Title

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

Permalink
https://escholarship.org/uc/item/8739c3x8

Journal
Journal of Marketing Research, 59(5)

ISSN
0022-2437

Authors
Li, Ye
Krefeld-Schwalb, Antonia
Wall, Daniel G
et al.

Publication Date
2022-10-01

DOI
10.1177/00222437211073581

Peer reviewed

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

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—i.e., 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.

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. But how many questions should we pose to respondents to maximize an elicitation task's *external validity*—i.e., 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 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. While it may be tempting to assume 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 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*—i.e., 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, in a conjoint choice task measuring new car preferences, information on car price, fuel economy, safety ratings, warranty, country of manufacture may be displayed side-by-side for easy comparison across a number of 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 or 5-star safety ratings). On the other hand, 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 only compare 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 paper, we examine consequences of the tradeoff between increasing measurement precision with more questions, on the one hand, and increasing mismatch in decision processes, on the other.

This tradeoff between precision and mismatch is relevant for at least two related applications, 1) designing elicitation tasks and 2) testing theories. Marketers, psychologists, economists, and policymakers 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 with 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 (R. 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 paper, we aim to understand the tradeoff 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 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 2001)—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 where 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. Since 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 towards values reflecting task-specific adapted decision processes that are potentially *mismatched* with the decision-making processes driving the behaviors we wish to predict.[i] (See Web Appendix A for a stylized conceptual model that formalizes this discussion.)

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 only comparing amounts instead.

We thus hypothesize:

H1: Respondents will adapt their decision processes as they answer more elicitation questions.

H2: Adaptation will be task-specific, reflecting idiosyncrasies of the elicitation task format.

Since H1 and H2 are likely to be true in sufficiently long tasks (Meißner et al. 2016), we incorporated H1 and H2 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, we hypothesize:

H3: Adaptation in decision processes will change how respondents make choices and therefore impact the preferences estimated from those choices.

H4: 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, 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 et al. 2016; Yang et al. 2015; 2018). (3) Conjoint studies typically include a hold-out 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 in today) and a larger outcome available later (e.g., $60 in 1 month). Conjoint tasks, in contrast, typically employ complex displays of three or more choice options, each of which typically vary on up to 10 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 H1-H3 by collecting process data to demonstrate task-specific adaptation in an intertemporal choice task. Studies 2a and 2b test H4 by searching for peaks and subsequent decreases in the external validity of time preferences in two existing datasets. 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 H1, H3, and H4 in a conjoint choice task by examining how decision processes change and how these changes affect external validity. Table 1 gives and 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 (H1), that it is task-specific (H2), and that changes in decision processes relate to changes in preferences (H3). 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., Goldstein et al. 2014; Costa-Gomes, Crawford, and Broseta 2001; 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 upon the choice process and provides data that is 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 versus 48 hours). Because the tradeoffs 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 allows us to examine three main predictions. First, respondents' decision processes will adapt as they answer more questions (H1). 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 only on the most relevant attributes (Payne 1976; Perkovic, Bown, and Kaptan 2018; Reisen, Hoffrage, and Mast 2008) and recent work suggests 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 (H2). In particular, we expected participants to compare the delay information more when it was presented as hours than as days, since 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 (H3), depending on which attribute is increasingly compared.

Methods

We recruited 353 participants from Amazon Mechanical Turk (47.3% female, ages from 18 to 74, M_{age} = 34.9). Since 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 (see Web Appendix B1 for full task details). 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 which delays were presented. Delays were presented in either the day format or in the equivalent number of hours, although we used “now” in both formats since “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 (H1) and between task format conditions (H2). We start by investigating trends in global search patterns (comparative versus integrative search) and then focus on attribute-specific search changes in order to identify task-specific adaptation. Finally, we investigate whether preferences changed accordingly (H3).

Decision process dynamics across questions (H1) and formats (H2). We first examined if 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 et al. 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. While suggestive of adaptation, the decrease 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 (H1) and between conditions (H2), 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:

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

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

As a robustness check, we also estimated a generalized additive model with cubic regression splines, which flexibly accounts for potential nonlinearity in question number effects. This model did not include clustered standard errors.

$$\text{Payne Index}_{\text{iq}} = \beta_0 + f(q) + \beta_{2,\text{format}_{\text{iq}}} + f_{1,\text{format}_{\text{iq}}}(q) \times \text{Ses1}_q + f_{2,\text{cond}_{\text{iq}}}(q) \times \text{Ses2}_q + \epsilon_{\text{iq}} \quad \text{where } \epsilon_{\text{iq}} \sim N(0, \sigma^2) \tag{2}$$

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 Payne Index for the analysis; results with untransformed DVs are similar. Since both models (1) and (2) led to similar conclusions, we present the results for model (1); adding regression splines did not improve model performance (BIC_{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 Payne Index decreased with question number (β_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 Payne Index, Δ_{PI}, on question number in each session and condition, averaged across the remaining predictors.

Column 1 of Table 3 shows that Payne Index significantly decreased in all four conditions and in both sessions, indicating more comparative search. The 95% confidence interval for the condition specific trends largely overlapped between conditions. That is, participants' information search became more comparative, consistent with H1, but was not sensitive to delay format overall, seemingly counter to H2's prediction of task-specific adaptation.

Attribute-specific transitions (H2). The fact that 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 (H1). But Payne Index treats all attribute transitions the same; it does not distinguish which attribute is being compared. Task-specific adaptation (H2) 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 of amount and delay transitions as dependent variables, both arcsine-transformed to approach normality. Results for these models are shown in columns 2 and 3 in Tables 1 and 2; results with untransformed DVs are similar. In line with H2, 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 will 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 ($\beta_{3,\text{hour}} = .002$, $p = .323$; $\Delta_{\text{Amt}[\text{format=day;Ses1=1}]} - \Delta_{\text{Amt}[\text{format=hour;Ses1=1}]} = .002$) 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 in session 1 ($\beta_{3,\text{day}} = .001$, $p = .741$, $\Delta_{\text{Amt}[\text{format=day;Ses1=1}]} - \Delta_{\text{Amt}[\text{format=hour;Ses1=1}]} = -.002$).

These results provide further evidence that adaptation is task-specific (H2); while search trended increasingly comparative overall, the delay format influenced *which* attribute was increasingly compared. These differences in search and presumably decision-making process should have consequences for the preferences observed in the task, a topic we explore next.

Change in preferences (H3). We now turn to the choices participants made to examine whether changes in search were associated with changes in preferences. Reeck and colleagues (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 choosing the bigger amount, and thus more patient choices, whereas comparing delays should lead to a preference for the shorter delay, and thus less patient choices. Based on 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 Eqs. 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 spline model fit the data similarly well (see Web Appendix Table B2). So, we only present results from 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 H3 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*).[ii] We then computed the correlation of the random slopes, S_{1i}, across the variables. As 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 H3, that changes in decision processes are associated with changes in preferences.

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 measured 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 (H1), while the patterns differed between delay formats (H2). In particular, a strategy in which respondents chose only based on comparing delays became more prominent in the hour format compared to 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 towards 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 to 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 H4 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, DEEP Time (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, an adaptive elicitation task offers an important advantage: they should provide more information about the underlying parameters for each question asked compared to static elicitation tasks. Adaptive tasks should provide, in theory, the best chances of parameter identification before any adaptation occurs.

Using two different datasets, we analyzed three different 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 dataset (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 1308 participants (41.4% female, ranging in age from 18 to 86, with a mean age of 40.9) included 603 recruited from Amazon Mechanical Turk and 705 recruited from a market research firm.[iii] We look at two sets of responses. The first consists of 12 static intertemporal choices designed by Bartels and colleagues 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 tradeoffs 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 afterwards.

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, 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 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 datasets, we used the hierarchical Bayes approach outlined by Toubia et al. (2013) to estimate the two parameters of the quasi-hyperbolic 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, etc., 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 quasi-hyperbolic 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.

$$BLB_i = b_0 + b_1\delta_{iq} + \epsilon_{iq} \tag{3}$$
$$\text{where } \epsilon_{iq} = \gamma_i + \eta_{iq} \text{ and } \gamma_i \sim N\left(0, \sigma_\gamma^2\right), \eta_{iq} \sim N\left(0, \sigma_\eta^2\right)$$

Taking each models' predictions, $\widehat{\mathrm{BLB}_{\mathrm{\imath q}}}$, 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.

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

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. In order to construct confidence intervals 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.

$$arctan\ (APE^{\delta}_{BLB_{iq}}) = b_0 + b_{1,q} + \epsilon_{iq} \tag{5}$$
$$\text{where } \mathrm{\epsilon_{iq}} = \mathrm{\gamma_i} + \mathrm{\eta_{iq}} \text{ and } \mathrm{\gamma_i} \sim \mathrm{N}\left(0, \sigma_\gamma^2\right), \mathrm{\eta_{iq}} \sim \mathrm{N}\left(0, \sigma_\eta^2\right)$$

The top panel 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).[iv] 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 H4, 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 (see Web Appendix C2 for a list of behaviors and their overall correlations with DEEP time preference estimates). 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. Since alpha for the remaining 26 items was .61, we averaged their *z*-scores into a behavior index.

Using the approach outlined in Eqs. 3 to 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. The middle panel of Figure 3 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 shown in the bottom panel of Figure 3, 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 may not only give diminishing returns for preference elicitation but 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 to Study 2a.

In sum, Study 2's results provide suggestive evidence that the decision processes measured in earlier questions might be a better match for the decision processes used in the target behaviors that we are trying to predict compared to 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 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 non-adaptive, static elicitation task.

STUDY 3: CONJOINT CHOICE FOR CONSUMER PREFERENCE MEASUREMENT

We now 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 allows us to test both changes in a single study. Second, we used a non-adaptive 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 dataset that tracked the eye movements of 70 participants in a conjoint choice task (Yang et al. 2015). We employed the fixations determined in the original analysis of the dataset 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 pre-randomized sequence of choice questions.

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 examine changes in participants' decision processes by using the eye tracking data (H1). Because the choice task is more complex, we employ a related, but different, method from the one we used in Study 1. We then investigate whether elicited preferences change as search changes (H3) by examining changes in estimated partworths. Finally, we assess whether external validity peaks (H4).

Decision processes. We first examined whether participants' search process changed as more questions were asked, reflecting changes in 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 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.[v] 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 interaction.

$$\begin{gathered}\mathrm{CV_{iq}} \; = \beta_0 + \beta_1\, \mathrm{q} + \beta_{2,\mathrm{position_i}} + \beta_{3,\mathrm{position_i}}\mathrm{q} + \epsilon_{\mathrm{iq}} \\ \text{where } \epsilon_{\mathrm{iq}} = \; \gamma_{\mathrm{i}} + \; \eta_{\mathrm{iq}} \text{ and } \gamma_{\mathrm{i}} \sim \mathrm{N}\left(0, \sigma_\gamma^2\right),\; \eta_{\mathrm{iq}} \sim \mathrm{N}\left(0, \sigma_\eta^2\right)\end{gathered} \tag{6}$$

$$\begin{gathered}\mathrm{CV_{iq}} \; = \beta_0 + \mathrm{f_1(q)} + \beta_{2,\mathrm{position_i}} + \mathrm{f}_{2,\mathrm{position_i}}(\mathrm{q}) + \epsilon_{\mathrm{iq}} \\ \text{where } \epsilon_{\mathrm{iq}} \sim \mathrm{N}(0, \sigma^2)\end{gathered} \tag{7}$$

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

Preferences. Next, to test whether these decision process changes were associated with preference changes (H3), we estimated the utilities of each attribute level (i.e., partworths) after each question. Using a Bayesian hierarchical multinomial model to allow 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.

While 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.[vi] 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 (BIC(spline) = -5238, R^2 = 23% vs. BIC(lm) = -5054, R^2 = 4%). Despite the initial decrease, the linear model, summarized in the right 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 (BIC(spline) = -5166, R^2 = 30% vs. BIC(lm) = -5170, R^2 = 29%).

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 (H4)? 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.

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 (Peterson 2009).

$$\mathrm{HR}_{\mathrm{iq}} = \beta_0 + \beta_{1,\mathrm{q}} + \beta_{2,\mathrm{position}_\mathrm{i}} + \beta_{3,\mathrm{position}_\mathrm{i},\mathrm{q}} + \epsilon_{\mathrm{iq}} \tag{8}$$
$$\text{where } \epsilon_{\mathrm{iq}} = \gamma_\mathrm{i} + \eta_{\mathrm{iq}} \text{ and } \gamma_\mathrm{i} \sim \mathrm{N}\left(0, \sigma_\gamma^2\right),\ \eta_{\mathrm{iq}} \sim \mathrm{N}\left(0, \sigma_\eta^2\right).$$

We then calculated Helmert and reversed 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 if 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. In order 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 Eq. 8.

$$\mathrm{HR}_{iq} = \beta_0 + \sum_{j=2}^{19} c_j \times Q_j + \beta_{1,\mathrm{position}_i} + \sum_{j=2}^{19} \beta_{2j,\mathrm{position}_i} \times Q_j + \epsilon_{iq} \qquad (9)$$
$$\text{where } \epsilon_{iq} = \gamma_i + \eta_{iq} \text{ and } \gamma_i \sim N(0, \sigma_\gamma^2),\ \eta_{iq} \sim N(0, \sigma_\eta^2).$$

For our mediator, we used the coefficient of variation (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 above). To estimate the effects of question number on the mediator variable, a_j, we estimated the following model:

$$\mathrm{mCV}_{iq} = \beta_0 + \sum_{j=2}^{19} a_j\, Q_j + \beta_{1,\mathrm{position}_i} + \sum_{j=2}^{19} \beta_{2j,\mathrm{position}_i}\, Q_j + \epsilon_{iq} \qquad (10)$$
$$\text{where } \epsilon_{iq} = \gamma_i + \eta_{iq} \text{ and } \gamma_i \sim N(0, \sigma_\gamma^2),\ \eta_{iq} \sim N(0, \sigma_\eta^2).$$

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:

$$\mathrm{HR}_{\mathrm{iq}} = \beta_0 + \sum_{j=2}^{19} \mathrm{c'}_{\mathrm{j}} \times \mathrm{Q}_{\mathrm{j}} + \beta_{1,\mathrm{position}_\mathrm{i}} + \sum_{j=2}^{19} \beta_{2\mathrm{j},\mathrm{position}_\mathrm{i}} \mathrm{Q}_{\mathrm{j}} + \mathrm{b} \times \mathrm{mCV}_{\mathrm{iq}} \tag{11}$$

$$+\beta_{3,\mathrm{position}_\mathrm{i}} \times \mathrm{mCV}_{\mathrm{iq}} + \epsilon_{\mathrm{iq}}$$

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

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 coefficient of variation (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% confidence intervals 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 question numbers and the model explained ΔR^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 (H1), which changed the estimated partworths with more questions asked (H3). Associated with that change in preferences was a peak and subsequent decrease in predictive accuracy for the external validity task (H4), 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, non-adaptive choices, and reduced search costs (due to the eye-tracking versus 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, these results were replicated 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 search, 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. Since 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 vs. 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 et al. 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 at studying 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 (Toubia et al. 2013; Cavagnaro et al. 2013). Second, process-tracing techniques can be used to diagnose adaptation, helping to 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 tradeoff 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 reversed-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.

Additionally, 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 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 since 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 can directly rely on the model's parameters to explore factors that should increase or decrease the likelihood of finding peaks in external validity and 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 and 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 paper 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 like 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 clearly 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 upon 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 tradeoffs, then the consumer may well adapt to the choice structure. For example, perhaps a new consumer starts off carefully weighing the price versus organic tradeoff when buying groceries but eventually forms a heuristic that as long as the organic version is less than 50% more expensive, she buys organic. One interesting prediction would be to see if 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

While humans are known to adapt to their environment, most methods in behavioral research used to measure preferences have underappreciated this fact. Although cognitive models and neuroscientific methods have started 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.

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), 383-392.

Atlas, Stephen A., Eric J. Johnson, and John W. Payne (2017), "Time Preferences and Mortgage Choice," *Journal of Marketing Research*, 54 (3), 415-429.

Ariely, Dan, Uri Gneezy, George Loewenstein, and Nina Mazar (2009), "Large stakes and big mistakes," *The Review of Economic Studies*, 76 (2), 451-469.

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," *University of Chicago Working Paper*.

Bettman, James R., Eric J. Johnson, and John W. Payne (1990), "A componential analysis of cognitive effort in choice," *Organizational Behavior and Human Decision Processes,* 45 (1), 111-139.

Bhatia, Sudeep (2014), "Sequential sampling and paradoxes of risky choice," *Psychonomic Bulletin & Review*, 21(5), 1095-1111.

Broomell, Stephen B., and Sudeep Bhatia (2014), "Parameter recovery for decision modeling using choice data," *Decision*, 1 (4), 252-274.

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), 358-375.

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), 237-269.

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

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.

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.

Dzyabura, Daria, and John R. Hauser. "Recommending products when consumers learn their preference weights," *Marketing Science* 38.3 (2019): 417-441.

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), 709-731.

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), 309-368.

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

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.

Gigerenzer, Gerd, and Wolfgang Gaissmaier (2011), "Heuristic decision making," *Annual Review of Psychology*, 62, 451-482.

Gigerenzer, Gerd, Peter M. Todd, and the ABC Research Group (1999), *Simple heuristics that make us smart*. New York: Oxford University Press, USA.

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), 742-752.

Green, Paul E., Abba M. Krieger, and Manoj K. Agarwal (1991), "Adaptive conjoint analysis: Some caveats and suggestions," *Journal of Marketing Research*, 28 (2), 215-222.

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

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.

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), 451–470.

Inman, J. Jeffrey (2001), "The role of sensory-specific satiety in attribute-level variety seeking," *Journal of Consumer Research* 28(1), 105-120.

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): 255-270.

Johnson, Richard M., and Bryan K. Orme (1996), "How many questions should you ask in choice-based conjoint studies?" In ART Forum, Beaver Creek, CO, 1-23.

Kable, Joseph W., and Paul W. Glimcher (2007), "The neural correlates of subjective value during intertemporal choice," *Nature Neuroscience*, 10 (12), 1625-1633.

Krefeld-Schwalb, Antonia, Chris Donkin, Ben R. Newell, and Benjamin Scheibehenne (2019), "Empirical comparison of the adjustable spanner and the adaptive toolbox models of choice," *Journal of Experimental Psychology: Learning, Memory, and Cognition,* 45 (7), 1151-1165.

Kvam, Peter D., and Jerome R. Busemeyer (2020), "A distributional and dynamic theory of pricing and preference," *Psychological Review,* 127(6), 1053-1078.

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.

Laibson, David (1997), "Golden eggs and hyperbolic discounting," *The Quarterly Journal of Economics*, 112 (2), 443-477.

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.

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.

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.

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), 826-833.

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.

Meyer, Robert, and Eric J. Johnson (1995), "Empirical generalizations in the modeling of consumer choice," *Marketing Science,* 14 (3 supplement), G180-G189.

Netzer, Oded, Olivier Toubia, Eric T. Bradlow, Ely Dahan, Theodoros Evgeniou, Fred M. Feinberg, Eleanor M. Feit, Sam K. Hui, Joseph Johnson, John C. Liechty, James B. Orlin, and Vithala R. Rao (2008), "Beyond conjoint analysis: Advances in preference measurement," *Marketing Letters,* 19 (3/4), 337-354.

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), 147-169.

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), 366-387.

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), 534-552.

Pearl, Judea, and Elias Bareinboim (2014), "External validity: From *do*-calculus to transportability across populations," *Statistical Science*, 29 (4), 579-595.

Rapoport, Amnon, and David V. Budescu (1997), "Randomization in individual choice behavior," *Psychological Review*, 104 (3), 603-617.

Read, Daniel, Shane Frederick, and Marc Scholten (2013), "DRIFT: An analysis of outcome framing in intertemporal choice," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39 (2), 573-588.

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), 11890-11895.

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), 973-978.

Russo, J. Edward, and Larry D. Rosen (1975), "An eye fixation analysis of multialternative choice," *Memory & Cognition* 3(3), 267-276.

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), 925-944.

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, 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, 242-251.

Schulte-Mecklenbeck, Michael, Joseph G. Johnson, Ulf Böckenholt, Daniel G. Goldstein, J. Edward Russo, Nicolette J. Sullivan, and Martijn C. Willemsen (2017), "Process-Tracing Methods in Decision Making: On Growing Up in the 70s," *Current Directions in Psychological Science*, 26(5), 442–450.

Shah, Anuj K., and Daniel M. Oppenheimer (2008), "Heuristics made easy: an effort-reduction framework," *Psychological Bulletin* 134(2), 207-222.

Shannon, Claude E. (1948), "A mathematical theory of communication," *The 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), 531–543.

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 (76), 1-20.

Swait, Joffre, and Rick L. Andrews (2003), "Enriching Scanner Panel Models with Choice Experiments," Marketing Science, 22(4), 442-460.

Toubia, Olivier, Martijn G. De Jong, Daniel Stieger, and Johann Füller (2012), "Measuring consumer preferences using conjoint poker," *Marketing Science*, 31 (1), 138-156.

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), 613-640.

Toubia, Olivier, Duncan I. Simester, John R. Hauser, and Ely Dahan (2003), "Fast polyhedral adaptive conjoint estimation," *Marketing Science*, 22 (3), 273-303.

Tversky, Amos, Shmuel Sattath, and Paul Slovic (1988), "Contingent weighting in judgment and choice," *Psychological Review*, 95 (3), 371-384.

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), 239-257.

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.

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), 166-183.

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), 127-135.

Zhao, Wenjia Joyce, Adele Diederich, Jennifer S. Trueblood, and Sudeep Bhatia (2019), "Automatic biases in intertemporal choice," *Psychonomic Bulletin & Review*, 26 (2), 661-668.

TABLES AND FIGURES

Table 1. STUDY DESIGNS AND HYPOTHESES TESTED

Study	Domain	Design	Manipulation	Process Measure	External Validity Measure	H1	H2	H3	H4
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	✓		✓	✓

Table 2. EFFECTS OF QUESTION NUMBER AND DELAY FORMAT ON SEARCH AND CHOICES IN STUDY 1

DV:	(1) Payne Index		(2) Prop. Amount Transitions		(3) Prop. Time Transitions		(4) Larger-later Choices	
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
$\beta_{2,\text{hour}}$	-.0277	.5082	.0342	.2354	.0171	.4663	-.6729	<.0001
$\beta_{3,\text{day}}$	-.0049	.1275	.0052	.0311	.0006	.7414	.0025	.7916
$\beta_{3,\text{hour}}$	-.0058	.0166	.0018	.3231	.0035	.0066	.0044	.5696
$\beta_{4,\text{day}-\text{day}}$	-.0027	.3367	.0041	.0632	.0005	.7352	.0052	.5312
$\beta_{4,\text{day}-\text{hour}}$	-.0049	.1063	.0053	.0215	.0015	.3405	.0325	.0027
$\beta_{4,\text{hour}-\text{day}}$	-.0050	.1333	.0055	.0186	.0021	.2373	.0027	.8198

Note. 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

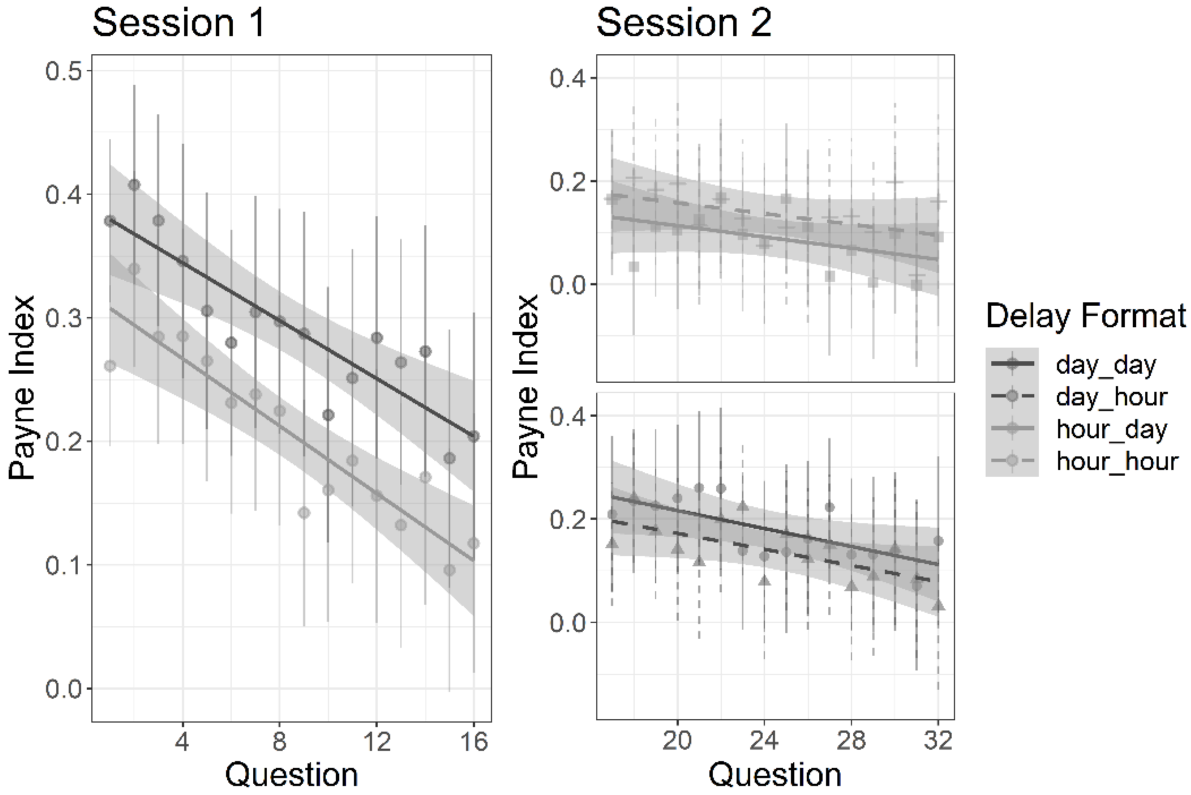
Session	Cond	(1) Payne Index Δ_{PI}		(2) Prop. Amount Transitions Δ_{Amt}		(3) Prop. Time Transitions Δ_{Time}		(4) Larger-later Choices Δ_{LL}	
		Est.	**95%CI**	**Est.**	**95%CI**	**Est.**	**95%CI**	**Est.**	**95%CI**
1	day	-.0074	[-.011, -.004]	.0034	[.001, .006]	.0020	[.000, .004]	-.0011	[-.003, .001]
	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]

Table 4. EFFECTS OF QUESTION NUMBER AND CONDITION ON VARIANCE IN SEARCH IN STUDY 3.

	Dependent Variable			
	Coefficient of variation of the number 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
$\beta_{2,\text{before}}$	-.060	.173	.005	.228
$\beta_{3,\text{before}}$	-.003	.377	-.0004	.194

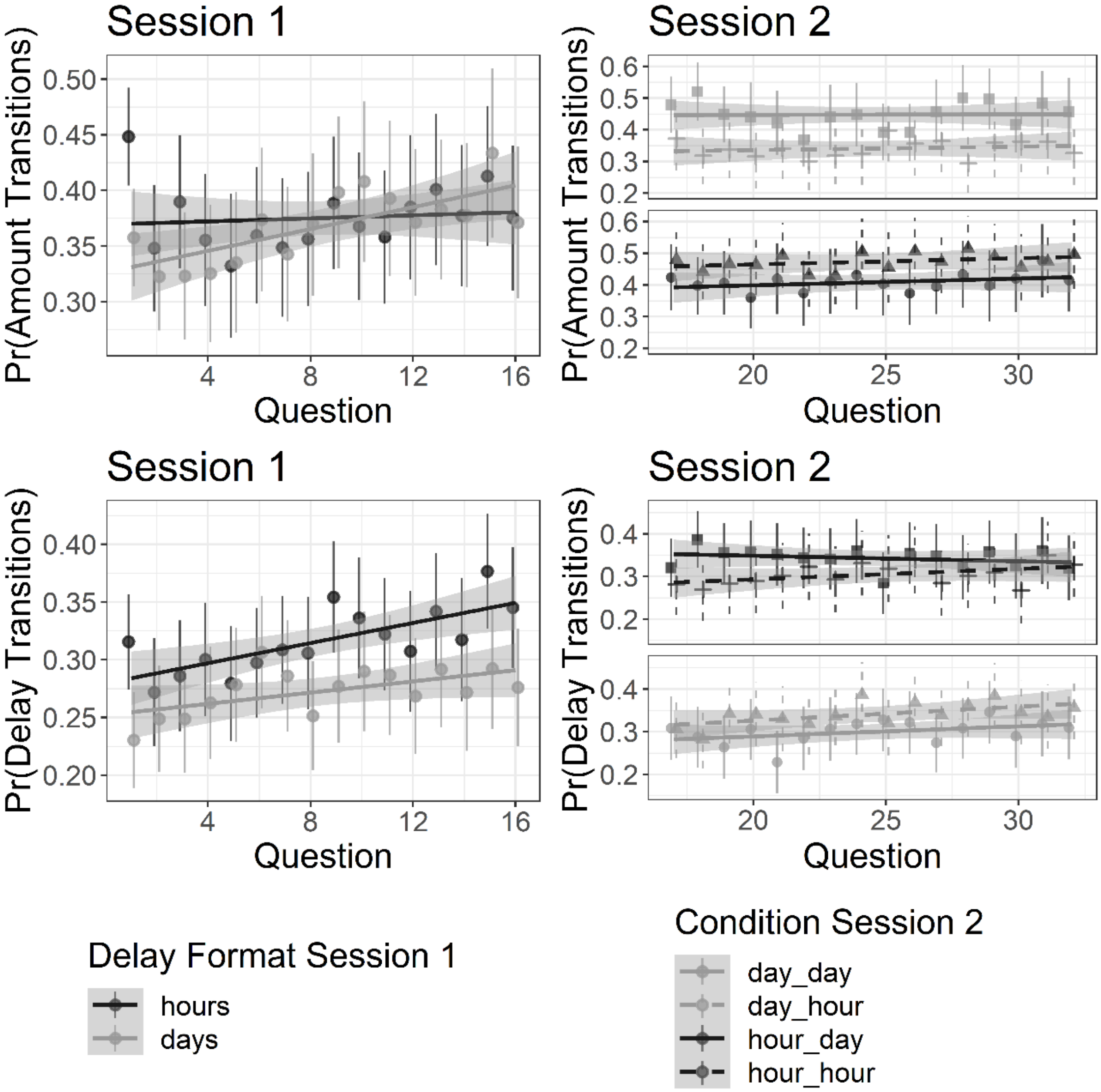
Note. All models also contained participant-level random intercepts and slopes on question. Standard errors were clustered by participant.

Figure 1. AVERAGE PAYNE INDEX AS A FUNCTION OF DELAY FORMAT.



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

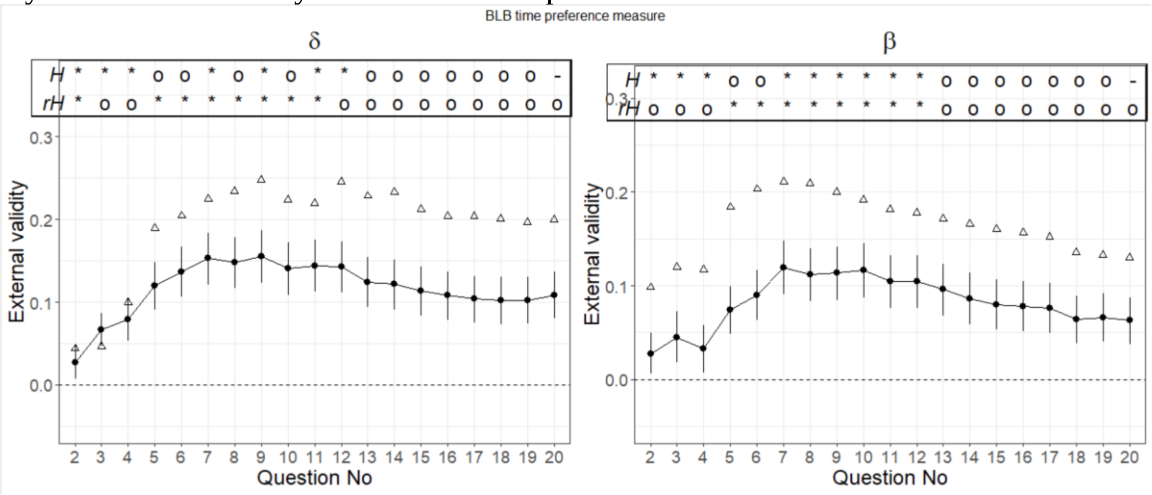
Figure 2. PROPORTION OF AMOUNT (UPPER ROW) AND DELAY (BOTTOM ROW) COMPARATIVE TRANSITIONS AS A FUNCTION OF THE DELAY FORMATS.



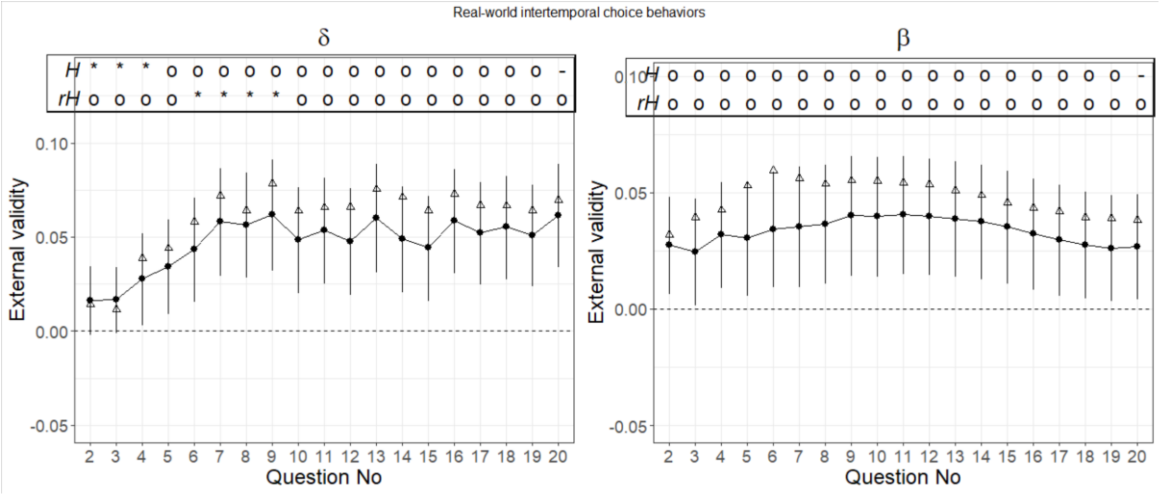
Note. The left plots illustrate proportions of each type of transition across questions 1-16 in session 1. The plots on the right do the same for questions 17-32 in session 2, split by the delay format in session 1. The error bars represent the 95% confidence interval around the mean and the lines illustrate the linear effect of question number, with the gray regions illustrating the confidence interval around the prediction.

Figure 3. EXTERNAL VALIDITY FOR DEEP TIME

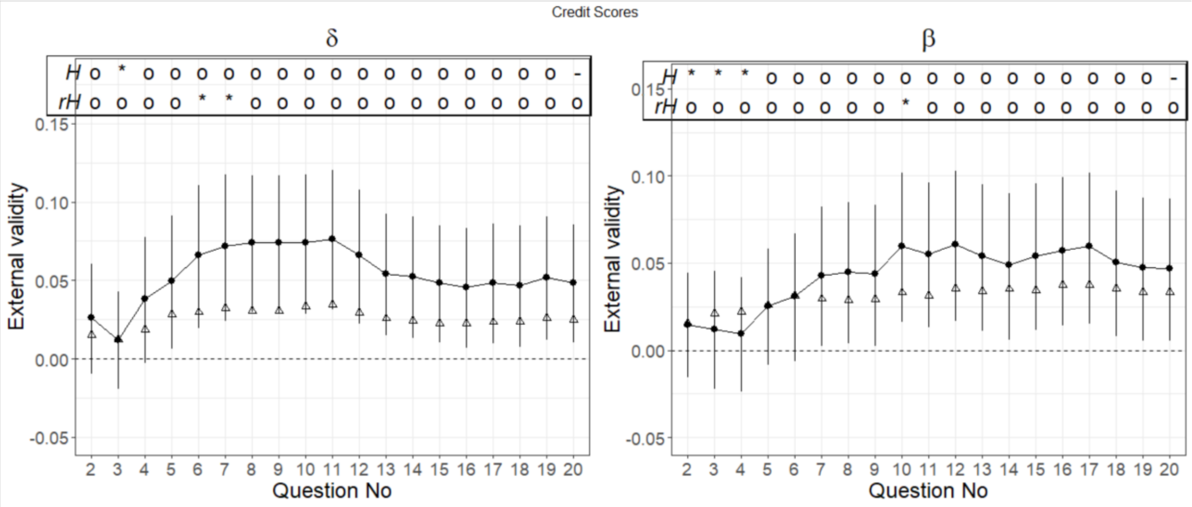
Study 2a: External validity for the BLB time preference measure.



Study 2a: External validity for real-world intertemporal choice behaviors.

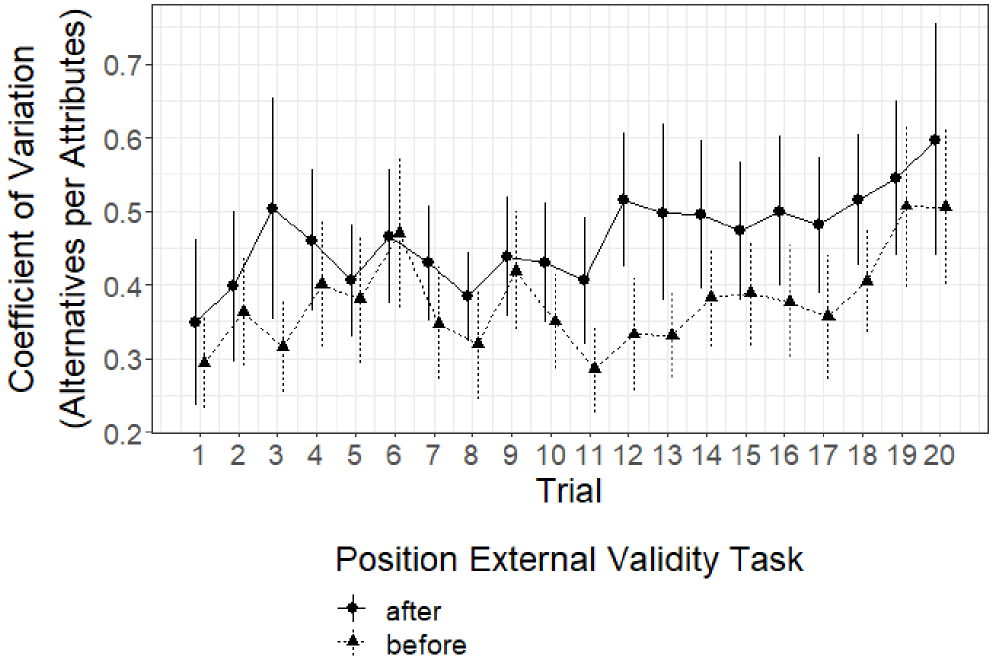


Study 2b: External validity for consumer credit scores.



Note. Top panel: Bartels, Li, and Bharti (2021) time preference measure. Middle panel: composite index of 26 real-world intertemporal choice behaviors. Bottom panel: credit score. The points depict question-by-question external validity measures (1 - mean absolute percentage error) and the error bars indicate 95% confidence intervals around it. The symbols above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant for each question number (*: $p < .05$; o: $p \geq .05$). Triangles show the explained variance (R^2) of δ/β for the DV.

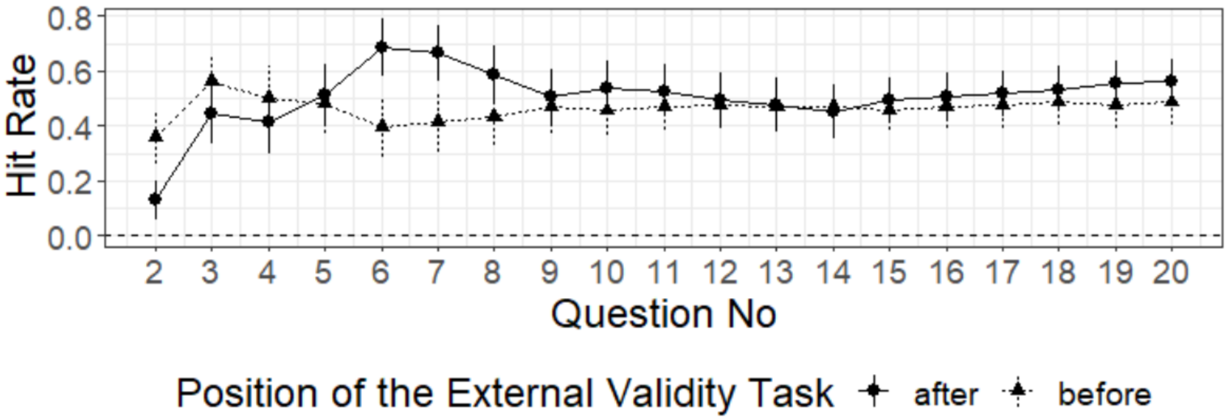
Figure 4. AVERAGE COEFFICIENT OF VARIATION OF OPTIONS VIEWED PER ATTRIBUTE IN EACH QUESTION AS A FUNCTION OF THE QUESTION NUMBER AND THE POSITION OF THE EXTERNAL VALIDITY TASK (BEFORE OR AFTER).



Note. The error bars represent the 95% CI around the means.

Figure 5. AVERAGE HIT RATE FOR PREDICTING THE EXTERNAL VALIDITY TASK.

	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



Note. 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 asterisks and circles above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant (*: *p* < .05; o: *p* ≥ .05) in the two conditions (after: dots and solid line; before: triangles and dashed line).

FOOTNOTES

[i] 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.

[ii] We used the following model to estimate the individual slopes for the proportion of amount transitions: $\mathrm{pr(Amount\ Transitions)}_{\mathrm{iq}} = \beta_0 + S_{0i} + (\beta_1 + S_{1i})q + \beta_{2,\mathrm{format}_{\mathrm{iq}}} + \beta_{3,\mathrm{format}_{\mathrm{iq}}} q \times \mathrm{Ses1}_{\mathrm{q}} + \beta_{4,\mathrm{cond}_{\mathrm{iq}}} q \times \mathrm{Ses2}_{\mathrm{q}} + \epsilon_{\mathrm{iq}}$ where $\epsilon_{\mathrm{iq}} \sim N(0, \sigma^2)$. The same model was used for delay transitions and larger-later choices with a logit link function. Since we are not testing the significance of the main effect coefficients nor 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.

[iii] Mechanical Turk participants were younger, more educated, and a higher percentage were male than participants from the market research firm. Since participant characteristics were not the focus of our paper, our analyses did not incorporate a panel variable or other demographics. As a robustness check, we also conducted the analysis separately on both subsamples. While we replicated the results in both subsamples for the BLB 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.

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

[v] 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 misleadingly indicating an integrative rather than comparative search.

[vi] 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.

WEB APPENDIX

The More You Ask, the Less You Get:

When Additional Questions Hurt External Validity

TABLE OF CONTENTS

WEB APPENDIX A: QUANTITATIVE MODEL OF ADAPTIVE CHOICE BEHAVIOR

We develop a stylized model to study how additional questions may affect the external validity of elicited preferences. In particular, we are interested in exploring conditions under which external validity may in fact decrease after some number of questions. The conditions we explore relate to the level of response error, the efficiency of questions, the amount of adaptation across respondents, and the speed with which adaptation occurs.

Formally, we denote the real-world choice we want to predict for respondent *i* as Y_i (e.g., a decision to pay off a credit card vs. carry a balance). We assume that Y_i reflects a true underlying preference, denoted by X_i (e.g., time preference), as well as idiosyncratic variations, reflected by a normally-distributed error term, ε_i. We acknowledge that preferences are constructed (Payne et al 1988), so we denote "true parameter" to simply mean the parameter that generates the target behavior. That is, we assume:

$$Y_i = X_i + \varepsilon_i, \text{ where } \varepsilon_i \sim N(0,\sigma_\varepsilon)$$

We consider a researcher who tries to predict Y_i based on the X_i estimated from respondents' answers to elicitation questions (e.g., a series of choices between smaller-sooner and larger-later amounts of money). We denote as $\hat{X}_{iq}$ the estimate of X_i for respondent *i* after question *q* that reflects the countervailing forces of precision and adaptation: Additional questions lead to convergence of the parameter estimate (i.e., precision), but the value to which the estimate converges may increasingly deviate from the true parameter X_i due to increasing reliance on task-specific decision processes (i.e., adaptation).

Although our model remains agnostic about the exact nature of adaptation, respondents will tend to rely on effort-reducing tactics that use a subset of the information or combine information using heuristics (Shah and Oppenheimer 2008) and therefore could produce responses that may systematically deviate from those produced by the true parameter X_i. If adaptation is increasing in *q*, asking additional questions will lead to divergence from the true parameter X_i, given that the different processes may produce different responses. We denote this deviation from the true parameter for respondent i as β_i. To represent respondents' increasing adaptation as they answer more questions, we weight β_{is} by a function $\sqrt{\alpha(q)}$ which is non-decreasing in the number of questions q. (We use the square root to simplify the expressions for external validity, which are functions of the square of the weight on the adapted process.)

With these assumptions, the combined effect of the true underlying value X_i and adaptation after q questions is: $X_i + \sqrt{\alpha(q)}\beta_i$. For convenience, we assume a specific functional form: $\alpha(q) = 1 - \exp\,(-r_\alpha q)$ with $r_\alpha > 0$. The parameter r_α captures how quickly $\alpha(q)$ increases: the larger r_α is, the faster $\alpha(q)$ converges to 1. For example, we may expect r_α to be smaller with preference measurement tools that make it harder for respondents to adapt to the task. While we implement this specification for convenience, our insights are not dependent on it, as long as $\alpha(q)$ is concave.

On the other hand, more questions reduce the standard error of the estimate, i.e., there is convergence within the session. Respondents likely report their preferences with error, leading to imperfect estimates. However, as more and more questions are asked, estimates converge to the underlying construct, $X_i + \sqrt{\alpha(q)}\beta_i$. We model this convergence using a normally distributed additional term, η_{iq}. To reflect the idea that estimates converge within a session, we model the variance of η_{iq} as decreasing in q; that is, we assume: $\eta_{iq} \sim N\left(0, \sigma_\eta\sqrt{\gamma(q)}\right)$, where $\gamma(q)$ is a non-increasing function of q. For convenience, we adopt a specific functional form: $\gamma(q) = \exp\,(-r_\gamma q)$. The parameter r_γ captures how quickly $\gamma(q)$ decreases: the larger r_γ is, the faster $\gamma(q)$ converges to 0. We can interpret r_γ as measuring the efficiency of the preference measurement method, i.e., how quickly the estimates converge. Again, while we adopt this specification for convenience, our insights are not dependent on it, as long as $\gamma(q)$ is convex.

In sum, we assume that our estimate of X_i for respondent *i* after question *q*, is as follows:

$$\hat{X}_{iq} = X_i + \sqrt{\alpha(q)}\beta_i + \eta_{iq}$$

Where $\alpha(q)$ is a non-decreasing function of *q*, $\eta_{iq} \sim N\left(0, \sigma_\eta\sqrt{\gamma(q)}\right)$, and $\gamma(q)$ is a non-increasing function of *q*. To finish our model specification, we assume X_i and β_i are normally distributed across respondents: $X_i \sim N(0, \sigma_X)$; $\beta_i \sim N(0, \sigma_\beta)$.

External validity is the correlation (across respondents) between the estimate $\hat{X}_{iq}$ and the behavior we wish to predict, Y_i.[1] We can show that this correlation after *q* questions is as follows:

$$Corr\left(\hat{X}_{iq}, Y_i\right) = \frac{Cov\left(\hat{X}_{iq}, Y_i\right)}{std(Y_i)std(\hat{X}_{iq})} = \frac{E(\hat{X}_{iq}Y_i)}{\sqrt{E(Y_i^2)E(\hat{X}_{iq}^2)}} = \frac{\sigma_X^2}{\sqrt{\sigma_X^2 + \sigma_\varepsilon^2}\sqrt{\sigma_X^2 + \alpha(q)\sigma_\beta^2 + \gamma(q)\sigma_\eta^2}}$$

The term $\alpha(q)\sigma_\beta^2$ captures deviation from the true parameter; it increases with *q*, leading to a decrease in external validity. The term $\gamma(q)\sigma_\eta^2$ captures convergence; it decreases with *q*, leading to an increase in external validity.

We now formally explore conditions under which the "deviation" force takes over the "convergence" force as the questionnaire progresses, i.e., conditions under which external validity may decrease after some point. We show the following propositions.

Proposition 1

$Corr\left(\hat{X}_{iq}, Y_i\right)$ is increasing for all q > 0 if one of the following holds:

(i) *σ_β is small enough* → adapted processes are not too mismatched with processes in the behavior we wish to predict

(ii) *σ_η is large enough* → there is a lot of response error

[1] Although we defined external validity broadly as the ability to use preferences measured in an elicitation task to make predictions about behaviors in other settings (Pearl and Bareinboim 2014), we use the narrower definition of correlation between preferences and some behavior to be predicted for convenience.

(iii) r_α *is large enough* → adaptation occurs quickly

Proposition 2

$Corr(\hat{X}_{iq}, Y_i)$ is decreasing for large q if one of the following holds:

(i) σ_β *is large enough* → adapted processes are more mismatched with processes in the behavior we wish to predict

(ii) σ_η *is small enough* → there is little response error

(iii) r_γ *is large enough* → estimates converge quickly

All conditions in Propositions 1 and 2 can be mapped onto the two opposing forces of convergence within the task and increasing adaptation. In particular, the incremental "convergence" benefit of additional questions is reduced if there is less response error (σ_η is small enough) or if estimates converge quickly (r_γ is large), leading the "deviation" force to take over. The incremental "deviation" effect diminishes more slowly if task-specific processes are more mismatched with the decision processes in the real-world behavior we wish to predict (i.e., σ_β is large), again leading the "deviation" force to take over. On the other hand, if adaptation occurs more quickly, the marginal impact of adaptation becomes smaller faster, and the "convergence" force dominates.

Next, we derive conditions under which external validity is inverted-U-shaped and derive a closed-form expression for the peak—the number of questions after which external validity starts decreasing. We show that:

Proposition 3

If $r_\gamma \sigma_\eta^2 > r_\alpha \sigma_\beta^2$ and $r_\alpha > r_\gamma$, then $Corr(\hat{X}_{isq}, Y_i)$ is increasing for all q > 0.

If $r_\gamma \sigma_\eta^2 > r_\alpha \sigma_\beta^2$ and $r_\gamma > r_\alpha$, then $Corr(\hat{X}_{isq}, Y_i)$ is inverted-U shaped, and it peaks at

$$q^* = \frac{\log\left(\frac{r_\gamma \sigma_\eta^2}{r_\alpha \sigma_\beta^2}\right)}{r_\gamma - r_\alpha}$$

In the region in which $Corr(\hat{X}_{iq}, Y_i)$ is inverted-U shaped, we can study how the peak, q^*, varies as a function of the model parameters. We find that:

Proposition 4

If $r_\gamma \sigma_\eta^2 > r_\alpha \sigma_\beta^2$ and $r_\gamma > r_\alpha$, then the peak in $Corr(\hat{X}_{iq}, Y_i)$, q^*, is:

(i) *Increasing in σ_η^2* → if responses are noisier, it takes more questions to hit the maximum correlation.

(ii) *Decreasing in σ_β^2* → if adapted processes are more mismatched from real-world processes, it takes fewer questions to hit the maximum correlation.

(iii) *Decreasing in r_γ* → if estimates converge quickly, it takes fewer questions to hit the maximum correlation.

The conditions in Proposition 4 are consistent with Propositions 1 and 2. In particular, higher response error helps the "convergence" force (i.e., increases the marginal benefit per question) and has no impact on the "deviation" force; more variation in task-specific processes increases the magnitude of the negative "deviation" force; and faster convergence of estimates reduces the marginal benefit of additional questions on "convergence," all making a decrease in external validity likely to occur with fewer questions.

Proof of Propositions 1 and 2

The derivative of $Corr(\hat{X}_{iq}, Y_i)$ with respect to q is of the same sign as: $-\alpha'(q)\sigma_\beta^2 - \gamma'(q)\sigma_\eta^2$. The first term is negative because $\alpha(q)$ is monotonically increasing and the second term is positive because $\gamma(q)$ is monotonically decreasing.

With our specification, $-\alpha'(q)\sigma_\beta^2 - \gamma'(q)\sigma_\eta^2 = -r_\alpha \exp(-r_\alpha q)\, \sigma_\beta^2 + r_\gamma \exp(-r_\gamma q)\, \sigma_\eta^2$

The first term is always negative, and the second term is always positive. We can therefore establish the above conditions under which external validity will decrease for high values of q. Note that we assume that q is bounded between 1 and some number Q, i.e., we do not let q go to 0 or ∞.

Proof of Proposition 3

We can establish some sufficient conditions that will ensure that $Corr(\hat{X}_{iq}, Y_i)$ is increasing at least for low values of q. This will be the case if $\alpha'(0)\sigma_\beta^2 + \gamma'(0)\sigma_\eta^2 < 0$. With our specification, this condition becomes: $r_\gamma \sigma_\eta^2 > r_\alpha \sigma_\beta^2$

Assuming this holds, next we can solve the first-order condition to find the optimal number of questions q^*: $r_\alpha \exp(-r_\alpha q^*)\, \sigma_\beta^2 = r_\gamma \exp(-r_\gamma q^*)\, \sigma_\eta^2$. This gives us:

$$q^* = \frac{\log\,(\frac{r_\gamma \sigma_\eta^2}{r_\alpha \sigma_\beta^2})}{r_\gamma - r_\alpha}$$

Note that $\log\left(\frac{r_\gamma \sigma_\eta^2}{r_\alpha \sigma_\beta^2}\right) > 0$ because $r_\gamma \sigma_\eta^2 > r_\alpha \sigma_\beta^2$. Therefore:

If $r_\gamma - r_\alpha < 0$, $q^* < 0$ and $Corr(\hat{X}_{iq}, Y_i)$ will be increasing for all $q > 0$.

If $r_\gamma - r_\alpha > 0$, $q^* > 0$ and $Corr(\hat{X}_{iq}, Y_i)$ will be inverted-U shaped.

Proof of Proposition 4

Proof: The first two conditions are obvious.

The sign of the derivative of $q*$ with respect to r_γ is the same as the sign of $1 - \frac{r_\alpha}{r_\gamma} + \log\,(\frac{r_\alpha \sigma_\beta^2}{r_\gamma \sigma_\eta^2})$. There are two cases.

Case 1: $\sigma_\beta^2 < \sigma_\eta^2$. Then we have:

$\log\left(\frac{r_\alpha \sigma_\beta^2}{r_\gamma \sigma_\eta^2}\right) < \frac{r_\alpha \sigma_\beta^2}{r_\gamma \sigma_\eta^2} - 1 < \frac{r_\alpha}{r_\gamma} - 1$ where the first inequality follows from log(1+x) < x for x > 0, and the second follows from $\sigma_\beta^2 < \sigma_\eta^2$. Therefore $1 - \frac{r_\alpha}{r_\gamma} + \log\left(\frac{r_\alpha \sigma_\beta^2}{r_\gamma \sigma_\eta^2}\right) < 0$.

Case 2: $\sigma_\beta^2 > \sigma_\eta^2$. Then the expression $1 - \frac{r_\alpha}{r_\gamma} + \log\,(\frac{r_\alpha \sigma_\beta^2}{r_\gamma \sigma_\eta^2})$ may be written as: $1 - x + \log\,(\alpha x)$ where $\alpha > 1$ and $x \epsilon [0, \frac{1}{\alpha}]$. This expression converges to -∞ as *x* approaches 0, and it is equal to 1 - $\frac{1}{\alpha}$, which is negative, when $x = \frac{1}{\alpha}$. The derivative of this expression is $-1 + \frac{1}{x}$, which is positive since x is always less than 1. Therefore, this expression is monotonically increasing in the $x \epsilon [0, \frac{1}{\alpha}]$ interval and it has a negative value when $x = \frac{1}{\alpha}$, which means it is always negative.

QED.

WEB APPENDIX B: STUDY 1 SUPPLEMENTAL MATERIALS

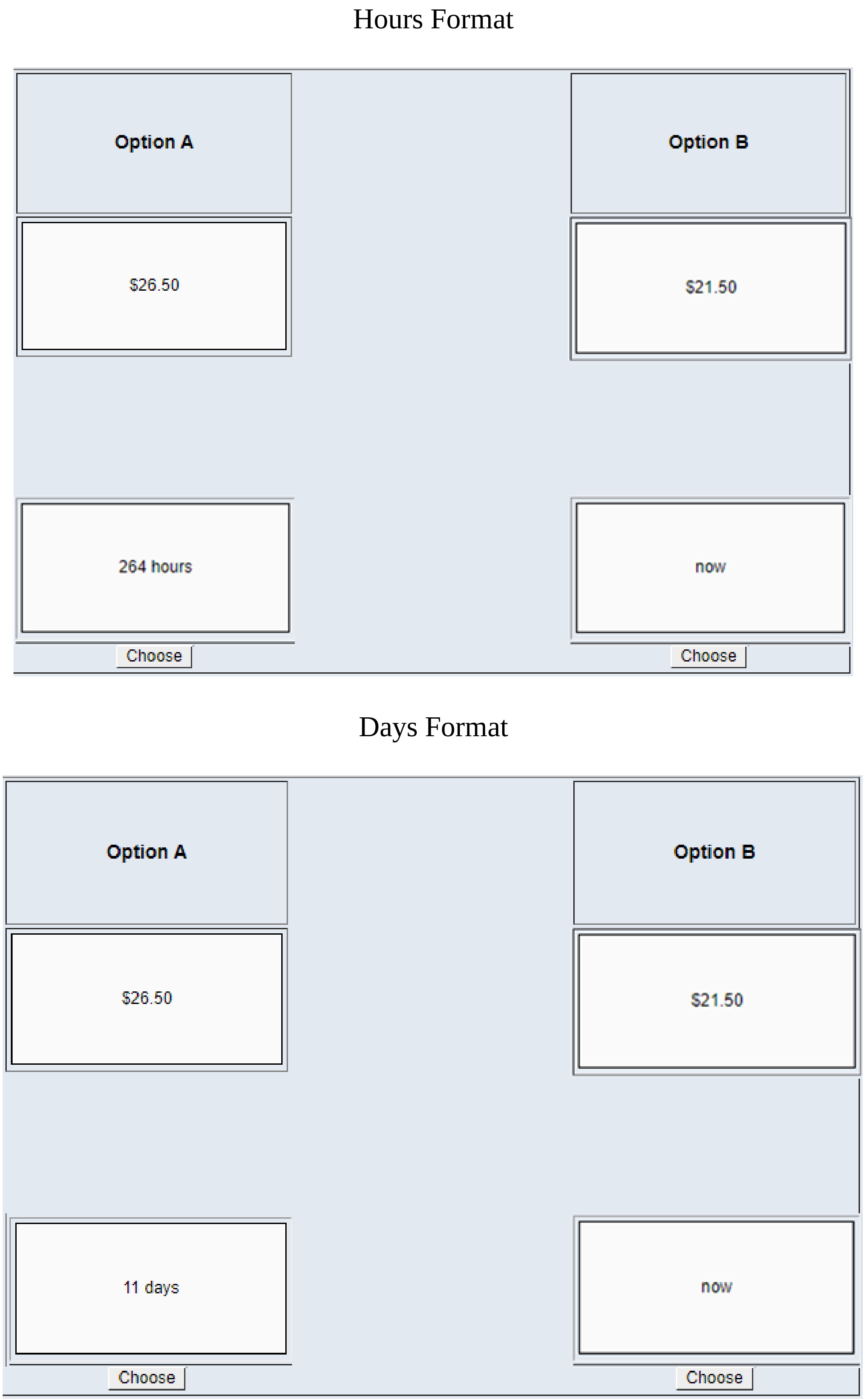
Web Appendix B1: Study 1 Task Details

Table B1. Study 1 elicitation task questions.

question	sooner amount	sooner time	larger amount	larger time
1	$21	now	$27	11 days
2	$21	1 day	$29	23 days
3	$21	3 days	$33	34 days
4	$21	7 days	$41	45 days
5	$22	now	$29	34 days
6	$22	1 day	$27	45 days
7	$22	3 days	$41	11 days
8	$22	7 days	$33	23 days
9	$24	now	$33	45 days
10	$24	1 day	$41	34 days
11	$24	3 days	$27	23 days
12	$24	7 days	$29	11 days
13	$26	now	$41	23 days
14	$26	1 day	$33	11 days
15	$26	3 days	$29	45 days
16	$26	7 days	$27	34 days

Note: For delays in hour format, we transformed the not-now times to days by multiplying by 24.

Figure B1. Examples of MouselabWeb display with all boxes opened


Hours Format

Days Format

Note: Options were randomized left-right but amounts always appeared on top and delays on bottom.

Web Appendix B2: Choice Consistency

As an additional robustness check for Study 1, we examined whether choices became less consistent with more questions in a way that suggests disengagement from the task (i.e., random responses; Howell et al 2021). To that end, we utilized the fact that participants answered essentially identical questions in the same order in Session 1 and Session 2. We therefore ran a generalized linear model with a logit link function to model choice consistency between pairs of identical questions as a function of question number (1-16), controlling for participant random effects and slopes:

$$p(cons)_{iq} = \frac{1}{1+e^{-(\beta_0+S_{0i}+(\beta_1+S_{1i})q+\beta_2 q_{iq}+\beta_{3,Cond_i}q+\beta_{4,Cond_i})}} \qquad (1)$$

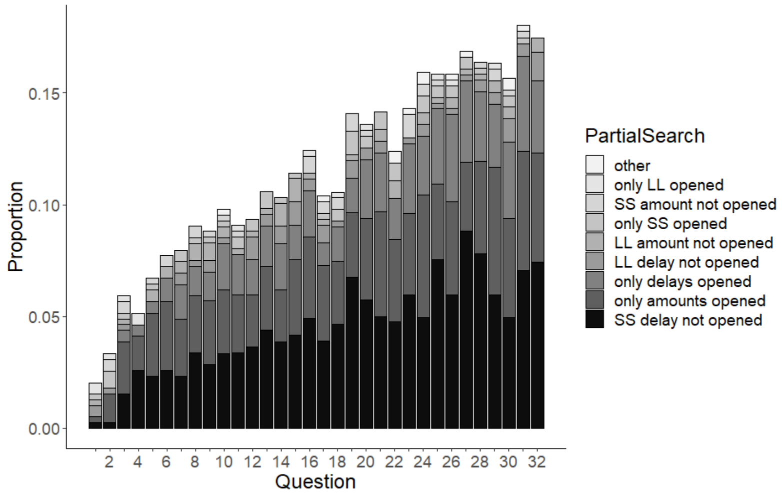
This analysis did not find a significant relationship between question number and choice consistency ($\beta_1 = 0.01$, p =.75), which suggests that choices do not become less consistent over time. That is, participants did not seem to become more disengaged with more questions.

Table B2. Model comparisons of the regression models with clustered errors, random intercepts and slopes, and splines for predicting the process and choice data in Study 1.

Dependent Variable					R^2
	Model	df	BIC	AIC	[Not adjusted]
Payne Index					
	Linear with Clust	9.000	15974	15909	.014
	Spline Regression	9.905	15998	15927	.013
Amount transitions					
	Linear with Clust	9.000	1157	1092	.013
	Spline Regression	10.048	1196	1124	.010
Delay transitions					
	Linear with Clust	9.000	-6038	-6103	.010
	Spline Regression	9.155	-6031	-6096	.009
LL choice					
	Linear with Clust	8.000	13087	13029	.019
	Spline Regression	14.893	13179	13072	.017

Web Appendix B3: The Toolbox Model

As alternative account to investigate adaptation to the elicitation task, we propose a Bayesian toolbox model (Scheibehenne, Rieskamp, and Wagenmakers 2013) to jointly fit the choice and search data in order to identify strategy use and measure systematic strategy shifts across questions. For example, a participant might compare the two amounts and choose the option with the larger amount in question one but might compare the delays and choose the option with the smaller delay in question two. As this example indicates, identifying the strategies participants used requires accounting for both information search and choices. Our toolbox model assumed that each decision can be described as selection among distinct decision strategies (i.e., “tools”) from a defined set (i.e., the “toolbox”) (Payne et al. 1988; Gigerenzer et al. 1999). The model estimated the probabilities that each of the strategies in the toolbox was applied for each question by considering both what information was searched and the option chosen. Applied across questions, the toolbox model could examine how each participant’s decision process adapted by identifying how the probabilities of each strategy changed with additional questions.



Partial Search	**Pr(LL)**
Full Search	.442
Only amounts	.985
Only delays	.012
SS delay not opened	.956

Figure B3.1. Proportion of partial search per question. The bars show the proportions of different types of partial search, by question. The table on the right summarizes the proportion of LL Choices (Pr(LL)) in questions where one of the three most frequent partial search patterns was observed.

Our first step was to identify common partial search patterns in Figure B3.1 (e.g., “only amounts”, “only delays” and “SS delay not opened”), and to classify a corresponding strategy (S1, S2, S3), as presented in Table B3.1. We also identified what option would be selected by that strategy. For example, S1, the “compare amounts” strategy, always chooses the option with the largest payoff whereas S2, the “compare delays” strategy, always chooses the shorter delay option. These strategies feature attribute-based comparisons that are a common feature of recent descriptive models of intertemporal choice (Ericson et al. 2015; Read, Frederick, and Scholten 2013; Scholten and Read 2010).

We also included two other strategies: Complete information search, which is required for any discounted utility model (S4) and no-information search, which implies guessing (S5). For S4, we used a standard exponential discounting model, setting the daily discount parameter to .017, the value which maximized the fit to the choices in the absence of process data. This is only one way of modeling temporal discounting and could be replaced by other models.

Table B3.1. *Strategies in toolbox and corresponding search behavior*

		Search behavior	
#	**Strategy**	**Acquisitions**	**Transitions**
S1	Compare amounts (“*Pick the bigger amount”)*	Amounts	Between amounts;
S2	Compare delays (“*Pick the sooner amount”)*	Delays	Between delays;
S3	Discount LL amount with LL delay and compare with SS amount	All but SS delay	Between amounts, within LL option;
S4	Discount amounts with delays and compare	All boxes	Between amounts, between delays within LL option, within SS option;
S5	Guess (i.e., random choice)	-	-

We expected participants’ use of simpler strategies that do not require integrating information (i.e., S1 and S2) would increase with question number. This could reflect participants’ adapting to the task as they learn the range of amounts and delays. Conversely, we expected participants' use of more complex strategies that require all four pieces of information (S4) would decrease with question number.

To examine adaptation over time, we implemented a novel computational toolbox model that extended prior models (Krefeld-Schwalb, Donkin, Newell, and Scheibehenne 2018; Scheibehenne, Rieskamp, and Wagenmakers 2013) by estimating strategy probabilities at the individual level over all questions. We can do this by considering choice and process data we extended prior work by estimating strategy probabilities for each individual at *each* question. Instead of identifying switching points in strategy use (Lee, Gluck, and Walsh 2019), we thus estimated the entire distribution of strategy probabilities across the task.

The probability to use a strategy $P(S_s)$ for respondent i in question q, was determined by means of a Luce choice rule (Luce 1952), such that the weight, ω, of one strategy, S, relative to the summed weights of all 5 strategies, determined that strategy's probability:

$$P_{qi}(S_s) = \frac{\omega_{siq}}{\sum_{j=1}^{5} \omega_{jiq}}$$

The weight of each strategy, ω_s, was modeled as a probit regression with an overall mean, μ_S, individual deviations from the mean λ_{Si}, individual specific question number effects τ_i, and eight regression weights for the search data: the number of times each of the four boxes was viewed and the number of transitions between amounts, between the delays, within the SS option, and within the LL option. We did not include a condition effect nor its interactions, since the focus of this analysis was estimating the process changes within each individual, whereas conditions were manipulated between individuals.

While the mean probability weight of each strategy μ and the regression weights for the process data b_1 to b_8 are group-level parameters, we estimate individual specific deviations from the mean λ and question number effects τ. For each of the individual level parameters, we used multivariate normal distributions as prior distributions, with normal hyperpriors for the group level means and set prior for the covariance matrix by using a scaled correlation matrix, with half-cauchy distribution as prior for a scaling factor and LKJ- distribution prior for the parameters' correlation matrix. The prior distributions of the parameters were determined in line with the recommendation in the manual for the software *stan* that we used for *mcmc sampling* from the models' posterior distributions (Stan development Team 2018).

We constrained the signs of the regression weights to correspond to search patterns assumed to occur for each strategy (see Table B3.1). For example, we expected S1 users to view amounts and transition between them, and so constrained the regression weights for the amount views and alternative-wise transitions to be positive for S1, while all remaining regression weights were constrained to be negative. The weight for strategy S1 on question *q* for participant *i* is therefore as follows, with all *b*'s constrained to be positive:

(1) $$\omega_{1qi} = \Phi(\mu_1 + \lambda_{1i} + \tau_{1i} \times q + b_{1\,1}LLAmt_{qi} + b_{1\,2}SSAmt_{qi} - b_{1\,3}LLDelay_{qi} - b_{1\,4}SSDelay_{qi} + b_{1\,5}Trans_{Amt_{qi}} - b_{1\,6}Trans_{Delay_{qi}} - b_{1\,7}Trans_{SS_{qi}} - b_{1\,8}Trans_{LL_{qi}}).$$

The model thus assumes that S1 is more likely with more amount acquisitions and transitions between amounts but less likely with more delay acquisitions and transitions between delays or within alternative. We make similar sets of assumptions for the remaining strategies.

The model's prediction $Pr_{iq}(LL)$ for respondent *i* in question *q*, is given by the summed predictions $P_q(LL|S_j)$ of the strategies weighted with the probability to use the strategies.

$$Pr_{iq}(LL) = \sum_{j=1}^{5} P_{iq}(S_j) \times P_q(LL|S_j)$$

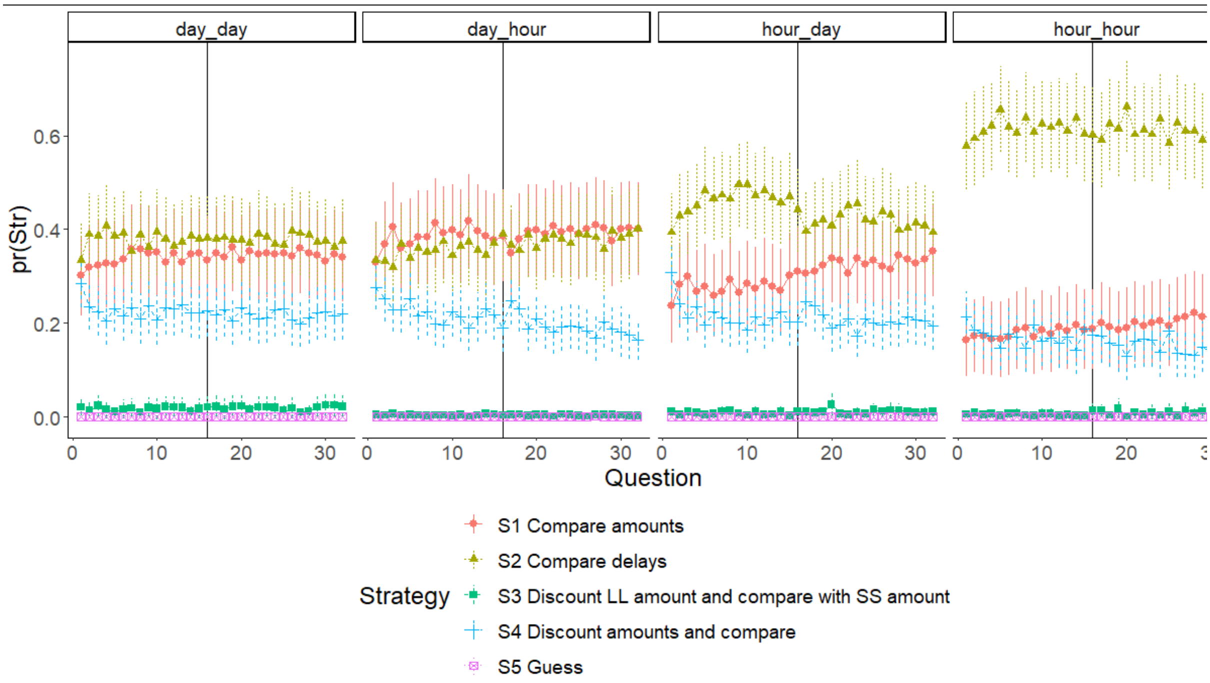


Figure B3.2. Estimated probabilities of the toolbox strategies by question and condition averaged across participants with the 95% CI around the mean. The vertical lines correspond to the break between session 1 and session 2. See Table B3.1 for details on strategies.

The toolbox model results were consistent with the inferences the behavioral and process data separately suggested—that participants adapted their decision processes to the task over time. Figure B3.2 plots the average estimated probability of each strategy by question and condition. We tested the development of each strategy's probabilities over time with GLM models, including question number, condition and their interaction as fixed effects and random slopes and intercepts per participants.

$$Prop(S)_{siq} = \beta_0 + S_{0i} + (\beta_1 + S_{1i})q + \beta_{2s,Cond_{iq}}q + \beta_{3Cond_{iq}} + \epsilon_{iq} \quad \text{where } \epsilon_{iq} \sim N(0,\sigma^2) \qquad (2)$$

The coefficients of the models' fixed effects are reported in Table B3.2. Participants adapted by using, on average, simpler strategies as they answered more questions, especially just comparing amounts (S1). S1 became significantly more likely with more question asked, no interaction or main effect of Cond was observed. We did not observe a main effect of trial number on S2, however S2 was, in line with the results reported in the main part of the manuscript more likely in the hour format conditions.

Table B3.2. Summary of fixed effects for the GLMs on the estimated probabilities of strategy use with the Toolbox model.

	Strategy									
	S1		**S2**		**S3**		**S4**		**S5**	
	Est.	***p***	**Est.**	***p***	**Est.**	***p***	**Est.**	***p***	**Est**	***p***
β_0	.209	<.001	.329	<.001	.008	0	.185	<.001	0	.575
β_1	.001	<.001	.000	.119	0	.322	-.001	<.001	0	.306
$\beta_{2,day-day}$	.004	.174	-.002	.678	0	.746	.001	.831	0	.65
$\beta_{2,day-hour}$	.002	.402	.002	.538	-.001	.528	-.003	.388	0	.587
$\beta_{2,hour-day}$	-.003	.178	.006	.094	0	.682	-.005	.095	0	.819
$\beta_{3,day-day}$	.000	.334	.000	.832	0	.643	0	.814	0	.326
$\beta_{3,day-hour}$	.000	.450	.000	.792	0	.788	0	.837	0	.357
$\beta_{3,hour-day}$	.000	.086	.000	.096	0	.977	0	.114	0	.466

In contrast, more complex strategies became less likely—especially integrating all information (S4). The analysis replicated the conclusions from the main text that the smaller proportion of larger-later choices when delays were displayed as hours, was at least in part because participants tended to only compare delays (S2) in this condition.

Additional References

Lee, Michael D., Kevin A. Gluck, and Matthew M. Walsh (2019), "Understanding the complexity of simple decisions: Modeling multiple behaviors and switching strategies," *Decision* 6(4), 335-368.

WEB APPENDIX C: STUDY 2 SUPPLEMENTAL MATERIALS

Web Appendix C1: Bartels, Li, and Bharti (2021) intertemporal choice task (BLB task)

Imagine that you can choose which of two sums of money you’d like to receive, one available sooner and the other available later.

For each choice below, please indicate which of these two payments you would prefer to receive. Imagine that each payment is guaranteed to occur when promised.

1. $816 in six months ——OR—— $860 in nine months
2. $213 today ——OR—— $281 in two years
3. $791 today ——OR—— $777 in one month
4. $457 today ——OR—— $551 in six months
5. $1064 today ——OR—— $1153 in one month
6. $600 today ——OR—— $611 in one month
7. $816 in six months ——OR—— $1028 in one year
8. $816 today ——OR—— $5440 in one year
9. $840 in six months ——OR—— $10,125 in two and a half years
10. $777 today ——OR—— $791 in one month
11. $816 today ——OR—— $860 in three months
12. $400 in six months ——OR—— 440 in one and a half years
13. $621 in six months ——OR—— $670 in six months
14. $504 today ——OR—— $524 in one month

Note: To assess time preferences, simply count the number of larger, later options chosen, excluding items 3 and 13, which feature dominated options and serve as attention checks.

Web Appendix C2: **Self-reported real-world behaviors and correlations with time preference parameters estimated with 20 DEEP questions - Study 2**

		Correlation with β		Correlation with δ	
Behavior	**Question Wording**	***r***	***p***	***r***	***p***
Diet	To what extent do you monitor your diet in terms of caloric, fat, carbohydrate, cholesterol, and/or sodium intake?	.03	.361	-.08	.003
Doctor Visits	How often do you visit a doctor for routine check-ups (physicals)?	.02	.513	-.03	.306
Sunscreen Use	How often do you use sunscreen when exposed to harsh sunlight?	.01	.675	-.08	.006
Tattoos	How many permanent tattoos do you have, if any?	-.03	.281	.09	.001
Driving	How often do you drive in a way that your driver's education teacher would consider "reckless"?	.08	.006	-.08	.004
Speeding Tickets	How many speeding tickets (or something similar) have you received in the last 5 years?	-.01	.623	-.03	.293
Credit Card - Friend	Compared to your friends who are close to you in age, how much have you taken out in loans for education (e.g., student loans or loans to cover job training or certification)?	.11	<.001	-.04	.159
Credit Card - Family	Compared to the other members of your family in your generation—brothers, sisters, and cousins close to your age—how much have you taken out in loans for education (e.g., student loans or loans to cover job training or certification)?	.10	<.001	-.03	.272
Mortgage - Friend	Compared to your friends who are close to you in age, how much money have you taken out for mortgage(s) to buy a home or homes?	.14	<.001	-.16	<.001
Mortgage - Family	Compared to the other members of your family in your generation—brothers, sisters, and cousins close to your age—how much money have you taken out for mortgage(s) to buy a home or homes?	.14	<.001	-.15	<.001
Credit Card Debt - Self	How much credit card debt do you currently have (total, across all of your credit cards)?	.03	.314	.00	.979
Tax Withholding	If you owe taxes on your income or salary, how much do you withhold (pay) with each paycheck?	.13	<.001	-.10	.001
Coupon Use	To what extent do you use coupons or rebate offers when you shop?	-.03	0.279	.00	.925
Age of first marriage	If you are currently or have been married, at what AGE did you first get married?	.00	>.99	-.03	.414
Age of first child	If you have children, at what AGE did you have your first child?	.03	.449	-.11	.005
Regular Bedtime	Do you go to bed the same time every night?	.04	.163	-.06	.037
Dishwashing	How often do you leave dirty dishes overnight?	.09	.002	-.04	.160
Punctuality	To what extent are you on time to appointments, engagements, or meetings (both personal- and business-related)?	-.02	.557	.01	.751
Procrastination	When given a long-term assignment or task, when do you tend to start it?	-.01	.591	.02	.438
BMI	Body-Mass Index calculated from height and weight	.03	.301	.05	.103
Activity Time (hours/week)	How many hours per week are you physically active (for example, walking, working around the house, working out)?	-.03	.314	.05	.060
Fitness Time (hours/week)	How many of those hours represent exercise primarily intended to improve or maintain your health or fitness?	-.06	.020	.07	.010
Exercise Intensity	If you do any exercise primarily for health or fitness, how would you rate its intensity?	.02	.396	-.04	.165

Overeating	In a typical week, how often do you eat more than you think you should eat?	.01	.713	-.03	.348
Diet plan	Are you currently following a specific diet plan?	.04	.201	-.06	.038
Dentist	How often do you visit your dentist for a check-up?	.02	.405	-.11	<.001
Floss	How often do you floss your teeth?	.05	.088	-.05	.059
Prescription Drugs	When your doctor gives you a prescription to fill at the drugstore (excluding birth control), do you follow it exactly (for example, by going to the drugstore, picking up the medication, taking all of the medication on schedule, and finishing the entire prescription)?	.04	.147	-.04	.205
Coffee	How would you describe your intake of coffee—how often do you consume it?	-.01	.857	.01	.762
Nicotine	How would you describe your intake of nicotine—how often do you consume it?	-.06	.023	.09	.002
Alcohol	How would you describe your intake of alcohol—how often do you consume it?	.05	.064	.00	.949
Drugs	How would you describe your intake of recreational drugs (e.g., marijuana)—how often do you consume them?	-.02	.560	.01	.640
Saving - Percent Income	Over the past three years, what percentage of your income have you saved? (Please include savings into retirement plans and any other form of savings that you do.)	.00	.870	-.07	.014
Wealth - Friends	Compared to your friends who are close to you in age, how much wealth have you accumulated? (Wealth includes retirement savings, stocks, bonds, and mutual funds you own, money in bank accounts, the value of your home minus the mortgage, etc.)	.04	.114	-.14	<.001
Wealth - Family	Compared to the other members of your family in your generation—brothers, sisters, and cousins close to your age-how much wealth have you accumulated?	.04	.110	-.13	<.001
Credit Card - Pay Late	If you have any credit cards, over the past two years how many times were you charged a late fee for making a credit card payment after the deadline?	.04	.191	-.06	.028
Credit card - Pay Full	If you have any credit cards, over the past two years, how often have you paid your credit card bill in full, as opposed to paying less than the full amount? (Paying in full means carrying no debt to the next month's bill.)	.04	.169	-.18	<.001
Gambling	On average, how many days per month do you gamble money, including visiting casinos, buying lottery tickets, betting on sports, playing poker, etc.	-.13	<.001	.05	.060

***Web Appendix C3:* Study 2 supplementary tables and figures**

Table C1

Linear model to test external validity for predicting BLB task time preferences - Study 2a.

	Prediction error δ				*Prediction error* β			
	arctan $(APE^{\delta}_{BLB_{iq}}) = b_0 + b_{1,q} + \epsilon_{iq}$				arctan $(APE^{\beta}_{BLB_{iq}}) = b_0 + b_{1,q} + \epsilon_{iq}$			
	b	*p*	2.5%	97.5%	b	*p*	2.5%	97.5%
b_0	.720	<.001	.706	.734	.774	<.001	.761	.787
$b_{1,2}$	.052	<.001	.034	.069	-.003	.454	-.009	.004
$b_{1,3}$	.031	<.001	.019	.043	-.012	.023	-.022	-.002
$b_{1,4}$	.024	.021	.004	.045	-.005	.276	-.014	.004
$b_{1,5}$	.002	.878	-.020	.023	-.027	<.001	-.038	-.016
$b_{1,6}$	-.008	.495	-.030	.015	-.036	<.001	-.048	-.024
$b_{1,7}$	-.017	.149	-.040	.006	-.052	<.001	-.066	-.038
$b_{1,8}$	-.014	.228	-.037	.009	-.048	<.001	-.062	-.034
$b_{1,9}$	-.018	.131	-.042	.005	-.049	<.001	-.063	-.034
$b_{1,10}$	-.010	.410	-.033	.014	-.050	<.001	-.066	-.035
$b_{1,11}$	-.012	.320	-.036	.012	-.044	<.001	-.058	-.029
$b_{1,12}$	-.011	.340	-.034	.012	-.044	<.001	-.059	-.029
$b_{1,13}$	-.001	.946	-.024	.022	-.039	<.001	-.054	-.024
$b_{1,14}$	.001	.940	-.022	.023	-.034	<.001	-.048	-.019
$b_{1,15}$	.005	.632	-.017	.027	-.030	<.001	-.045	-.016
$b_{1,16}$	.009	.443	-.013	.030	-.029	<.001	-.044	-.015
$b_{1,17}$	.011	.318	-.010	.032	-.028	<.001	-.043	-.013
$b_{1,18}$	.012	.291	-.010	.033	-.022	.003	-.036	-.007
$b_{1,19}$	.012	.287	-.010	.033	-.023	.002	-.037	-.008
$b_{1,20}$	.008	.456	-.013	.029	-.021	.004	-.035	-.007

Note. Question numbers are dummy coded, the errors are clustered per individual.

Table C2

Helmert and reverse-Helmert contrasts for external validity for predicting BLB time preferences - Study 2

	δ		β			δ		β	
Helmert contrasts	Δ	*p*	Δ	*p*	Reverse-Helmert contrasts	Δ	*p*	Δ	*p*
Q1 vs. Q2-20	.00	.31	.03	<.01	-	-	-	-	-
Q2 vs. Q3-20	.05	<.01	.03	<.01	Q2 vs. Q1	.05	<.01	.00	.34
Q3 vs. Q4-20	.03	<.01	.02	<.01	Q3 vs. Q1-2	.01	.19	-.01	.06
Q4 vs. Q5-20	.03	<.01	.03	<.01	Q4 vs. Q1-3	.00	.31	.00	.48
Q5 vs. Q6-20	.00	.32	.01	.10	Q5 vs. Q1-4	-.03	<.01	-.02	<.01
Q6 vs. Q7-20	-.01	.25	.00	.46	Q6 vs. Q1-5	-.03	<.01	-.03	<.01
Q7 vs. Q8-20	-.02	.03	-.02	.02	Q7 vs. Q1-6	-.03	<.01	-.04	<.01
Q8 vs. Q9-20	-.01	.05	-.01	.04	Q8 vs. Q1-7	-.03	<.01	-.03	<.01
Q9 vs. Q10-20	-.02	.01	-.02	.02	Q9 vs. Q1-8	-.03	<.01	-.03	<.01
Q10 vs. Q11-20	-.01	.07	-.02	.01	Q10 vs. Q1-9	-.02	.04	-.02	<.01
Q11 vs. Q12-20	-.02	.03	-.01	.04	Q11 vs. Q1-10	-.02	.03	-.02	.02
Q12 vs. Q13-20	-.02	.02	-.02	.02	Q12 vs. Q1-11	-.01	.05	-.01	.03
Q13 vs. Q14-20	-.01	.15	-.01	.05	Q13 vs. Q1-12	.00	.39	-.01	.14
Q14 vs. Q15-20	-.01	.16	-.01	.13	Q14 vs. Q1-13	.00	.48	.00	.38
Q15 vs. Q16-20	.00	.28	-.01	.22	Q15 vs. Q1-14	.00	.31	.00	.43
Q16 vs. Q17-20	.00	.40	-.01	.21	Q16 vs. Q1-15	.01	.20	.00	.38
Q17 vs. Q8-20	.00	.48	-.01	.19	Q17 vs. Q1-16	.01	.14	.00	.33
Q18 vs. Q19-20	.00	.41	.00	.48	Q18 vs. Q1-17	.01	.13	.01	.10
Q19 vs. Q20	.00	.33	.00	.40	Q19 vs. Q1-18	.01	.15	.01	.14
-	-	-	-	-	Q20 vs. Q1-19	.00	.29	.01	.10

Note. Δ = Difference in $M(\text{arctan}\ (APE^{\delta}_{BLB_{iq}}))$, p= p-value for t-test of Δ against 0.

Table C3

Linear model to test external validity for predicting real-world behavior index - Study 2a.

	Prediction error δ arctan $\left(APE^{\delta}_{beh_{iq}}\right) = b_0 + b_{1,q} + \epsilon_{iq}$				*Prediction error* β arctan $(APE^{\beta}_{beh_{iq}}) = b_0 + b_{1,q} + \epsilon_{iq}$			
	b	***p***	**2.5%**	**97.5%**	**b**	***p***	**2.5%**	**97.5%**
b_0	.780	<.001	.771	.789	.772	<.001	.760	.784
$b_{1,2}$	-.003	.641	-.014	.009	-.001	.848	-.008	.007
$b_{1,3}$	-.003	.508	-.011	.006	.001	.874	-.007	.009
$b_{1,4}$	-.009	.217	-.022	.005	-.003	.479	-.012	.005
$b_{1,5}$	-.012	.078	-.025	.001	-.002	.683	-.014	.009
$b_{1,6}$	-.017	.024	-.031	-.002	-.004	.470	-.016	.007
$b_{1,7}$	-.024	.001	-.039	-.010	-.005	.426	-.017	.007
$b_{1,8}$	-.024	.002	-.039	-.009	-.006	.369	-.017	.006
$b_{1,9}$	-.026	.001	-.042	-.011	-.007	.240	-.020	.005
$b_{1,10}$	-.019	.010	-.034	-.005	-.007	.263	-.020	.005
$b_{1,11}$	-.022	.004	-.037	-.007	-.008	.236	-.020	.005
$b_{1,12}$	-.019	.011	-.034	-.004	-.007	.268	-.020	.005
$b_{1,13}$	-.026	.001	-.041	-.010	-.007	.296	-.019	.006
$b_{1,14}$	-.020	.008	-.034	-.005	-.006	.349	-.018	.007
$b_{1,15}$	-.017	.022	-.032	-.002	-.005	.447	-.017	.008
$b_{1,16}$	-.025	.001	-.040	-.010	-.003	.599	-.016	.009
$b_{1,17}$	-.021	.003	-.036	-.007	-.002	.751	-.014	.010
$b_{1,18}$	-.023	.002	-.037	-.008	-.001	.890	-.013	.011
$b_{1,19}$	-.021	.005	-.035	-.006	.000	.981	-.012	.012
$b_{1,20}$	-.026	.000	-.041	-.012	-.001	.935	-.013	.012

Table C4

Helmert and reverse-Helmert contrasts for predicting real-world behavior index - Study 2a.

	δ		β			δ		B	
Helmert contrasts	Δ	*p*	Δ	*p*	Reverse-Helmert contrasts	Δ	*p*	Δ	*p*
Q1 vs. Q2-20	.019	.001	.004	.266	-	-	-	-	-
Q2 vs. Q3-20	.017	.003	.003	.290	Q2 vs. Q1	-.003	.282	-.001	.449
Q3 vs. Q4-20	.018	.002	.005	.213	Q3 vs. Q1-2	-.002	.367	.001	.431
Q4 vs. Q5-20	.013	.033	.001	.418	Q4 vs. Q1-3	-.007	.114	-.003	.301
Q5 vs. Q6-20	.010	.077	.002	.366	Q5 vs. Q1-4	-.008	.078	-.002	.402
Q6 vs. Q7-20	.006	.224	.000	.483	Q6 vs. Q1-5	-.012	.037	-.003	.307
Q7 vs. Q8-20	-.002	.383	.000	.478	Q7 vs. Q1-6	-.017	.005	-.003	.307
Q8 vs. Q9-20	-.002	.419	-.001	.436	Q8 vs. Q1-7	-.014	.020	-.003	.299
Q9 vs. Q10-20	-.005	.270	-.003	.312	Q9 vs. Q1-8	-.015	.018	-.005	.228
Q10 vs. Q11-20	.003	.369	-.003	.311	Q10 vs. Q1-9	-.006	.183	-.004	.268
Q11 vs. Q12-20	.000	.492	-.004	.267	Q11 vs. Q1-10	-.008	.119	-.004	.268
Q12 vs. Q13-20	.003	.333	-.004	.264	Q12 vs. Q1-11	-.005	.261	-.003	.309
Q13 vs. Q14-20	-.004	.314	-.004	.259	Q13 vs. Q1-12	-.011	.072	-.003	.349
Q14 vs. Q15-20	.003	.367	-.004	.261	Q14 vs. Q1-13	-.004	.289	-.002	.398
Q15 vs. Q16-20	.006	.208	-.003	.289	Q15 vs. Q1-14	-.001	.431	.000	.475
Q16 vs. Q17-20	-.002	.393	-.002	.346	Q16 vs. Q1-15	-.009	.112	.001	.428
Q17 vs. Q8-20	.002	.395	-.001	.405	Q17 vs. Q1-16	-.005	.253	.002	.352
Q18 vs. Q19-20	.001	.472	-.001	.465	Q18 vs. Q1-17	-.006	.199	.003	.293
Q19 vs. Q20	.006	.220	.000	.476	Q19 vs. Q1-18	-.003	.315	.004	.264
-	-	-	-	-	Q20 vs. Q1-19	-.009	.109	.003	.294

$Note.\Delta\ =$ Difference in $M(\text{arctan}(APE^{\delta}_{beh_iQ}))$, *p*= *p*-value for t-test of Δ against 0.

Table C5

Linear model to test external validity for predicting FICO credit score - Study 2b.

	Prediction error δ				*Prediction error* δ			
	$\arctan(APE^{\delta}_{FICOQ}) = b_0 + b_{1,q} + \epsilon_{iq}$				$\arctan(APE^{\delta}_{FICOQ}) = b_0 + b_{1,q} + \epsilon_{iq}$			
	b	*p*	2.5%	97.5%	b	*p*	2.5%	97.5%
b_0	.780	.000	.763	.796	.779	.000	.763	.795
$b_{1,2}$	-.007	.502	-.029	.014	-.001	.871	-.017	.014
$b_{1,3}$	.000	.970	-.018	.017	.000	.992	-.013	.012
$b_{1,4}$	-.014	.120	-.031	.004	.001	.839	-.011	.014
$b_{1,5}$	-.020	.035	-.038	-.001	-.007	.434	-.024	.010
$b_{1,6}$	-.028	.004	-.048	-.009	-.010	.291	-.027	.008
$b_{1,7}$	-.032	.002	-.051	-.012	-.016	.119	-.036	.004
$b_{1,8}$	-.033	.001	-.052	-.013	-.017	.095	-.037	.003
$b_{1,9}$	-.033	.001	-.052	-.013	-.016	.116	-.036	.004
$b_{1,10}$	-.033	.001	-.052	-.013	-.025	.023	-.046	-.003
$b_{1,11}$	-.034	.000	-.053	-.015	-.022	.031	-.043	-.002
$b_{1,12}$	-.028	.003	-.047	-.010	-.025	.017	-.046	-.004
$b_{1,13}$	-.022	.009	-.039	-.005	-.022	.031	-.041	-.002
$b_{1,14}$	-.021	.013	-.038	-.004	-.019	.058	-.038	.001
$b_{1,15}$	-.019	.026	-.036	-.002	-.022	.033	-.042	-.002
$b_{1,16}$	-.018	.035	-.034	-.001	-.023	.027	-.044	-.003
$b_{1,17}$	-.019	.026	-.036	-.002	-.025	.027	-.046	-.003
$b_{1,18}$	-.018	.033	-.035	-.001	-.020	.065	-.041	.001
$b_{1,19}$	-.021	.015	-.038	-.004	-.018	.088	-.039	.003
$b_{1,20}$	-.019	.022	-.035	-.003	-.018	.096	-.039	.003

Note. Individual-specific effects are not included in this table for parsimony. The coefficients were inverted for δ so that higher values are more predictive.

Table C6

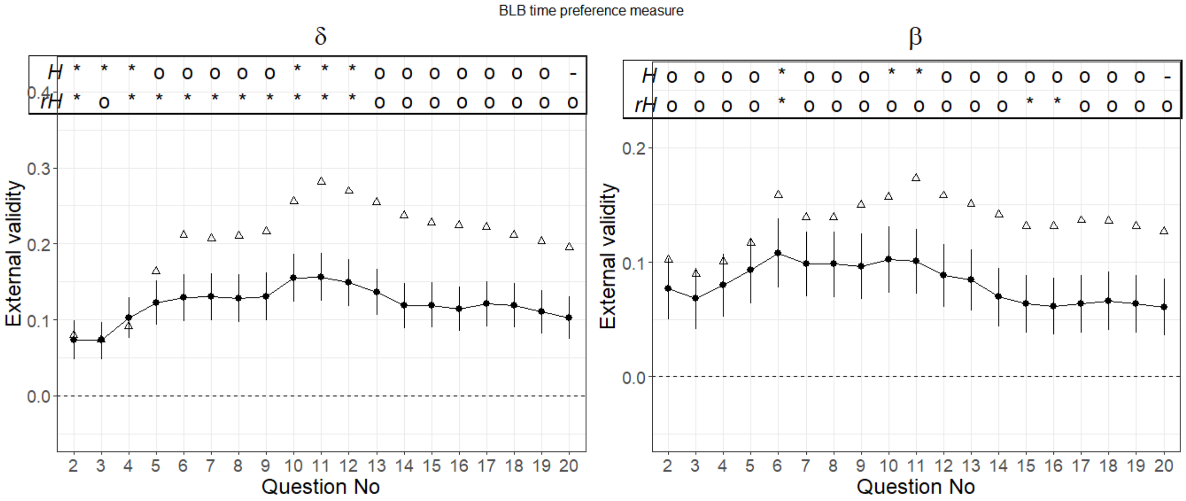
Helmert and reverse-Helmert contrasts tests for predicting FICO credit scores – Study 2b.

	Δ		β			δ		β	
Helmert contrasts	Δ	*p*	Δ	*p*	Reverse-Helmert contrasts	Δ	*p*	Δ	*p*
Q1 vs. Q2-20	.022	.012	.016	.045	-	-	-	-	-
Q2 vs. Q3-20	.015	.064	.016	.049	Q2 vs. Q1	-.007	.202	-.001	.435
Q3 vs. Q4-20	.024	.007	.018	.035	Q3 vs. Q1-2	.003	.345	.001	.472
Q4 vs. Q5-20	.011	.152	.020	.019	Q4 vs. Q1-3	-.011	.127	.002	.417
Q5 vs. Q6-20	.005	.314	.013	.095	Q5 vs. Q1-4	-.014	.082	-.007	.214
Q6 vs. Q7-20	-.003	.383	.011	.147	Q6 vs. Q1-5	-.020	.034	-.008	.181
Q7 vs. Q8-20	-.007	.276	.005	.322	Q7 vs. Q1-6	-.020	.042	-.013	.088
Q8 vs. Q9-20	-.009	.215	.004	.349	Q8 vs. Q1-7	-.018	.052	-.012	.106
Q9 vs. Q10-20	-.010	.197	.006	.310	Q9 vs. Q1-8	-.016	.081	-.010	.158
Q10 vs. Q11-20	-.011	.177	-.003	.386	Q10 vs. Q1-9	-.014	.112	-.017	.049
Q11 vs. Q12-20	-.013	.119	-.001	.462	Q11 vs. Q1-10	-.014	.113	-.013	.100
Q12 vs. Q13-20	-.009	.216	-.004	.352	Q12 vs. Q1-11	-.007	.266	-.015	.083
Q13 vs. Q14-20	-.003	.391	-.001	.466	Q13 vs. Q1-12	.000	.485	-.010	.169
Q14 vs. Q15-20	-.002	.415	.002	.426	Q14 vs. Q1-13	.001	.478	-.007	.266
Q15 vs. Q16-20	.000	.496	-.001	.466	Q15 vs. Q1-14	.003	.394	-.009	.199
Q16 vs. Q17-20	.002	.433	-.003	.386	Q16 vs. Q1-15	.004	.355	-.010	.176
Q17 vs. Q8-20	.000	.488	-.006	.303	Q17 vs. Q1-16	.002	.418	-.011	.168
Q18 vs. Q19-20	.002	.434	-.002	.437	Q18 vs. Q1-17	.003	.392	-.005	.312
Q19 vs. Q20	-.002	.433	.000	.492	Q19 vs. Q1-18	.000	.494	-.003	.378
-	-	-	-	-	Q20 vs. Q1-19	.002	.428	-.003	.391

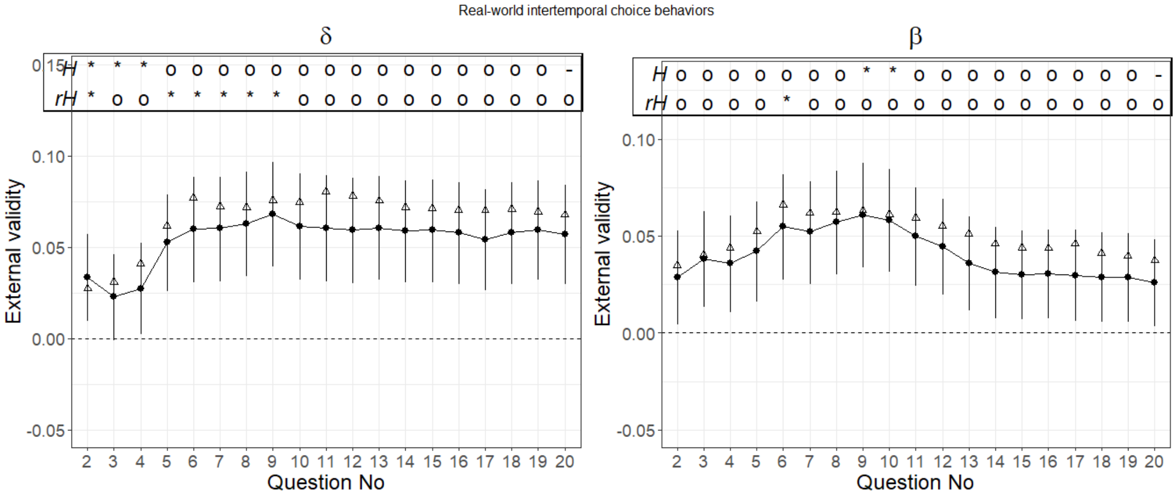
$Note. \Delta\ =$ Difference in $M(\arctan(APE^{\delta}_{FICOQ}))$, *p*= p-value for t-test of Δ against 0.

Figure C2. ***External validity for DEEP Time estimates with questions in random order***

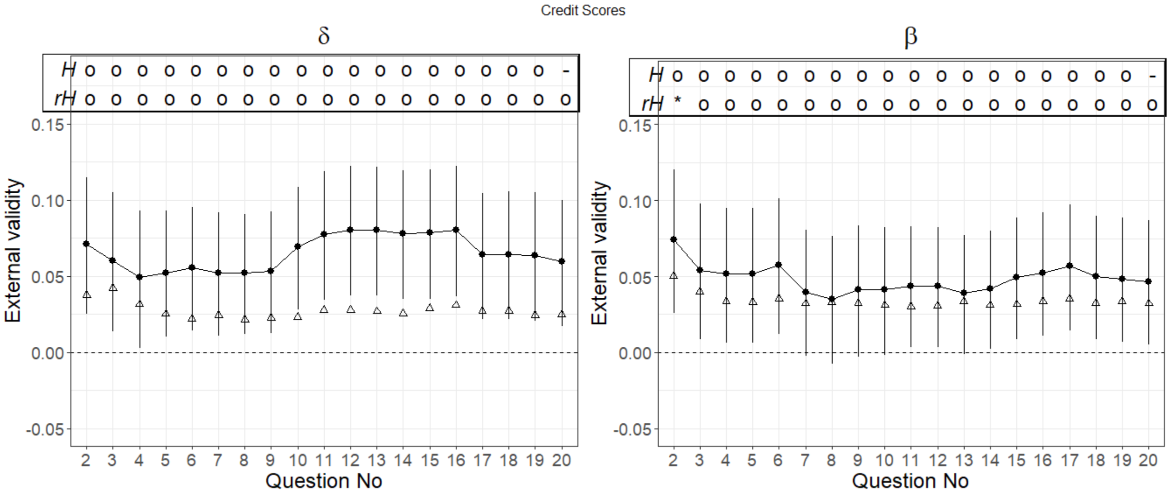
Study 2a: External validity for the BLB time preference measure.



Study 2a: External validity for real-world intertemporal choice behaviors.



Study 2b: External validity for consumer credit scores.



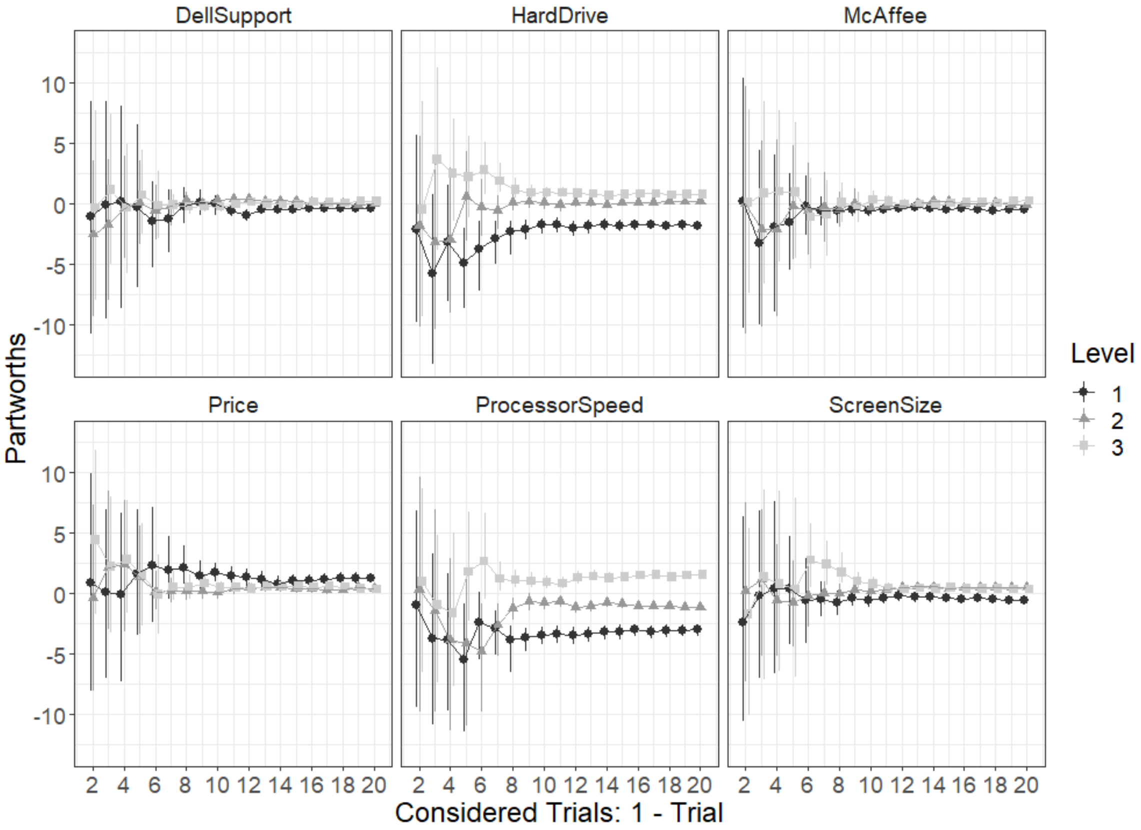
Note: External validity for DEEP Time estimates with questions in random order, for the BLB (Bartels, Li, and Bharti 2021) time preference measure (top panel) and the composite index of 26 real-world intertemporal choice behaviors in Study 2a (middle panel), as well as for credit scores in Study 3 (bottom panel). The points depict question-by-question external validity measures (1 - mean absolute percentage error) and the error bars indicate 95% confidence intervals around it. The symbols above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant for each question number (*: $p < .05$; o: $p \geq .05$). Triangles show the explained variance (R^2) of δ/β for the DV.

WEB APPENDIX D: STUDY 3 SUPPLEMENTAL MATERIALS

Stan model code for estimating the individual partworths in Study 3

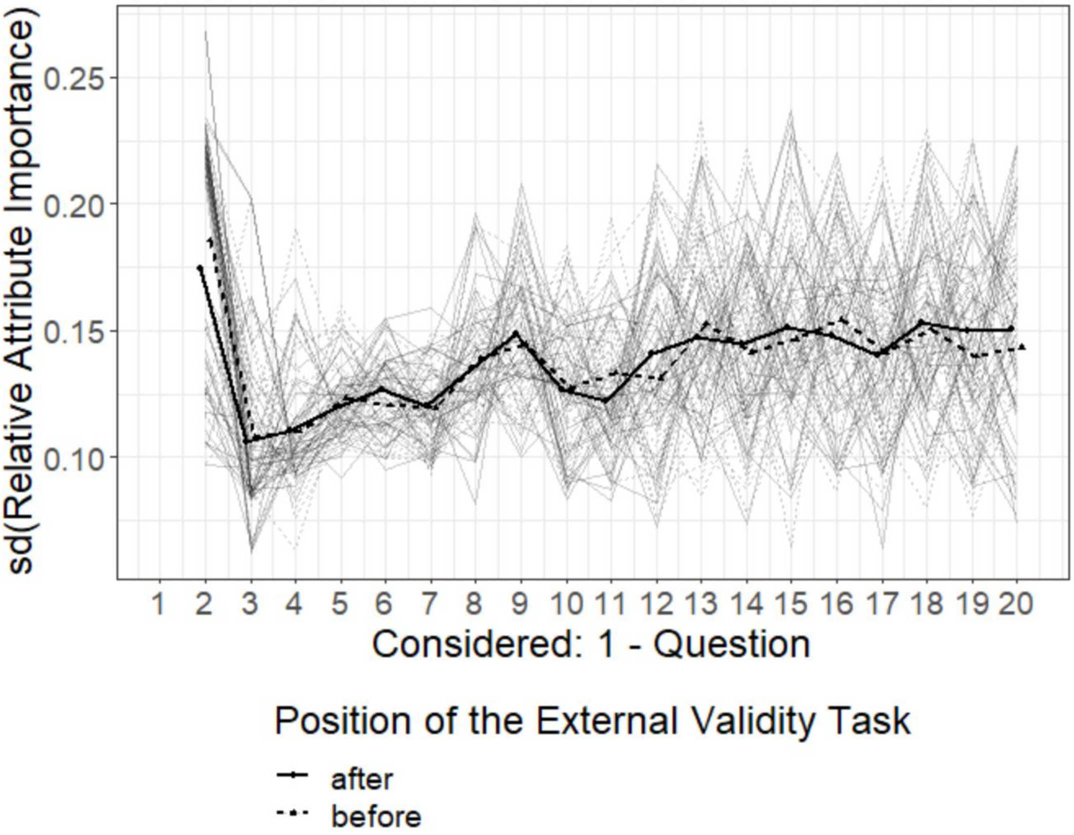
```
data {
     int<lower=2> C; // No of alternatives (choices) in each scenario
     int<lower=1> K; // No of covariates of alternatives
     int<lower=1> R; // No of respondents
     int<lower=1> S; // No of scenarios per respondent

     int<lower=1,upper=C> Y[R, S]; // observed choices
     matrix[C, K] X[S]; // matrix of attributes for each obs
   }
parameters {
     vector[K] Theta;
     matrix[K,R] alpha;
     corr_matrix[K] Omega;
     vector<lower=0>[K] tau;
   }
transformed parameters {
     matrix[K, R] Beta;
     matrix[K,K] L;
     cov_matrix[K] Sigma = quad_form_diag(Omega, tau);
     L = cholesky_decompose(Sigma);
     for (r in 1:R) {
       Beta[,r] = Theta + L * alpha[,r];
     }

   }
   model {
     //priors
     to_vector(Theta) ~ normal(0, 5); //to_vector(Theta)~ normal(0, 10);
     tau ~ cauchy(0, 1);
     Omega ~ lkj_corr(2);
     //likelihood
     for (r in 1:R) {
       alpha[,r] ~ std_normal();
       //Beta[,r] ~ multi_normal(Theta, Sigma);
       for (s in 1:S)
         Y[r,s] ~ categorical_logit(X[s]*Beta[,r]);
     }
   }
```

Figure D1. Population-level partworths for the three levels of each attribute



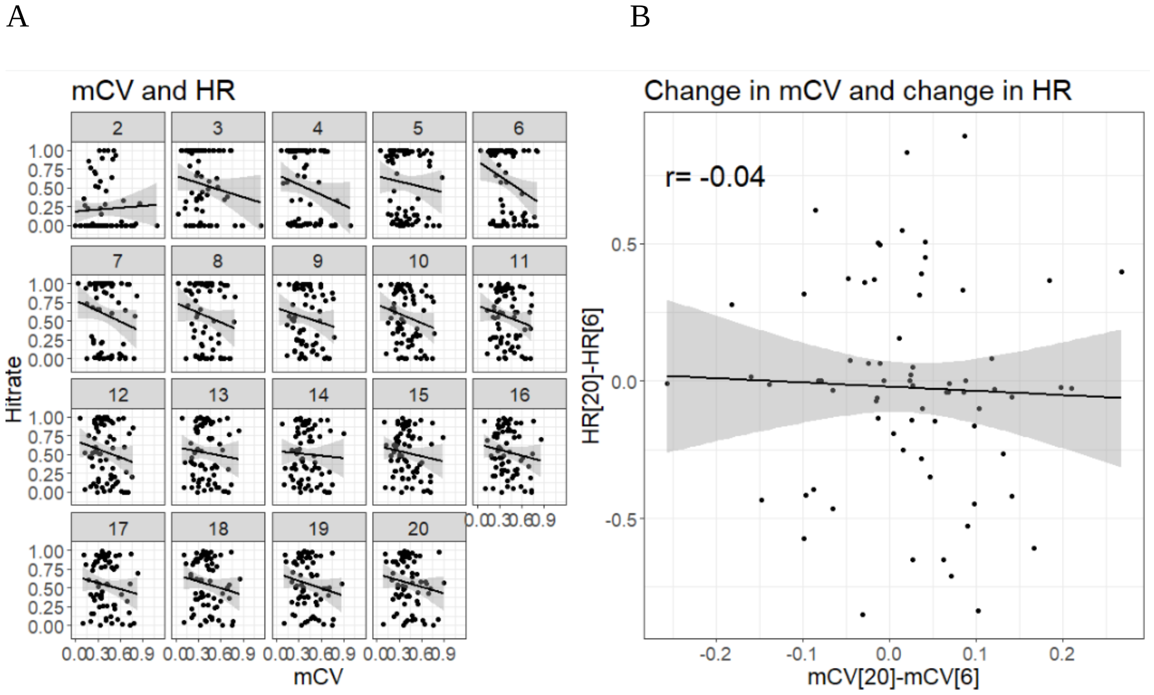
Note: Population-level partworths for the three levels of each attribute as a function of the number of questions considered for the estimation. E.g., the points at question 10 represents the population-level partworth estimates based on participants' choices in the first 10 questions. The points (dots, squares, and triangles) show the 50th percentile and error bars represent the 95% highest density interval (HDI) of the population-level posterior distribution.

Figure D2. Standard deviation of the relative attribute importance



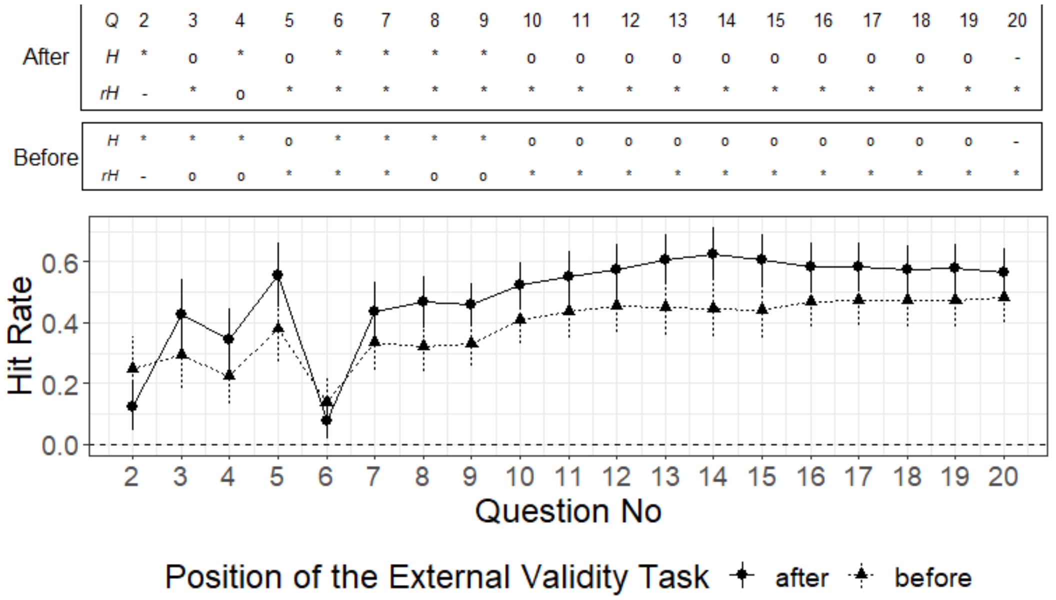
Note: Standard deviation of the relative attribute importance across attributes plotted as a function of the number of questions considered for the partworth estimation. Each gray line illustrates one participant’s standard deviation of the relative attribute importance for each question number. The bold lines illustrate the averages across participants as a function of whether the external validity task was before or after the measurement task.

Figure D3. Coefficient of variation of visited alternatives per attribute versus hit rate



Note: Scatterplot of coefficient of variation (mCV) of visited alternatives per attribute and hit rate (HR) per question number in panel A and change in mCV and HR between the peak in hit rate at question six and the end of the conjoint task in panel B.

Figure D4. Average hit rate for predicting the external validity task



	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	-
	rH	-	*	o	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*	*
Before	*H*	*	*	*	o	*	*	*	*	o	o	o	o	o	o	o	o	o	o	-
	rH	-	o	o	*	*	*	o	o	*	*	*	*	*	*	*	*	*	*	*

Note: Average hit rate for predicting the external validity task, as a function of the number of questions in randomized order considered for the partworth estimation, and the position of the external validity task. The asterisks and circles above the plot illustrate whether the Helmert (H) and reverse-Helmert (rH) contrasts are significant (*: $p < .05$; o: $p \geq .05$) in the two conditions (after: dots and solid line; before: triangles and dashed line).

Table D1. Helmert and reverse-Helmert contrasts tests for predicting hit rate.

	Position of the External Validity Task								
	Before		After			Before		After	
Helmert contrasts	Δ	*P*	Δ	*p*	Reverse-Helmert contrasts	Δ	*p*	Δ	*p*
Q2 vs. Q3-20	-.111	.012	-.398	<.001	-	-	-	-	-
Q3 vs. Q4-20	.099	.022	-.084	.057	Q3 vs. Q2	.204	.001	.318	<.001
Q4 vs. Q5-20	.041	.219	-.126	.008	Q4 vs. Q2-3	.042	.218	.124	.009
Q5 vs. Q6-20	.025	.316	-.025	.312	Q5 vs. Q2-4	.010	.425	.186	<.001
Q6 vs. Q7-20	-.070	.089	.156	.001	Q6 vs. Q2-5	-.082	.062	.310	<.001
Q7 vs. Q8-20	-.052	.152	.145	.002	Q7 vs. Q2-6	-.045	.201	.226	<.001
Q8 vs. Q9-20	-.037	.226	.075	.062	Q8 vs. Q2-7	-.020	.355	.113	.014
Q9 vs. Q10-20	-.003	.470	-.008	.436	Q9 vs. Q2-8	.017	.364	.015	.388
Q10 vs. Q11-20	-.018	.340	.029	.266	Q10 vs. Q2-9	.002	.481	.047	.172
Q11 vs. Q12-20	-.006	.449	.017	.358	Q11 vs. Q2-10	.015	.374	.028	.287
Q12 vs. Q13-20	.002	.479	-.018	.352	Q12 vs. Q2-11	.021	.327	-.008	.438
Q13 vs. Q14-20	-.006	.443	-.044	.178	Q13 vs. Q2-12	.012	.401	-.027	.296
Q14 vs. Q15-20	-.007	.433	-.078	.043	Q14 vs. Q2-13	.011	.406	-.048	.166
Q15 vs. Q16-20	-.018	.332	-.044	.150	Q15 vs. Q2-14	.002	.481	-.003	.474
Q16 vs. Q17-20	-.011	.394	-.037	.191	Q16 vs. Q2-15	.011	.406	.012	.403
Q17 vs. Q8-20	-.008	.424	-.036	.197	Q17 vs. Q2-16	.015	.371	.021	.328
Q18 vs. Q19-20	.005	.458	-.027	.255	Q18 vs. Q2-17	.025	.291	.038	.205
Q19 vs. Q20	-.010	.412	-.007	.431	Q19 vs. Q2-18	.014	.376	.059	.095
-	-	-	-	-	Q20 vs. Q2v-19	.023	.305	.063	.079

Note. Δ = Difference in Hit rate, *p* = *p*-value for t-test of Δ against 0.

Table D2. Results of the mediation analysis

	Total effects (c)		**Direct effects (c')**		**Effects of *q* on mCV (a)**		**Indirect effects (a×b)**
Q	**Est.**	***p***	**Est.**	***p***	**Est.**	***p***	**Est. [95% CI]**
3	-.121	.016	-.144	.005	-.048	.162	.014[-.004,.057]
4	-.148	.010	-.165	.003	-.037	.174	.011[-.003,.041]
5	-.017	.776	-.037	.558	-.041	.066	.012[.001,.042]
6	.116	.083	.100	.142	-.034	.073	.010[.000,.035]
7	.113	.057	.097	.114	-.034	.05	.010[.001,.033]
8	.055	.287	.036	.509	-.04	.011	.012[.002,.037]
9	-.029	.500	-.048	.290	-.039	.007	.012[.002,.035]
10	.005	.903	-.013	.744	-.038	.005	.011[.002,.036]
11	.007	.824	-.012	.701	-.04	<.001	.012[.002,.036]
12	-.045	.199	-.060	.094	-.032	.001	.010[.002,.029]
13	-.066	.072	-.079	.036	-.027	.002	.008[.001,.026]
14	-.088	.015	-.099	.007	-.023	.003	.007[.001,.023]
15	-.049	.032	-.059	.013	-.021	.001	.006[.001,.021]
16	-.037	.035	-.045	.013	-.018	<.001	.005[.001,.017]
17	-.035	.045	-.043	.016	-.016	.001	.005[.001,.014]
18	-.022	.024	-.028	.005	-.012	.005	.004[.001,.011]
19	-.005	.191	-.008	.052	-.007	.027	.002[.000,.006]

Effect of mCV on hit rate (b)		Effect of position on hit rate (β_1)		Interaction of position with mCV ($\beta_{3,before}$)	
Est.	*p*	Est.	*p*	Est.	*p*
-.477	.059	-.193	.293	.380	.362

Table D3. Model comparison for Study 3

Dependent Variable **Model**	**df**	**BIC**	**AIC**	**R^2[Not adjusted]**
CV (Options per attribute)				
Linear with Clustered SE	5.000	324.259	298.038	.044
Spline Regression	8.170	369.923	327.078	.028
Variance in relative attribute importance (Q2-20)				
Linear with Clustered SE	5.000	-5128.533	-5154.498	.038
Spline Regression	29.697	-5239.788	-5394.000	.226
Variance in relative attribute importance (Q3-20)				
Linear with Clustered SE	5.000	-5170.026	-5195.721	.289
Spline Regression	9.097	-5165.842	-5212.590	.303