, a trend that continued in session 2. Delay format also seemed to matter, with more comparative search (i.e., lower Payne Index) when delays were displayed as hours than as days, at least in session 1.

To test for differences in Payne Index across questions (H_1) and between conditions (H_2), we estimated regression models predicting Payne Index for participant i on questions q (1–32). These models must account for the fact that the session 1 delay format can influence the question effect in both sessions 1 and 2, while the session 2 delay format can only influence the question effect in session 2. We therefore introduced a fixed effect of delay format, format_{iq}, and a condition categorical variable for session 2 (cond_{iq} = day-day, day-hour, hour-day, or hour-hour), as well as two session dummy variables, Ses1 and Ses2. We also included interaction effects of question number with delay format in session 1 and with condition in session 2. To account for correlations between residuals within participant, our main model clustered standard errors by participant:

$$\begin{aligned}\text{Payne Index}_{iq} &= \beta_0 + \beta_1 q + \beta_{2,\text{format}_{iq}} + \beta_{3,\text{format}_{iq}} q \times \text{Ses1}_q \\ &\quad + \beta_{4,\text{cond}_{iq}} q \times \text{Ses2}_q + \epsilon_{iq}, \qquad (1)\end{aligned}$$

$$\text{where } \epsilon_{iq} = \gamma_i + \eta_{iq} \text{ and } \gamma_i \sim N(0, \sigma_\gamma^2), \quad \eta_{iq} \sim N(0, \sigma_\eta^2).$$

As a robustness check, we also estimated a generalized additive model with cubic regression splines, which flexibly accounts for

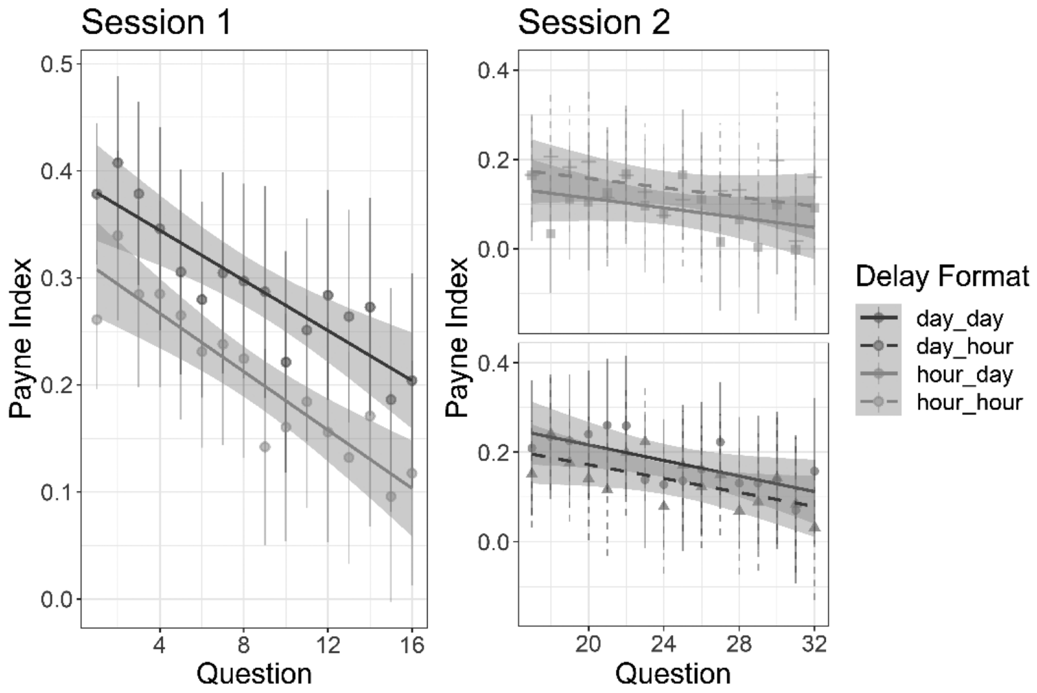


Figure 1. Average Payne Index as a function of delay format.

Notes: The Payne Index is calculated as (#integrative – #comparative)/(#integrative + #comparative). Smaller values correspond to more comparative search. The left plot illustrates the Payne Index across questions 1–16 in session 1, collapsed across the delay format in that session (days = gray and hours = black). The right plot illustrates the Payne Index across questions 17–32 in session 2, separately by the delay format in session 1. The error bars represent the 95% CI around the mean and the lines illustrate the linear effect of question number, with the gray regions illustrating the CI around the prediction.

potential nonlinearity in question number effects. This model did not include clustered standard errors.

$$\text{Payne Index}_{iq} = \beta_0 + f(q) + \beta_{2,\text{format}_{iq}} + f_{1,\text{format}_{iq}}(q) \times \text{Ses1}_q + f_{2,\text{cond}_{iq}}(q) \times \text{Ses2}_q + \epsilon_{iq}, \quad (2)$$

where $\epsilon_{iq} \sim N(0, \sigma^2)$.

We fit both of these models, as well as analogous models in subsequent analyses, in R version 4.1.1, using the *miceadds* package version 3.11-6 to estimate clustered errors and the *mgcv* package version 1.8-34 to estimate the spline regressions. To approach normality, we arctan-transformed the Payne Index for the analysis; results with untransformed dependent variables (DVs) are similar. Because both Models 1 and 2 led to similar conclusions, we present the results for Model 1; adding regression splines did not improve model performance (Bayesian information criterion $[\text{BIC}]_{\text{lm}} = 15{,}974$ vs. $\text{BIC}_{\text{splines}} = 15{,}998$; see Web Appendix Table B2).

Table 2 summarizes the main effects for this and subsequent analyses. Column 1 shows that the Payne Index decreased with question number ($\beta_1 = -.003$, $p = .086$) but was not significantly affected by delay format ($\beta_{2,\text{hour}} = -.028$, $p = .508$). That is, the arctan-transformed Payne Index decreased by .003 with every additional question.

To further explore the effect of delay format and facilitate the interpretation of coefficients, we estimated the marginal mean trends of question number on Payne index for the different combinations of session number and condition. That is, we calculated the predicted slopes of the Payne Index, Δ_{PI}, on question number in each session and condition, averaged across the remaining predictors.

Column 1 of Table 3 shows that the Payne Index significantly decreased in all four conditions and in both sessions, indicating more comparative search. The 95% confidence intervals (CIs) for the condition-specific trends largely overlapped between conditions. That is, participants' information search became more comparative, consistent with H_1, but was not sensitive to delay format overall, seemingly counter to H_2's prediction of task-specific adaptation.

Attribute-specific transitions (H_2). The fact that the Payne Index decreased with more questions suggests that participants increasingly shifted their decision process from integrating information within each option toward comparing attributes between options. This finding is a first indicator of participants' adaptation to the task (H_1). But the Payne Index treats all attribute transitions the same; it does not distinguish which attribute is being compared. Task-specific adaptation (H_2) could manifest via increasing reliance on comparing just one of the attributes. Figure 2 shows how the proportions of amount and delay transitions changed over time.

To examine attribute-specific adaptation, we estimated models analogous to Equations 1 and 2 but using the proportions

Table 2. Effects of Question Number and Delay Format on Search and Choices in Study 1.

	(1) Payne Index		(2) Prop. Amount Transitions		(3) Prop. Time Transitions		(4) Larger-Later Choices	
DV: Coefficient	Est.	*p*	Est.	*p*	Est.	*p*	Est.	*p*
β_0	.2955	<.0001	.3355	<.0001	.2591	<.0001	.2229	.0622
β_1 (question)	−.0034	.0857	−.0011	.4180	.0012	.2432	−.0103	.1394
$\beta_{2,hour}$	−.0277	.5082	.0342	.2354	.0171	.4663	−.6729	<.0001
$\beta_{3,day}$	−.0049	.1275	.0052	.0311	.0006	.7414	.0025	.7916
$\beta_{3,hour}$	−.0058	.0166	.0018	.3231	.0035	.0066	.0044	.5696
$\beta_{4,day\text{-}day}$	−.0027	.3367	.0041	.0632	.0005	.7352	.0052	.5312
$\beta_{4,day\text{-}hour}$	−.0049	.1063	.0053	.0215	.0015	.3405	.0325	.0027
$\beta_{4,hour\text{-}day}$	−.0050	.1333	.0055	.0186	.0021	.2373	.0027	.8198

Notes: Models clustered standard errors per participant. The Payne Index was arctan-transformed, and the proportions of amount and time transitions were arcsine-transformed for these analyses. Larger-later choice was a logistic regression.

Table 3. Marginal Mean Trends of Question Number on Information Search and Choice in Study 1.

		(1) Payne Index Δ_{PI}		(2) Prop. Amount Transitions Δ_{Amt}		(3) Prop. Time Transitions Δ_{Time}		(4) Larger-later Choices Δ_{LL}	
Session	Cond.	Est.	95% CI	Est.	95% CI	Est.	95% CI	Est.	95% CI
1	Day	−.0074	[−.011, −.004]	.0034	[.001, .006]	.0020	[.000, .004]	−.0011	[−.003, .001]
	Hour	−.0079	[−.011, −.005]	.0017	[.000, .004]	.0035	[.002, .005]	−.0006	[−.003, .001]
2	Day-day	−.0074	[−.011, −.004]	.0027	[.000, .005]	.0025	[.001, .004]	−.0014	[−.003, .000]
	Day-hour	−.0085	[−.012, −.005]	.0033	[.001, .006]	.0029	[.001, .005]	.0017	[−.001, .004]
	Hour-day	−.0085	[−.012, −.005]	.0034	[.001, .006]	.0032	[.001, .005]	−.0017	[−.004, .001]
	Hour-hour	−.0060	[−.009, −.003]	.00065	[−.001, .003]	.0022	[.001, .004]	−.0020	[−.004, .000]

of amount and delay transitions as dependent variables, both arcsine-transformed to approach normality. We show results for these models in columns 2 and 3 in Tables 1 and 2; results with untransformed DVs are similar. In line with H_2, we expected task-specific adaptation to manifest in terms of different trends for amount and delay transitions across the different delay formats. Linear and spline models again performed similarly, so we focus on the linear model with clustered standard errors. Overall, we observed no main effect of question number on the proportion of transitions for delays ($\beta_1 = .001$, $p = .243$) or for amounts ($\beta_1 = -.001$, $p = .418$). More importantly, we observed significant interactions with delay format in session 1 for both amount and delay transitions. Table 3 shows the slopes by condition, showing that amount transitions increased more in the day format ($\beta_{3,\text{day}} = .005$, $p = .031$) than in the hour format ($\beta_{3,\text{hour}} = .002$, $p = .323$; $\Delta_{\text{Amt[format=day;Ses1=1]}} - \Delta_{\text{Amt[format=hour;Ses1=1]}} = .002$) in session 1. The opposite was true for delay transitions, with delay transitions increasing more in the hour format ($\beta_{3,\text{hour}} = .004$, $p = .007$) than in the day format ($\beta_{3,\text{day}} = .001$, $p = .741$, $\Delta_{\text{Amt[format=day;Ses1=1]}} - \Delta_{\text{Amt[format=hour;Ses1=1]}} = -.002$) in session 1.

These results provide further evidence that adaptation is task specific (H_2); while search trended increasingly comparative overall, the delay format influenced *which* attribute was increasingly compared. These differences in the search (and presumably decision-making) process should have consequences for the preferences observed in the task, a topic we explore next.

Change in preferences (H_3). We next turn to the choices participants made to examine whether changes in search were associated with changes in preferences. Reeck, Wall, and Johnson (2017) suggested that more comparative search is associated with more patient choices; however, they did not distinguish between amount and delay transitions. We reasoned that comparing amounts should predict that a person will choose the larger amount and thus make more patient choices, whereas comparing delays should lead to a preference for the shorter delay and thus less patient choices.

From this reasoning, we expected participants' choices to become more patient (i.e., more likely to choose the large-later option) with more questions when delays were presented as days, due to the increase in amount transitions as the number of questions increases. However, we expected this effect to weaken or even reverse when delays were presented as hours, due to the increase in delay transitions as the number of questions increases.

To test these predictions, we estimated models analogous to those in Equations 1 and 2 with a logit link function to account for the binary outcome variable. Again, the model with a linear effect of question number with clustered standard errors and the

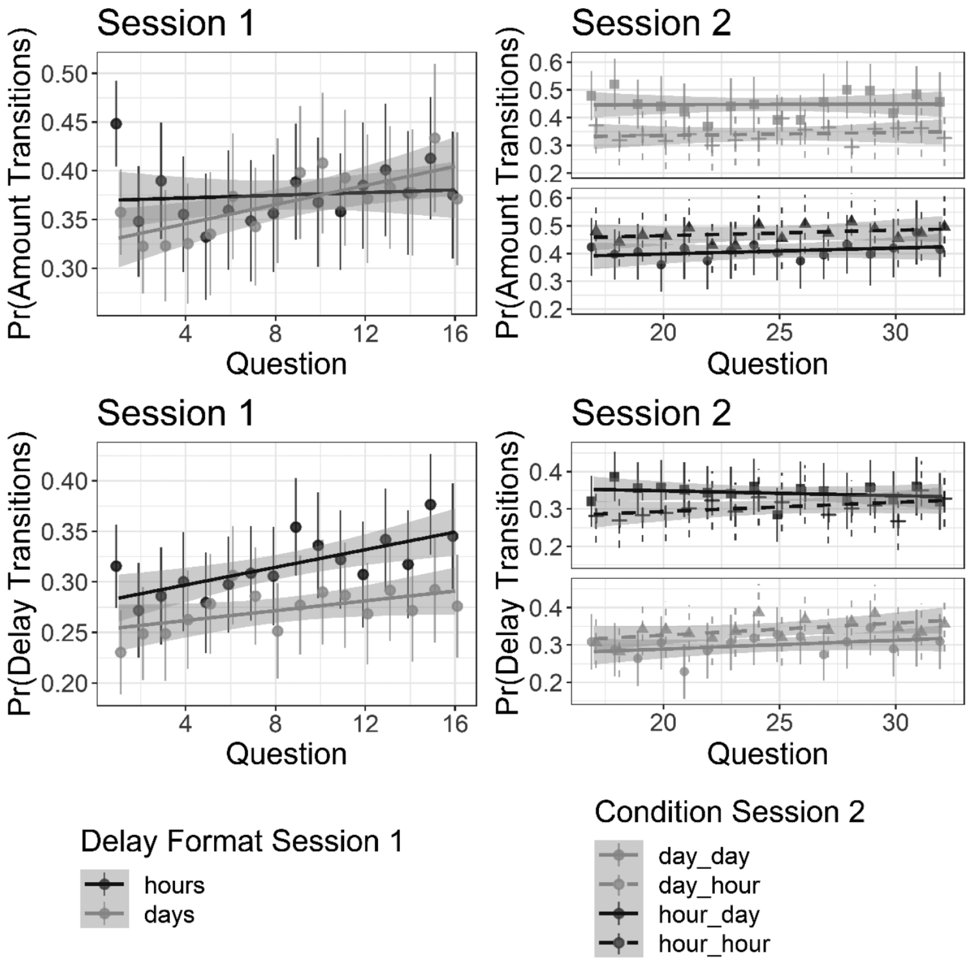


Figure 2. Proportion of amount (upper row) and delay (bottom row) comparative transitions as a function of the delay formats.
Notes: The left plots illustrate proportions of each type of transition across questions 1–16 in session 1. The plots on the right do the same for questions 17–32 in session 2, split by the delay format in session 1. The error bars represent the 95% CI around the mean, and the lines illustrate the linear effect of question number, with the gray regions illustrating the CI around the prediction.

spline model fit the data similarly well (see Web Appendix Table B2). So, we present results from only the linear model. Model 4 of Table 2 shows that participants made less patient choices (i.e., lower probability of choosing larger-later option) when delays were displayed as hours than as days ($\beta_{2,\text{hour}} = -.673$, $p < .001$), which is consistent with H_3 and with the increase in delay transitions in that format.

Recall that we expected differences in decision processes to result in changes in choices. The nature of the changes in choices, however, will depend on how decision processes change across questions. To clarify the relationship between decision process changes and choice changes, we investigated whether each participant's changes in proportion of amount and delay transitions would correlate with changes in choices. We thus estimated individual-level changes in larger-later choices and in proportions of amount and delay transitions by extracting the linear random slopes of question number from mixed models on these variables (with a logit link for predicting larger-later choices; models with random slopes and intercepts per participant were estimated using the R package *lme4*).[2] We then computed the correlation of the random slopes, S_{1i}, across the variables. As we expected, participants who increased their proportion of amount transitions more also increased more in patience ($r = .31$, $p < .0001$), while participants who increased their proportion of delay transitions more also decreased more in patience ($r = -.07$, $p = .25$), although this latter effect was not significant. Taken together, these results support H_3, that changes in decision processes are associated with changes in preferences.

[2] We used the following model to estimate the individual slopes for the proportion of amount transitions: $\text{pr(Amount Transitions)}_{iq} = \beta_0 + S_{0i} + (\beta_1 + S_{1i})q + \beta_{2,\text{format}_{iq}} + \beta_{3,\text{format}_{iq}} q \times \text{Ses1}_q + \beta_{4,\text{cond}_{iq}} q \times \text{Ses2}_q + \epsilon_{iq}$, where $\epsilon_{iq} \sim N(0, \sigma^2)$. We used the same model for delay transitions and larger-later choices with a logit link function. Because we are not testing the significance of the main effect coefficients or comparing marginal means or marginal trends, clustered standard errors are unnecessary. Correlating the individual slopes across models, as we do here, does not depend on the standard error of the estimates.

Cognitive toolbox model. Finally, as an alternative analysis, we also implemented a model that analyzes search and choice data together. In particular, we implemented a Bayesian toolbox model (Scheibehenne, Rieskamp, and Wagenmakers 2013) to jointly fit the search and choice data to identify strategy use and measure systematic strategy shifts across questions. The model assumes that decision strategies are associated with certain patterns of information search. For example, a participant might compare the two amounts and choose the option with the larger amount in question 1 but might compare the delays and choose the option with the smaller delay in question 2. Studying which decision strategies became more or less likely with more questions revealed similar results. We found that comparative strategies became more likely, and integrative strategies became less likely (H_1), while the patterns differed between delay formats (H_2). In particular, a strategy in which respondents chose only based on comparing delays became more prominent in the hour format compared with the day format. The toolbox model results further supported our inference that one reason why we do not observe a question number effect on preferences is that adaptation may differ between conditions and between participants, which can lead to opposite effects on choices. For full details, see Web Appendix B3.

Discussion

Study 1 finds that the decision processes underlying intertemporal choices shifted from more integrative toward more comparative decision strategies as participants made more choices. Furthermore, participants' adaptations were task specific: when delays were presented as hours, participants made increasingly more delay transitions and increasingly less patient choices compared with when the same delays were presented as days.

We next extend our results to the potential downsides of adaptation for an elicitation task's external validity. We thus turned to an elicitation method designed to provide precise, valid estimates of time preferences, something our elicitation task in Study 1 was not designed to do.

Studies 2a and 2b: External Validity of Time Preferences

In Studies 2a and 2b, we tested H_4 by examining the external validity of time preference estimates for a possible peak followed by a decrease with the number of questions asked. We hypothesized that this decrease results from participants adapting their decision-making processes to the specific task. That is, additional questions would lead to participants increasingly relying on a task-specific decision process that mismatches the decision processes used in other elicitation tasks and in real-world decisions.

Studies 2a and 2b both used an established time preference elicitation task, the Dynamic Experiments for Estimating Preferences (DEEP) Time task (Toubia et al. 2013), in which respondents answer a series of 20 binary intertemporal choices dynamically selected to maximize their informativeness for estimating time preferences. Toubia et al. (2013) found that time preferences estimated from DEEP Time have higher external validity than those estimated using other common time preference elicitation tasks, while also taking fewer questions to collect. It is important to note that for our purposes, adaptive elicitation tasks offer an important advantage: they should provide more information about the underlying parameters for each question asked compared with static elicitation tasks. Adaptive tasks should provide, in theory, the best chances of parameter identification before any adaptation occurs.

Using two data sets, we analyzed three external validity measures: (1) a different intertemporal choice task, (2) an index of self-reported behaviors that potentially involve trade-offs between costs and rewards over time, and (3) the respondents' subsequent credit scores.

Methods

The two studies used different measures of external validity. The first data set (Study 2a) was part of a large study on how time preferences relate to real-world intertemporal choice behaviors (Bartels, Li, and Bharti 2021). The 1,308 participants (41.4% female; age range = 18 to 86 years, $M_{age} = 40.9$ years) included 603 recruited from MTurk and 705 recruited from a market research firm.[3] We look at two sets of responses. The first consists of 12 static intertemporal choices designed by Bartels, Li, and Bharti (2021) using item response theory (we refer to this as the BLB task; see Web Appendix C1 for details). The second consists of participants' self-reports of the degree to which they exhibited a variety of 36 behaviors involving trade-offs between costs or benefits that occur across time and are thus theoretically related to time preferences (e.g., flossing, smoking, credit card repayment; see Web Appendix C2). Participants completed the DEEP Time task afterward.

Study 2b used a community sample of 478 participants who were recruited as part of a larger project on decision making across the life span (for more detail, see Li et al. [2015]). Participants ranged in age from 18 to 86 years, with roughly equal numbers in the 18–30, 31–45, 46–60, and over 60 age groups. All participants completed the DEEP Time task, and participants' credit scores were obtained from a major

[3] MTurk participants were younger and more educated, and a higher percentage were male compared with participants from the market research firm. Because participant characteristics were not the focus of our article, our analyses did not incorporate a panel variable or other demographics. As a robustness check, we also conducted the analysis separately on both subsamples. Although we replicated the results in both subsamples for the Bartels, Li, and Bharti (2021) time preference measure, the correlation between time preference and self-reported intertemporal choice behaviors was smaller in the market research subsample, which may explain why we did not replicate the same peak pattern in the market research subsample.

credit-reporting bureau for 417 of the participants (with informed consent). Time preferences have been shown to be predictive of credit scores (Li et al. 2015; Meier and Sprenger 2012).

Results

DEEP Time preference estimation. For both data sets, we used the hierarchical Bayes approach outlined by Toubia et al. (2013) to estimate the two parameters of the quasihyperbolic discounting model ($d(t) = \beta\delta^t$ for $t > 0$, $d(t) = 1$ for $t = 0$; Laibson 1997): β, present bias (i.e., how much any amount of delay from the present discounts values) and δ, the exponential discount factor (i.e., the proportion of value an outcome retains as it is delayed from the present—essentially the inverse of discount rate).

To estimate the evolution of preferences as the number of questions increases, we estimated β and δ after each of the 20 DEEP Time questions. That is, we estimated parameters after only the first DEEP Time question, the first 2 questions, and so on, up to all 20 questions, thus generating 20 pairs of time preference estimates for each participant, δ_{iq} and β_{iq}, for q from 1 to 20. Estimation based on only a few questions is possible due to DEEP's design combined with the hierarchical Bayesian implementation of the quasihyperbolic discounting model (for details, see Toubia et al. [2013]).

Study 2a: External validity for another time preference measure. To assess how external validity evolved with more questions, we used the DEEP Time preference estimates after each question, δ_{iq} and β_{iq}, to predict the time preferences derived from the BLB task, BLB_i, which were simple counts of the number of larger, later choices out of the 12 questions on the BLB task. We ran separate linear mixed models to predict BLB_i using the 20 estimates of δ for each participant, δ_{iq}, and did the same for $\beta_{iq.}$, with parameters standardized per question. Thus, we repeatedly estimated the following model for each q.

$$BLB_i = b_0 + b_1\delta_{iq} + \epsilon_{iq}, \tag{3}$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma^2_\gamma)$, $\eta_{iq} \sim N(0, \sigma^2_\eta)$.

Taking each models' predictions, $\widehat{BLB}_{iq}$, we next calculated the absolute percentage error (APE) per individual. These are essentially scaled residuals for predicting the BLB task, such that larger APEs correspond to lower external validity.

$$APE^{\delta}_{BLB_{iq}} = \left|\frac{\widehat{BLB}_{iq} - BLB_i}{BLB_i}\right|. \tag{4}$$

We used the APEs to compare the external validity of time preference estimates after 1 DEEP question, after 2 DEEP questions, and so on, up to all 20 DEEP questions. To construct CIs for external validity, we estimated a model including main effects for question number q (treated as a factor; i.e., we estimated a separate coefficient for each question number) and standard errors clustered by participant to account for correlated residuals. The APEs were arctan-transformed to approach normality. Results were similar with log-transforms and without transformations.

$$\arctan(APE^{\delta}_{BLB_{iq}}) = b_0 + b_{1,q} + \epsilon_{iq}, \tag{5}$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma^2_\gamma)$, $\eta_{iq} \sim N(0, \sigma^2_\eta)$.

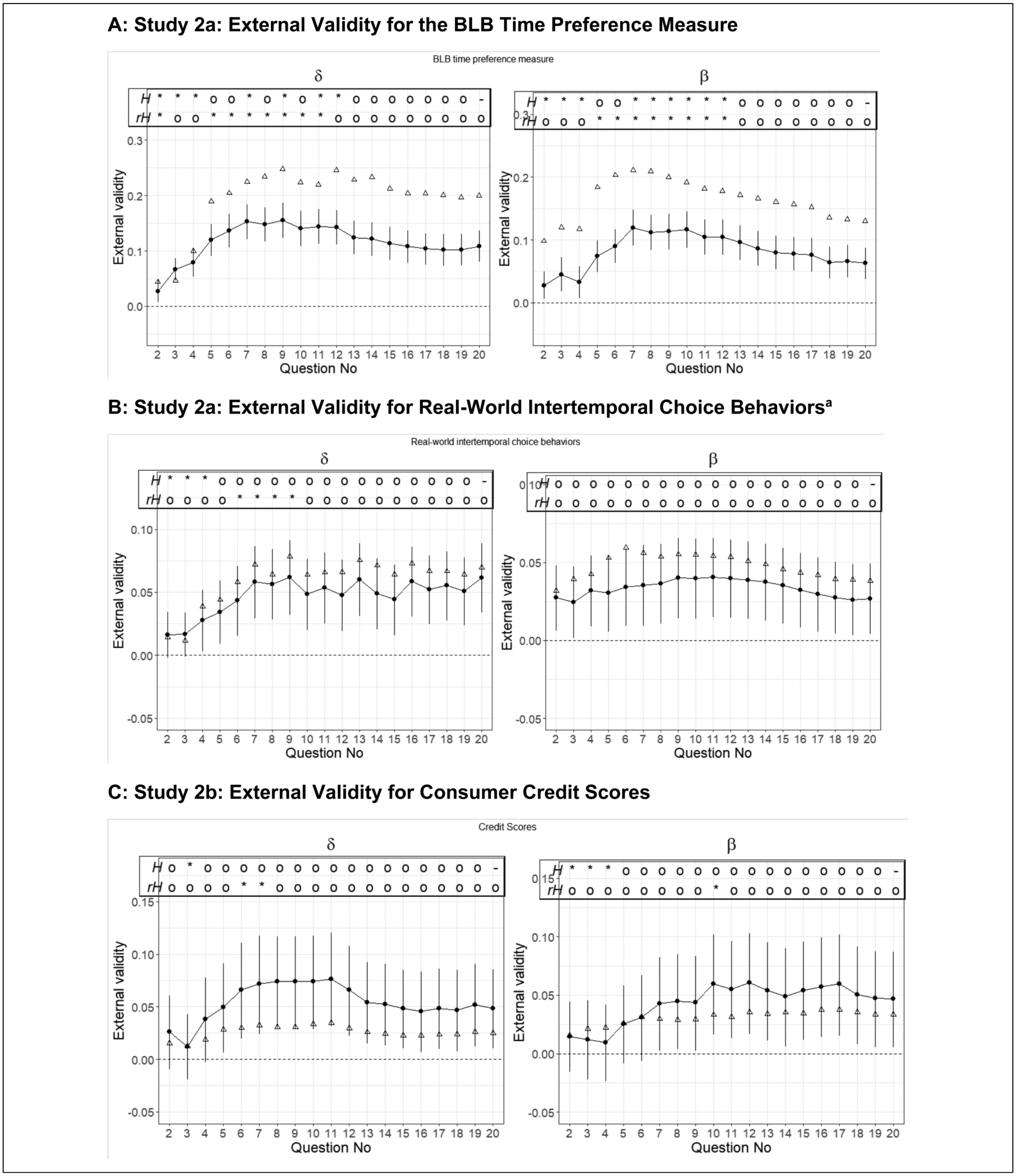
Panel A of Figure 3 shows how the external validity of the parameter estimates changed with the number of DEEP questions (see Web Appendix Table C1 for more details). For ease of interpretation, we plotted external validity as one minus the mean absolute percentage error (1 − MAPE). To assess the significance of these differences, we used Helmert and reverse-Helmert contrast tests. These contrasts (whose significance is depicted as asterisks and circles in Figure 3) compare the external validity at each question with the average external validity for all subsequent (Helmert) and all prior questions (reverse-Helmert), respectively (see Web Appendix Table C2).[4] For comparison, Figure 3 also shows the explained variance, as triangles, when regressing the external validity measure on the question-specific parameters (Equation 3).

We defined a peak to occur at question q if both the Helmert and reverse-Helmert contrast tests at that question are significant and the external validity is larger than at other questions. We defined a plateau to occur starting at question q if the reverse-Helmert contrast test is significant but the Helmert contrast and all subsequent Helmert contrasts are not. One way of thinking about this approach is as an iterative test of the incremental gain or loss in external validity as the number of questions increases.

The external validity of δ for predicting the BLB measure peaked at question 9 and the external validity of β peaked at question 7. That is, these results provide initial support for H_4 by finding a peak in external validity in which asking additional questions past question 9 actually reduced external validity.

Study 2a: External validity for real-world intertemporal choice behaviors. We performed a similar analysis to predict participants' self-reports of 36 real-world intertemporal choice behaviors with items such as smoking, flossing, and credit card debt (for a list of behaviors and their overall correlations with DEEP Time preference estimates, see Web Appendix C2). To make these behavioral measures comparable, we z-scored and oriented all 36 items such that higher numbers indicate more impatient behavior. We dropped 8 items that did not significantly correlate (at $p < .01$) with either DEEP Time preference estimates after any of the 20 questions. Because alpha for the remaining 26 items was .61, we averaged their z-scores into a behavior index.

[4] This analysis is preferable to pairwise tests because the entire set of coefficients is considered, which makes the test less sensitive to random differences in the coefficients. This process goes back to the matrices introduced by Friedrich Robert Helmert, which described the matrix used for contrasting estimated means across a series of observations.



A: Study 2a: External Validity for the BLB Time Preference Measure

B: Study 2a: External Validity for Real-World Intertemporal Choice Behaviors[a]

C: Study 2b: External Validity for Consumer Credit Scores

Figure 3. External validity for DEEP Time.
*$p < .05$.
o = $p \geq .05$.
[a]Composite index of 26 real-world intertemporal choice behaviors.
Notes: The points depict question-by-question external validity measures (1 − MAPE) and the error bars indicate 95% CIs around it. The symbols above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant for each question number. Triangles show the explained variance (R^2) of δ/β for the DV.

Using the approach outlined in Equations 3–5, we estimated linear models to predict the behavior index with the question-specific time preference estimates from DEEP and then predicted the arctan-transformed APEs from these models on question number with standard errors clustered by participant. Figure 3, Panel B, shows how the external validity of the DEEP estimates for predicting the behavior index changed with number of questions (see also Web Appendix Table C3). Using Helmert and reverse-Helmert contrast tests (see also Web Appendix Table C4), we found a plateau for the external validity of δ starting at question 6. For β estimates, none of the contrast tests were significant despite the appearance of a small peak at question 9.

Study 2b: External validity for credit scores. We followed the same procedures as used in Study 2a, using the hierarchical Bayes procedure to estimate 20 sets of time preference estimates for each participant, δ_{iq} and β_{iq}, for q from 1 to 20, and then using these estimates to evaluate external validity by estimating the question-by-question external validity of the task for predicting participants' credit scores (see Web Appendix Table C5). As we show in Figure 3, Panel C, the external validity of the DEEP δ estimates for predicting credit scores again appeared to peak between questions 7 and 11. However, only the reverse-Helmert contrast was significant at question 7, suggesting only a plateau (see Web Appendix Table C6). We also observed what appears to be a peak at question 10 for β, but only the significant reverse-Helmert contrast was significant, again suggesting a plateau.

Discussion

Studies 2a and 2b found that more questions not only may give diminishing returns for preference elicitation but also can even potentially reduce the external validity of the elicitation task. The external validity of DEEP—an efficient measure of time preference parameters—peaked as early as question 7 for predicting choices in another intertemporal choice task and plateaued after question 6 for predicting both self-reported behaviors and credit scores. This peak was most pronounced for the exponential discounting parameter, δ. The external validity of the present bias parameter, β, while suggesting a similar trend, was more modest in general and was more stable with more questions.

Although the measures of external validity were categorically different from each other—another intertemporal choice task, self-reported behavior, and credit scores—the external validity of DEEP's time preference estimates was reduced with more questions asked for all three measures. While the peak was not statistically significant for credit scores and real-world intertemporal choice behaviors, this lack of significance may have arisen because credit scores and self-reported behaviors are measures that reflect various other factors that are unrelated to time preferences and due to Study 2b's smaller sample size compared with Study 2a.

In summary, Study 2's results provide suggestive evidence that the decision processes measured in previous questions might be a better match for the decision processes used in the target behaviors that we are trying to predict compared with the task-specific decision processes respondents adapt to in later questions. These results must nonetheless be taken with a grain of salt as some features of the adaptive elicitation task may have contributed to the dynamics we observed. In particular, the questions' difficulties change along the task: questions are chosen with regard to each participant's preferences such that the options' attractiveness tend to become more and more similar with each additional question asked. Although these aspects of adaptive tasks cannot fully explain the effects observed in Study 2, they may have contributed to the reduction in external validity. Study 3 therefore uses a nonadaptive static elicitation task.

Study 3: Conjoint Choice for Consumer Preference Measurement

We next broaden the scope of our exploration by turning to conjoint analysis. Aside from studying a new preference measurement domain, Study 3 extends our previous studies in four important ways. First, rather than examining process changes (as in Study 1) and changes in the external validity of the measured preferences (as in Study 2) separately, Study 3 enables us to test both changes in a single study. Second, we used a nonadaptive conjoint choice task to measure preferences, unlike the adaptive task used in Study 2. Third, we used eye tracking as the process tracing method, which does not impose additional search costs, making it potentially more natural than mouse tracking, especially for more complex choices with many options and attributes. Finally, choices were incentive aligned, which addresses potential concerns about whether the Study 2 results may have been driven by unmotivated respondents.

Methods

We reanalyzed a data set that tracked the eye movements of 70 participants in a conjoint choice task (Yang, Toubia, and De Jong 2015). We employed the fixations determined in the original analysis of the data set with velocity-based fixation detection (Van der Lans, Wedel, and Pieters 2011).

In the measurement task, participants made 20 choices, each between four Dell computers that varied on six attributes with four levels each: processor speed (1.6 GHz, 1.9 GHz, 2.7 GHz, and 3.2 GHz), screen size (26 cm, 35.6 cm, 40 cm, and 43 cm), hard drive capacity (160 GB, 320 GB, 500 GB, and 750 GB), Dell support subscription (1 year, 2 years, 3 years, and 4 years), McAfee antivirus subscription (30 days, 1 year, 2 years, and 3 years), and price (€350, €500, €650, and €800). All participants saw the same prerandomized sequence of choice questions.

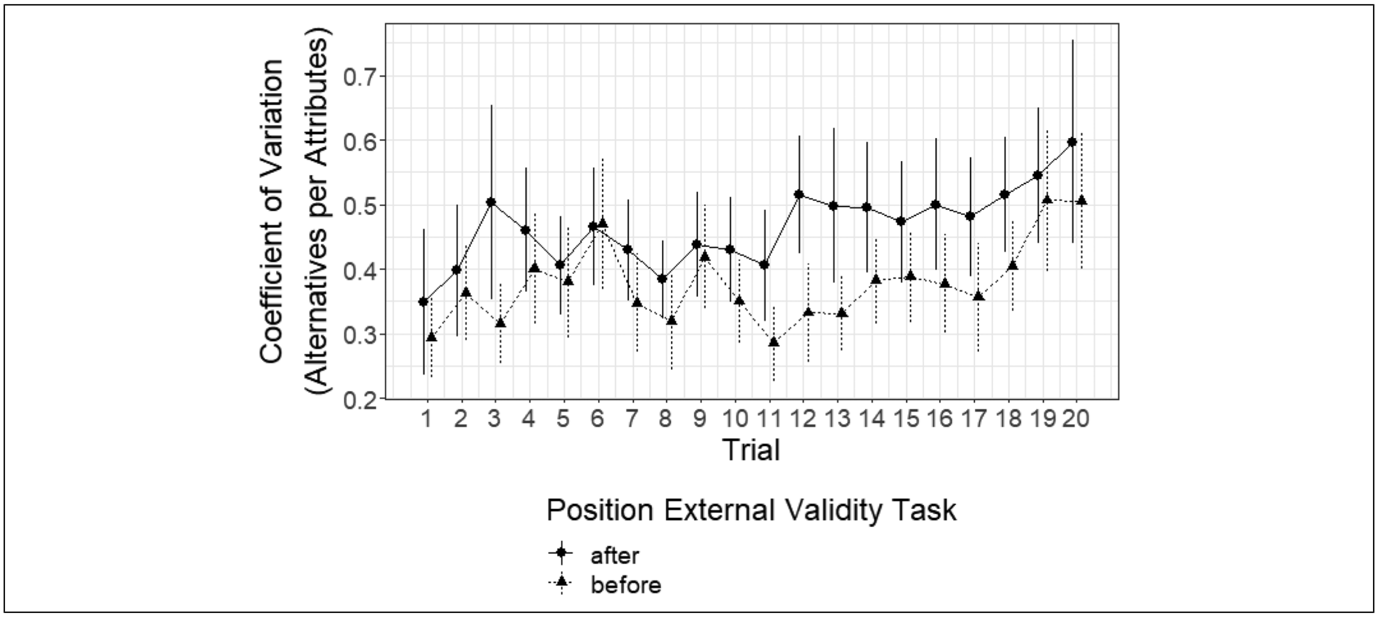


Figure 4. Average coefficient of variation of options viewed per attribute in each question as a function of the question number and the position of the external validity task (before or after).
Notes: The error bars represent the 95% CI around the means.

Participants also completed an external validity task consisting of a choice between six Dell computers (vs. four in the main task) that varied on the same attributes. The positions of the external validity task and measurement task were counterbalanced, with the external validity task being administered either before or after the measurement task. To make choices incentive aligned, one randomly drawn participant received a chosen laptop from either the external validity task (50% chance) or one of the measurement task's choices (each 2.5% chance), as well as the difference between €800 and the price of the chosen laptop.

Results

To test our hypotheses, we first examined changes in participants' decision processes by using the eye-tracking data (H_1). Because the choice task is more complex, we employed a related, but different, method from the one we used in Study 1. We then investigated whether elicited preferences change as search changes (H_3) by examining changes in estimated partworths. Finally, we assessed whether external validity peaks (H_4).

Decision processes. We first examined whether participants' search process changed as more questions were asked, reflecting changes in the decision-making process. Participants' search decreased from viewing an average of 77% of available information on the first question to 61% on the last question. We expected this reduction in search to correspond to participants focusing on selected attributes, as few as one, while only skimming the others (Jenke et al. 2021; Payne 1976; Russo and Rosen 1975). To assess whether this decrease in viewed information reflects changes in their decision-making process, while controlling for any decrease in total search, we calculated the coefficient of variation (CV; i.e., standard deviation divided by the mean) as a scale-independent summary of the variation in the number of options viewed across attributes.[5] Figure 4 plots the evolution of CV with more questions asked. To provide some benchmarks: The minimum CV of 0 corresponds to all attributes being searched equally (regardless of how many options are viewed). The maximum CV of 2.45 corresponds to comparing all four options on a single attribute while ignoring all other information.

We tested the development of CV across the 20 conjoint choices using similar methods as used in Study 1 by fitting regression models with standard errors clustered by participant and either a linear effect of question number or cubic regression splines to predict the CV for participant i on question q (1–20). These models also included a main effect of the position of the external validity task, position_i (before or after the measurement task) and the question × position

[5] The Payne Index used as a proxy for decision-making process in Study 1 is not sufficiently sensitive for a choice matrix consisting of four options with six attributes. For example, imagine a choice in which the respondent compares all options on only one attribute and subsequently scans the values of the remaining attributes of the chosen option (a common pattern in the data). This search pattern—four comparative and six integrative comparisons—generates a Payne Index that misleadingly indicates an integrative rather than comparative search.

interaction.

$$\mathrm{CV}_{iq} = \beta_0 + \beta_1 q + \beta_{2,\mathrm{position}_i} + \beta_{3,\mathrm{position}_i} q + \epsilon_{iq}, \tag{6}$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0,\ \sigma_\gamma^2)$, $\eta_{iq} \sim N(0, \sigma_\eta^2)$,

$$\mathrm{CV}_{iq} = \beta_0 + f_1(q) + \beta_{2,\mathrm{position}_i} + f_{2,\mathrm{position}_i}(q) + \epsilon_{iq}, \tag{7}$$

where $\epsilon_{iq} \sim N(0, \sigma^2)$.

Because the spline model did not improve model performance (see Web Appendix Table D3), we focus on the results of the linear model. As Table 4 shows, the linear model found a main effect of question number ($\beta_1 = .008$, $p = .003$), meaning that search increasingly focused on a few attributes with more questions.

Preferences. Next, to test whether these decision process changes were associated with preference changes (H_3), we estimated the utilities of each attribute level (i.e., partworths) after each question. Using a Bayesian hierarchical multinomial model to allow for parameter estimation with a small number of questions, we estimated individual- and population-level partworths independently after each question, starting with a minimum of two questions. We used the R package rstan (version 2.21.2; Stan Development Team 2020) to sample from the model's posterior distributions, following standard recommendations for setting the prior distributions for the individual- and group-level parameters (see code in Web Appendix D). Web Appendix Figure D1 plots the estimated population-level partworths as a function of the number of questions included in the estimation, revealing significant changes in the partworths with more questions asked. For example, while a larger screen size had a higher partworth than smaller screen sizes until question 8, this superiority vanished with more questions.

Although some of the changes in the attribute partworths may reflect greater uncertainty in the parameter estimation and thus higher variance when considering fewer questions, other changes in the partworths may be explained by changes in the decision process. To illustrate this development, Web Appendix Figure D2 plots each participant's standard deviation of relative attribute importance across attributes as a function of the number of questions considered for the partworth estimation.[6] This measure describes how much variance there is in attribute importance: lower variance means uniform attribute importance, whereas higher variance means some attributes are more important than others.

The results suggest that after a steep decrease in variance, suggestive of convergence in parameter estimates, the variance of relative attribute importance then increased linearly with more questions considered. To test this development, we again fit linear models with clustered standard errors and spline regression models to test the effect of question number. As suggested by Web Appendix Figure D2, the spline model provides a better description of the data accounting for the decrease between question 2 and 3 and subsequent increase in variance ($\mathrm{BIC}_{\mathrm{spline}} = -5{,}238$, $R^2 = 23\%$ vs. $\mathrm{BIC}_{\mathrm{lm}} = -5{,}054$, $R^2 = 4\%$). Despite the initial decrease, the linear model, summarized in the right-most columns of Table 4, estimates a significant positive effect of question number on the variance ($\beta_1 = .001$, $p < .001$). After question 2, the increase in variance was equally well described by the linear model, which we confirmed by fitting the models to this subset of questions only ($\mathrm{BIC}_{\mathrm{spline}} = -5{,}166$, $R^2 = 30\%$ vs. $\mathrm{BIC}_{\mathrm{lm}} = -5{,}170$, $R^2 = 29\%$).

Table 4. Effects of Question Number and Condition on Variance in Search in Study 3.

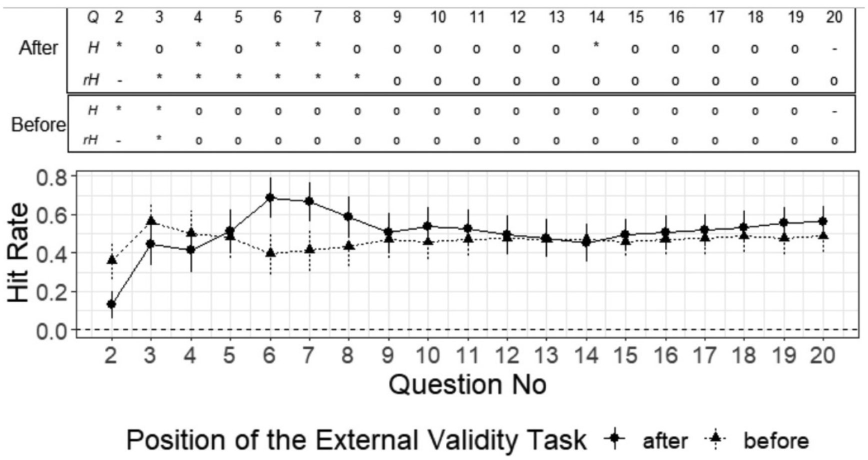
	DV			
	CV of the No. of Viewed Options per Attribute		Variance in Relative Attribute Importance	
Coefficients	**Est.**	***p***	**Est.**	***p***
β_0	.385	<.001	.122	<.001
β_1 (question)	.008	.003	.001	<.001
$\beta_{2,\mathrm{before}}$	−.060	.173	.005	.228
$\beta_{3,\mathrm{before}}$	−.003	.377	−.0004	.194

Notes: All models also contained participant-level random intercepts and slopes on question. Standard errors were clustered by participant.

Taken together with the process changes, these findings are consistent with the idea that participants increasingly compare options on selected attributes as the number of questions increases and that this leads to changes in preferences. Next, we examine whether these adapted preferences would be more or less useful for predicting the participants' choices on the external validity task.

External validity. Does the external validity of the preference estimates reach a peak (followed by decrease) with the number of questions asked (H_4)? Figure 5 plots, by condition, the evolution of average "hit rate" across participants. The individual hit rate is defined as each participant's predicted probability for their chosen option using the individual-level partworth estimates (i.e., the medians of the individual posterior distributions). When the external validity task came after the measurement task, the maximum average hit rate of 69% was reached after considering only the first six conjoint questions. When the external validity task came before, the maximum average hit rate of 56% was achieved after only three questions. These early peaks in external validity suggest that participants' adapted decision process did not match their decision process on the external validity task (which had a somewhat different format with six options and had 20 times higher likelihood of being chosen for incentive payments), and that the latter choice questions were not only unnecessary but actually hurt external validity.

[6] Relative attribute importance is defined as the difference in utility between the highest and lowest partworths for that attribute relative to the sum of those ranges for all attributes.



		Q	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
After		H	*	o	*	o	*	*	o	o	o	o	o	o	*	o	o	o	o	o	-
		rH	-	*	*	*	*	*	*	o	o	o	o	o	o	o	o	o	o	o	o
Before		H	*	*	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	-
		rH	-	*	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o	o

Figure 5. Average hit rate for predicting the external validity task.
$^{*}p < .05$.
o = $p \geq .05$.
Notes: Hit rate plotted as a function of the number of questions considered for the partworth estimation and the position of the external validity task. The symbols above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant.

We next describe and test this pattern with a similar model as in Study 2. We predicted the hit rate in the external validity task with fixed effects for question number, q (treated as a factor; i.e., we estimate a separate coefficient for each question), external validity task position, position_i, and their interaction (captured by a separate factor), and standard errors clustered by participant (Petersen 2009).

$$\mathrm{HR}_{iq} = \beta_0 + \beta_{1,q} + \beta_{2,\mathrm{position}_i} + \beta_{3,\mathrm{position}_i,q} + \epsilon_{iq}, \qquad (8)$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\eta_{iq} \sim N(0, \sigma_\eta^2)$.

We then calculated Helmert and reverse-Helmert contrasts on the estimated marginal mean hit rates predicted with this model. This analysis verified a statistical peak at question 6 when the external validity task was after the measurement task and a peak at question 3 when it was before (see Web Appendix Table D1 for contrast tests).

While these results corroborate the peaks found in Study 2, we can go one step further by explaining the peak and subsequent drop in external validity because we have both process and choice data for the same task. We can therefore test whether the decrease in hit rate as the number of questions increased was mediated by the observed changes in information search. For this analysis, we focused on questions 3 to 20 to reduce the amount of unexplained variance in the data that occurred due to the lack of convergence in question 2. To estimate the indirect effects of question number implemented as a factor (as in Equation 8), we first reparametrized the factor as dummy-coded binary variables Q_j for each question number (Hayes and Preacher 2014) to estimate the effects of each question number on hit rate, c_j, while maintaining the other parts of Equation 8.

$$\mathrm{HR}_{iq} = \beta_0 + \sum_{j=2}^{19} c_j \times Q_j + \beta_{1,\mathrm{position}_i} + \sum_{j=2}^{19} \beta_{2j,\mathrm{position}_i} \times Q_j + \epsilon_{iq}, \qquad (9)$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\eta_{iq} \sim N(0, \sigma_\eta^2)$.

For our mediator, we used the CV of visited alternatives per attribute. However, to account for the fact that the partworth estimates underlying the predicted hit rates are based on all previous questions, we calculated the average CV (mCV) of visited alternatives per attributes for the questions until q (as opposed to only the CV for question q, as we analyzed previously). To estimate the effects of question number on the mediator variable, a_j, we estimated the following model:

$$\mathrm{mCV}_{iq} = \beta_0 + \sum_{j=2}^{19} a_j Q_j + \beta_{1,\mathrm{position}_i} + \sum_{j=2}^{19} \beta_{2j,\mathrm{position}_i} Q_j + \epsilon_{iq}, \qquad (10)$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\eta_{iq} \sim N(0, \sigma_\eta^2)$.

Finally, we included fixed effects for the position of the external validity task as well as its interaction with mCV to estimate the effect of mCV (treated as a continuous variable) on the hit rate, b, which subsequently allowed us to estimate the indirect effects of q on the hit rate, a × b:

$$\mathrm{HR}_{iq} = \beta_0 + \sum_{j=2}^{19} c'_j \times Q_j + \beta_{1,\mathrm{position}_i} + \sum_{j=2}^{19} \beta_{2j,\ \mathrm{position}_i} Q_j + b \times \mathrm{mCV}_{iq} \quad + \beta_{3,\mathrm{position}_i} \times \mathrm{mCV}_{iq} + \epsilon_{iq}, \qquad (11)$$

where $\epsilon_{iq} = \gamma_i + \eta_{iq}$ and $\gamma_i \sim N(0, \sigma_\gamma^2)$, $\eta_{iq} \sim N(0, \sigma_\eta^2)$.

The analysis revealed that the evolution of hit rate over the questions was indeed mediated by changes in participants' decision processes (see Web Appendix Table D2). First, the mCV of visited alternatives per attributes was negatively related to hit rate ($b = -.477$, $p = .059$). We next tested the significance of the indirect effects (ab) by computing unstandardized indirect effects for each of 5,000 bootstrapped samples, deriving 95% CIs for the indirect effect from the 2.5th and 97.5th percentiles of the bootstrapped estimates. The indirect effects of question number on hit rate via mCV were significant for all but two questions, and the model explained $\Delta R^2 = 3\%$ more variance in hit rate than the model without the mediator. However, a significant question number effect (c) in the model without the mediator remained significant when including the mediators (c′), consistent with partial mediation (see Web Appendix Figure D3 for an illustration).

Discussion

Study 3 found that adaptation in a conjoint choice task occurred in a similar fashion as observed in Study 1. Eye-tracking data revealed that participants increasingly focused on comparing a few selected attributes (H_1), which changed the estimated partworths with more questions asked (H_3). Associated with that change in preferences was a peak and subsequent decrease in predictive accuracy for the external validity task (H_4), corroborating the results of Study 2. We further found that the decrease in external validity was mediated by the search process changes becoming more focused on comparing fewer attributes. Thus, incentive compatibility, nonadaptive choices, and reduced search costs (due to the eye-tracking vs. mouse-tracking technology) did not mitigate the effect of adaptation on respondents' decision process and preferences observed in Study 1 nor the peak in external validity observed in Study 2. Moreover, we replicated these results in a different domain with a significantly more complex task.

General Discussion

In four studies, we found that asking more questions is not always better for improving the external validity of a preference elicitation task. Instead, our studies revealed that respondents adapt to the task, increasingly relying on task-specific decision processes across repeated elicitation questions, which in turn reduces the task's external validity. Moreover, we found these effects in elicitation tasks of fairly typical lengths; longer tasks may exhibit exacerbated effects.

Our results illustrate that the standard "more is better" assumption for gathering data may not hold in preference elicitation tasks. While information theory suggests that more data should be better, it requires respondents' behavior in experimental tasks to be generated by the same process across questions (Fisher 1922). Instead, humans are adaptive decision makers (Payne, Bettman, and Johnson 1988), meaning they use task-specific processes that reflect their learning about the task structure and range of parameters. These task-specific processes, however, might deviate from the decision processes that produce the behavior we wish to predict.

Why does adaptation lead to increased mismatch with the behaviors we wish to predict? In Studies 1 and 3, we found that participants not only reduced how much they searched but also became more comparative and focused on fewer attributes with an increasing number of questions. As a consequence, the importance of attributes in later questions is different from earlier trials. These adaptations are task specific, as the effect of delay formats in Study 1 illustrates. Because elicitation tasks tend to have presentation formats and choice options that differ from the target behaviors we wish to predict (such as the external validity measures in Studies 2 and 3), adaptation will lead to decision processes that are likely to be less representative of the target behavior. For this reason, adaptive decision making in elicitation tasks may mean that collecting more data is not always more informative and can sometimes reduce our ability to predict behaviors outside the lab. After a certain point, we start to learn less about respondents' preferences and more about the strategies they used to get through a repetitive task.

Increasing Reliance on Strategies versus Additional Response Error

We initially anticipated that we may observe an increase in response error or noise with more questions asked. Our results instead suggested systematic changes, such that more questions led to less complete but more focused search patterns indicative of adaptations in decision processes (Jenke et al. 2021; Meißner, Musalem, and Huber 2016). Boredom and fatigue might nonetheless play a role here as a simplified search pattern is easier to generate than a random pattern, just as generating random choices is surprisingly hard for people to do (Rapoport and Budescu 1997).

Although the current results do not suggest that individuals produced more random responses—the consistency in the repeated choices in Study 1 did not change with more questions asked (see Web Appendix B2)—we cannot rule out increasing random responses for other task designs. While our studies asked at most 32 questions, other studies ask many more questions, in the hundreds (e.g., Amasino et al. 2019; Kable and Glimcher 2007; Kvam and Busemeyer 2020; Zhao et al. 2019) or even thousands (e.g., Konstantinidis et al. 2020; Nosofsky and Palmeri 1997). Indeed, studies in fields such as neuroscience often require *at least* 100 questions. At some point, respondents may adapt to an extremely simple strategy to quickly finish the task (i.e., straight-lining; Zhang and Conrad 2014). Given counterbalanced stimuli, similar response strategies could result in extremely noisy data with more questions. Studying the dynamics in tasks with so many questions is worth exploring in future research.

Implications

Research in marketing often aims to study preferences and behavior for predicting and understanding behaviors in the real world, such as consumer choices. Our results suggest some practices that can increase the validity of our measures while also saving time and money. First, some adaptive methods exist that provide better estimation with fewer questions. More research is needed, but it may be that asking as few as six questions can be sufficient to maximize external validity in some contexts (Cavagnaro et al. 2013; Toubia et al. 2013). Second, process-tracing techniques can be used to diagnose adaptation, helping identify when it is a threat to external validity. For instance, researchers could use process tracing to monitor changes in search as a proxy for changes in decision-making process; such changes could indicate potential mismatch between the decision processes used in the task and in the target behavior to predict, which may increase with further questions. This is particularly relevant if the cognitive models or neurological methods require a large number of data points per respondent for precise measurement. Optimizing the trade-off between potential bias introduced by adding questions versus the benefits of increased precision is an important question for future research.

If researchers' goal is to use an elicitation task (e.g., an elicitation task for measuring risk preferences) to predict real-world behavior (e.g., health behaviors), we suggest using the method we used in Studies 2 and 3 for identifying peaks. That is, designers of the elicitation task can use increasing subsets of the questions to estimate individual parameters and test for peaks in external validity by calculating Helmert and reverse-Helmert contrasts between question-specific predictions. If a similar dynamic is observed and peaks are identified, the number of questions included for estimation can be reduced ex post to maximize external validity.

In addition, we encourage the development of methods for mitigating adaptation to the task. For example, adaptation could be reduced or delayed by repeatedly changing the format of the task or adding filler questions or breaks. Incentives could be another way to reduce adaptation, although it is unclear whether a more motivated respondent would be more or less likely to adapt by using effort-saving strategies and high-power incentives that could even backfire (Ariely et al. 2009). Moreover, Yang, Toubia, and De Jong (2018) showed that incentive alignment in preference measurement is not sufficient to create a perfect match with real-world behavior. Generally speaking, our results suggest designing elicitation tasks that avoid the development of simplified strategies because such adaptations are unlikely to apply to more varied real-world decision contexts.

Our conceptual model (presented in Web Appendix A) can help researchers think about the optimal number of questions to ask while being aware of the trade-off between measurement precision and adaptation. Future studies could directly rely on the model's parameters to explore factors that should increase or decrease the likelihood of finding peaks in external validity and determine after how many questions that peak occurs. For example, studies could manipulate the efficiency of the measurement task to study whether more efficient tasks are indeed more likely to find peaks in external validity.

Finally, our research suggests that if the goal of preference measurement is to maximize external validity, researchers might use an ensemble of methods, preferably using multiple measurement modalities (e.g., intertemporal choices, matching questions) and a variety of contexts. For instance, data from preference elicitation tasks can be enriched by pairing them with market data and real consumer choices (Ellickson, Lovett, and Ranjan 2019; Feit, Beltramo, and Feinberg 2010; Swait and Andrews 2003). Multiple methods might allow researchers to identify which components of responses are associated with task-specific differences and which are associated with preferences, akin to identifying latent variables. Further, combining multiple methods may allow researchers to reduce the number of questions per task to mitigate the development of task-specific decision processes.

The focus of this article was on time preference measurement and conjoint analysis, which served as important complementary paradigms for testing our hypotheses. However, the results should extend to any choice or judgment task using similar, repeated decisions over a set of well-defined attributes. Further research should extend this question-by-question analysis of external validity to measurement of other preferences, such as risk aversion and contingent valuation.

Limitations

One potential concern with Studies 2a and 2b is that they both draw their conclusions from the DEEP Time task. In adaptive tasks such as DEEP Time, the difficulty of questions may increase with more questions answered. From a statistical perspective, increasing difficulty has the benefit of gathering more information from each elicitation question and can result in needing fewer questions to achieve precise estimates. However, increasing question difficulty could increase response error. This would suggest that our findings in Studies 2a and 2b may arise in part because later DEEP questions were harder for participants to answer precisely in line with their preferences. This decreased precision (or increased response error) in these harder questions might counter any additional information gained from those responses. This is an interesting question for further research, but the results of Studies 1 and 3, which both used static choice tasks, suggests our results do not depend on the use of adaptive methods.

Further, despite the limitations of the adaptive task, the very fact that the task is adaptive lets parameter estimation converge more quickly than in static elicitation tasks, which makes it easier to detect if respondents start off with some initial decision process before adapting to a task-specific decision process. In other words, if parameter estimates are not sufficiently converged before adaptation occurs, we may fail to detect adaptation. Conversely, not finding a peak in validity does not mean adaptation did not occur; it may just have occurred before sufficient convergence of parameter estimates was reached.

The current studies focused on adaptation in a behavioral task. Another question is whether the adaptation that happens in the tasks also happens in some of the targeted behaviors. Learning and adaptation are possible in real-world choices as well, and to the extent that people repeatedly encounter the same choices in life, they may start to adapt to them as well.

If the target behavior is also frequently repeated and features explicit trade-offs, then the consumer may well adapt to the choice structure. For example, perhaps a new consumer starts off carefully weighing the price versus organic trade-off when buying groceries but eventually forms a heuristic that as long as the organic version is less than 50% more expensive, they buy organic. It would be interesting to examine whether later elicitation questions are better at predicting repetitive behaviors, especially clearly structured ones, such as ordering at a favorite fast-food restaurant. We leave such empirical questions to future research in the field.

Conclusion

Humans are known to adapt to their environment, but most methods in behavioral research used to measure preferences have underappreciated this fact. Although cognitive models and neuroscientific methods have started to carefully characterize how preferences are accessed and/or assembled, measurement methods are still potentially clouded by the fact that researchers usually assume individual behavior in experimental tasks to be static, that is, free of sequential dependencies between questions. To make valid and reliable predictions for real-world behavior, we must see humans as the adaptive beings that they are, not the static decision makers we assume them to be.

Acknowledgments

The authors thank Elke Weber and audience members at ACR Berlin, BDRM Toronto, SPUDM Amsterdam, and SJDM Boston for helpful comments and Salah Chafik, Zeynep Enkavi, Jie Gao, and Brian Huh for research assistance.

Author Contributions

EJJ and OT contributed equally.

Associate Editor

Fred Feinberg

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: AKS gratefully acknowledges funding from the Swiss National Science Foundation (P400PS_186660). EJJ gratefully acknowledges funding from the Alfred P. Sloan Foundation (G-2018-11114) and from the National Science Foundation (UV-GA11247-155597).

ORCID iDs

Daniel G. Wall https://orcid.org/0000-0001-6897-3077
Eric J. Johnson https://orcid.org/0000-0001-7797-8347

References

Amasino, Dianna R., Nicolette J. Sullivan, Rachel E. Kranton, and Scott A. Huettel (2019), "Amount and Time Exert Independent Influences on Intertemporal Choice," *Nature Human Behaviour*, 3 (4), 383–92.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar (2009), "Large Stakes and Big Mistakes," *Review of Economic Studies*, 76 (2), 451–69.

Atlas, Stephen A., Eric J. Johnson, and John W. Payne (2017), "Time Preferences and Mortgage Choice," *Journal of Marketing Research*, 54 (3), 415–29.

Bartels, Daniel M., Ye Li, and Soaham Bharti (2021), "How Well Do Laboratory-Derived Estimates of Time Preference Predict Real-World Behavior? Comparisons to Four Benchmarks," working paper, University of Chicago.

Birnbaum, Michael H. (2013), "True-and-Error Models Violate Independence and Yet They Are Testable," *Judgment and Decision Making*, 8 (6), 717–37.

Broomell, Stephen B. and Sudeep Bhatia (2014), "Parameter Recovery for Decision Modeling Using Choice Data," *Decision*, 1 (4), 252–74.

Brucks, Merrie (1985), "The Effects of Product Class Knowledge on Information Search Behavior," *Journal of Consumer Research*, 12 (1), 1–16.

Cavagnaro, Daniel R., Richard Gonzalez, Jay I. Myung, and Mark A. Pitt (2013), "Optimal Decision Stimuli for Risky Choice Experiments: An Adaptive Approach," *Management Science*, 59 (2), 358–75.

Chabris, Christopher F., David Laibson, Carrie L. Morris, Jonathon P. Schuldt, and Dmitry Taubinsky (2008), "Individual Laboratory-Measured Discount Rates Predict Field Behavior," *Journal of Risk and Uncertainty*, 37 (2/3), 237–69.

Cohen, Jonathan, Keith Marzilli Ericson, David Laibson, and John Myles White (2020), "Measuring Time Preferences," *Journal of Economic Literature*, 58 (2), 299–347.

Costa-Gomes, Miguel, Vincent P. Crawford, and Bruno Broseta (2001), "Cognition and Behavior in Normal-Form Games: An Experimental Study," *Econometrica*, 69 (5), 1193–1235.

Coulter, Keith S. and Robin A. Coulter (2005), "Size Does Matter: The Effects of Magnitude Representation Congruency on Price Perceptions and Purchase Likelihood," *Journal of Consumer Psychology*, 15 (1), 64–76.

Dzyabura, Daria and John R. Hauser (2019), "Recommending Products When Consumers Learn Their Preference Weights," *Marketing Science*, 38 (3), 417–41.

Ellickson, Paul B., Mitchell J. Lovett, and Bhoomija Ranjan (2019), "Product Launches with New Attributes: A Hybrid Conjoint–Consumer Panel Technique for Estimating Demand," *Journal of Marketing Research*, 56 (5), 709–31.

Feit, Eleanor McDonnell, Mark A. Beltramo, and Fred M. Feinberg (2010), "Reality Check: Combining Choice Experiments with Market Data to Estimate the Importance of Product Attributes," *Management Science*, 56 (5), 785–800.

Fisher, Ronald A. (1922), "On the Mathematical Foundations of Theoretical Statistics," *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 222 (594–604), 309–68.

Frederick, Shane, George Loewenstein, and Ted O'Donoghue (2002), "Time Discounting and Time Preference: A Critical Review," *Journal of Economic Literature*, 40 (2), 351–401.

Freeman, A. Myrick, III, Joseph A. Herriges, and Catherine L. Kling (2014), *The Measurement of Environmental and Resource Values: Theory and Methods*. New York: Routledge.

Gigerenzer, Gerd and Wolfgang Gaissmaier (2011), "Heuristic Decision Making," *Annual Review of Psychology*, 62, 451–82.

Goldstein, Daniel G., Siddharth Suri, R. Preston McAfee, Matthew Ekstrand-Abueg, and Fernando Diaz (2014), "The Economic and Cognitive Costs of Annoying Display Advertisements," *Journal of Marketing Research*, 51 (6), 742–52.

Green, Paul E., Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive Conjoint Analysis: Some Caveats and Suggestions," *Journal of Marketing Research*, 28 (2), 215–22.

Green, Paul E. and Venkatachary Srinivasan (1978), "Conjoint Analysis in Consumer Research: Issues and Outlook," *Journal of Consumer Research*, 5 (2), 103–23.

Green, Paul E. and Venkat Srinivasan (1990), "Conjoint Analysis in Marketing: New Developments with Implications for Research and Practice," *Journal of Marketing*, 54 (4), 3–19.

Gustafsson, Anders, Andreas Herrmann, and Frank Huber (2013), *Conjoint Measurement: Methods and Applications*. Berlin: Springer Science & Business Media.

Hayes, Andrew F. and Kristopher J. Preacher (2014), "Statistical Mediation Analysis with a Multicategorical Independent Variable," *British Journal of Mathematical and Statistical Psychology*, 67 (3), 451–70.

Howell, John R., Peter Ebbes, and John C. Liechty (2021), "Gremlins in the Data: Identifying the Information Content of Research Subjects," *Journal of Marketing Research*, 58 (1), 74–94.

Inman, J. Jeffrey (2001), "The Role of Sensory-Specific Satiety in Attribute-Level Variety Seeking," *Journal of Consumer Research*, 28 (1), 105–20.

Jenke, Libby, Kirk Bansak, Jens Hainmueller, and Dominik Hangartner (2021), "Using Eye-Tracking to Understand Decision-Making in Conjoint Experiments," *Political Analysis*, 29 (1), 75–101.

Johnson, Eric J., Steven Bellman, and Gerald L. Lohse (2003), "Cognitive Lock-In and the Power Law of Practice," *Journal of Marketing*, 67 (2), 62–75.

Johnson, Eric J., Robert J. Meyer, and Sanjoy Ghose (1989), "When Choice Models Fail: Compensatory Models in Negatively Correlated Environments," *Journal of Marketing Research*, 26 (3), 255–70.

Johnson, Richard M. and Bryan K. Orme (1996), "How Many Questions Should You Ask in Choice-Based Conjoint Studies?" paper presented at ART Forum, Beaver Creek, CO, https://sawtoothsoftware.com/resources/technical-papers/how-many-questions-should-you-ask-in-choice-based-conjoint-studies.

Kable, Joseph W. and Paul W. Glimcher (2007), "The Neural Correlates of Subjective Value During Intertemporal Choice," *Nature Neuroscience*, 10 (12), 1625–33.

Konstantinidis, Emmanouil, Don van Ravenzwaaij, Şule Güney, and Ben R. Newell (2020), "Now for Sure or Later with a Risk? Modeling Risky Intertemporal Choice as Accumulated Preference," *Decision*, 7 (2), 91–120.

Kvam, Peter D. and Jerome R. Busemeyer (2020), "A Distributional and Dynamic Theory of Pricing and Preference," *Psychological Review*, 127 (6), 1053–78.

Laibson, David (1997), "Golden Eggs and Hyperbolic Discounting," *Quarterly Journal of Economics*, 112 (2), 443–77.

Li, Ye, Jie Gao, A. Zeynep Enkavi, Lisa Zaval, Elke U. Weber, and Eric J. Johnson (2015), "Sound Credit Scores and Financial Decisions Despite Cognitive Aging," *Proceedings of the National Academy of Sciences*, 112 (1), 65–69.

Lohse, Gerald and Eric J. Johnson (1996), "A Comparison of Two Process Tracing Methods for Choice Tasks," *Organizational Behavior and Human Decision Processes*, 68 (1), 28–43.

Ly, Alexander, Maarten Marsman, Josine Verhagen, Raoul Grasman, and Eric-Jan Wagenmakers (2017), "A Tutorial on Fisher Information," *Journal of Mathematical Psychology*, 80 (October), 40–55.

Marzilli Ericson, Keith M., John M. White, David Laibson, and Jonathan D. Cohen (2015), "Money Earlier or Later? Simple Strategies Explain Intertemporal Choices Better Than Delay Discounting Does," *Psychological Science*, 26 (6), 826–33.

Meier, Stephan and Charles D. Sprenger (2012), "Time Discounting Predicts Creditworthiness," *Psychological Science*, 23 (1), 56–58.

Meißner, Martin, Andres Musalem, and Joel Huber (2016), "Eye Tracking Reveals Processes That Enable Conjoint Choices to Become Increasingly Efficient with Practice," *Journal of Marketing Research*, 53 (1), 1–17.

Netzer, Oded, Olivier Toubia, Eric T. Bradlow, Ely Dahan, Theodoros Evgeniou,, Fred M. Feinberg, et al. (2008), "Beyond Conjoint Analysis: Advances in Preference Measurement," *Marketing Letters*, 19 (3/4), 337–54.

Nosofsky, Robert M. and Thomas J. Palmeri (1997), "An Exemplar-Based Random Walk Model of Speeded Classification," *Psychological Review*, 104 (2), 266–300.

Pachur, Thorsten, Michael Schulte-Mecklenbeck, Ryan O. Murphy, and Ralph Hertwig (2018), "Prospect Theory Reflects Selective Allocation of Attention," *Journal of Experimental Psychology: General*, 147 (2), 147–69.

Payne, John W. (1976), "Task Complexity and Contingent Processing in Decision Making: An Information Search and Protocol Analysis," *Organizational Behavior and Human Performance*, 16 (2), 366–87.

Payne, John W., James R. Bettman, and Eric J. Johnson (1988), "Adaptive Strategy Selection in Decision Making," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 14 (3), 534–52.

Pearl, Judea and Elias Bareinboim (2014), "External Validity: From Do-Calculus to Transportability Across Populations," *Statistical Science*, 29 (4), 579–95.

Perkovic, Sonja, Nicola J. Bown, and Gulbanu Kaptan (2018), "Systematicity of Search Index: A New Measure for Exploring Information Search Patterns," *Journal of Behavioral Decision Making*, 31 (5), 673–85.

Petersen, Mitchell A. (2009), "Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches," *Review of Financial Studies*, 22 (1), 435–80.

Rapoport, Amnon and David V. Budescu (1997), "Randomization in Individual Choice Behavior," *Psychological Review*, 104 (3), 603–17.

Reeck, Crystal, Daniel Wall, and Eric J. Johnson (2017), "Search Predicts and Changes Patience in Intertemporal Choice," *Proceedings of the National Academy of Sciences*, 114 (45), 11890–95.

Reimers, Stian, Elizabeth A. Maylor, Neil Stewart, and Nick Chater (2009), "Associations Between a One-Shot Delay Discounting Measure and Age, Income, Education and Real-World Impulsive Behavior," *Personality and Individual Differences*, 47 (8), 973–78.

Reisen, Nils, Ulrich Hoffrage, and Fred W. Mast (2008), "Identifying Decision Strategies in a Consumer Choice Situation," *Judgment and Decision Making*, 3 (8), 641–58.

Russo, J. Edward and Larry D. Rosen (1975), "An Eye Fixation Analysis of Multialternative Choice," *Memory & Cognition*, 3 (3), 267–76.

Scheibehenne, Benjamin, Jörg Rieskamp, and Eric-Jan Wagenmakers (2013), "Testing Adaptive Toolbox Models: A Bayesian Hierarchical Approach," *Psychological Review*, 120 (1), 39–64.

Scholten, Marc and Daniel Read (2010), "The Psychology of Intertemporal Tradeoffs," *Psychological Review*, 117 (3), 925–44.

Scholten, Marc, Daniel Read, and Adam Sanborn (2014), "Weighing Outcomes by Time or Against Time? Evaluation Rules in Intertemporal Choice," *Cognitive Science*, 38 (3), 399–438.

Schulte-Mecklenbeck, Michael, Joseph G. Johnson, Ulf Böckenholt, Daniel G. Goldstein, J. Edward Russo,, Nicolette J. Sullivan, et al. (2017), "Process-Tracing Methods in Decision Making: On Growing Up in the 70s," *Current Directions in Psychological Science*, 26 (5), 442–50.

Schulte-Mecklenbeck, Michael, Matthias Sohn, Emanuel de Bellis, Nathalie Martin, and Ralph Hertwig (2013), "A Lack of Appetite for Information and Computation. Simple Heuristics in Food Choice," *Appetite*, 71, 242–51.

Shah, Anuj K. and Daniel M. Oppenheimer (2008), "Heuristics Made Easy: An Effort-Reduction Framework," *Psychological Bulletin*, 134 (2), 207–22.

Shannon, Claude E. (1948), "A Mathematical Theory of Communication," *Bell System Technical Journal*, 27 (3), 379–423.

Stan Development Team (2020), *Stan Modeling Language: User's Guide and Reference Manual*. Version 2.27.0.

Stillman, Paul E., Xi Shen, and Melissa J. Ferguson (2018), "How Mouse-Tracking Can Advance Social Cognitive Theory," *Trends in Cognitive Sciences*, 22 (6), 531–43.

Story, Giles, Ivo Vlaev, Ben Seymour, Ara Darzi, and Ray Dolan (2014), "Does Temporal Discounting Explain Unhealthy Behavior? A Systematic Review and Reinforcement Learning Perspective," *Frontiers in Behavioral Neuroscience*, 8, 1–20.

Swait, Joffre and Rick L. Andrews (2003), "Enriching Scanner Panel Models with Choice Experiments," *Marketing Science*, 22 (4), 442–60.

Toubia, Olivier, Martijn G. De Jong, Daniel Stieger, and Johann Füller (2012), "Measuring Consumer Preferences Using Conjoint Poker," *Marketing Science*, 31 (1), 138–56.

Toubia, Olivier, Eric Johnson, Theodoros Evgeniou, and Philippe Delquié (2013), "Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters," *Management Science*, 59 (3), 613–40.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast Polyhedral Adaptive Conjoint Estimation," *Marketing Science*, 22 (3), 273–303.

Van der Lans, Ralf, Michel Wedel, and Rik Pieters (2011), "Defining Eye-Fixation Sequences Across Individuals and Tasks: The Binocular-Individual Threshold (BIT) Algorithm," *Behavior Research Methods*, 43 (1), 239–57.

Willemsen, Martijn C. and Eric J. Johnson (2010), "Visiting the Decision Factory: Observing Cognition with MouselabWEB and Other Information Acquisition Methods," in *A Handbook of Process Tracing Methods for Decision Research*, M. Schulte-Mecklenbeck, A. Kühberger, and R. Ranyard, eds. New York: Taylor & Francis, 21–42.

Yang, Liu, Olivier Toubia, and Martijn G. de Jong (2015), "A Bounded Rationality Model of Information Search and Choice in Preference Measurement," *Journal of Marketing Research*, 52 (2), 166–83.

Yang, Liu, Olivier Toubia, and Martijn G. de Jong (2018), "Attention, Information Processing and Choice in Incentive-Aligned Choice Experiments," *Journal of Marketing Research*, 55 (6), 783–800.

Zhang, Chan and Frederick Conrad (2014), "Speeding in Web Surveys: The Tendency to Answer Very Fast and Its Association with Straightlining," *Survey Research Methods*, 8 (2), 127–35.

Zhao, Wenjia Joyce, Adele Diederich, Jennifer S. Trueblood, and Sudeep Bhatia (2019), "Automatic Biases in Intertemporal Choice," *Psychonomic Bulletin & Review*, 26 (2), 661–68.