Cognitively Optimized Measurement of Preferences

Ye Li, University of California Riverside, USA
Daniel Wall, Rutgers University, USA
Eric Johnson, Columbia University, USA
Olivier Toubia, Columbia University, USA

Every question respondents answer provides data but consumes limited attention—increasing reliance on heuristics. We examine this tradeoff using a formal model and test predictions in three studies. In an adaptive time preference measure, reliability and validity peak after only 8 questions; MouseLab tracing shows evidence of increasing heuristic use.

[to cite]:

[url]:
http://www.acrwebsite.org/volumes/1022597/volumes/v44/NA-44

[copyright notice]:
This work is copyrighted by The Association for Consumer Research. For permission to copy or use this work in whole or in part, please contact the Copyright Clearance Center at http://www.copyright.com/.
Cognitively Optimized Measurement of Preferences

Ye Li, University of California Riverside, USA
Daniel Wall, Rutgers University, USA
Eric Johnson, Columbia University, USA
Olivier Toubia, Columbia University, USA

EXTENDED ABSTRACT

Researchers, managers, and policy-makers are increasingly interested in measuring individual differences to predict behavior in the field. From measuring time and risk preferences in economics to contingent valuation in environmental science and policy, and from consumer surveys to conjoint analysis in marketing, predicting people’s behavior is a big deal.

Every time we ask a respondent a question, we do two things: (1) We obtain data about some parameter(s) of interest, such as their discount rate or the part worth of a product attribute, and (2) We expend some of the respondent’s limited attentional resources. This paper focuses on the tradeoff between additional precision and the depletion of attentional resources.

The intersection of precision and depletion is of increasing interest because of two related applications: The first is preference elicitation. Marketing has expended considerable effort to study how to model choices in one environment in order to predict choices in another. A good example of this is the use of conjoint analysis in new product design. The problem of how many questions to ask to get the best model is central to designing elicitation procedures. A related version of this question arises when we use adaptive procedures thought to minimize respondent fatigue (e.g., Green et al 1991; Toubia et al 2003).

The second application is more general: To test theories, researchers in marketing and psychology have increasingly turned to asking respondents to make many decisions. Several different trends have led to an increased use of many elicitation questions: More complex models require more observations as the number of parameters increases, and new techniques, such as functional magnetic resonance imaging (fMRI), demand many decisions to overcome noise in the desired physiological signals.

Both of these developments raise a general concern about the changes that occur as we ask respondents more questions. By asking more questions, are we changing the underlying process by which respondents answer them? A specific concern about the effects of answering questions on attention is ego depletion (Baumeister, Bratslavsky, Muraven, and Tice 1998), an idea originating in the social psychological literature. Perhaps, as respondents use this scarce executive resource to answer more questions (Vohs et al 2008; Wang et al 2010), they will move toward simpler, heuristic processes that may differ from those used for decisions in the real world application (Pocheptsova et al 2009). For example, Levav and colleagues (2010) found that asking more questions increases the acceptance of defaults by the decision-maker, particularly when those questions are cognitively difficult.

In this paper, we develop a framework for understanding the tradeoff between additional signal and noise. Our goal is to help identify a priori changes in the underlying decision processes that might occur with more questions answered and how those changes will affect the reliability and validity of the models that are estimated on the data. One more specific goal of this research to help researchers think about the best number of questions to ask, a term we will call $q^*$, in various situations.

We first develop a conceptual model that will generate predictions and insights to test in a series of three experiments that measure time preferences. We focus on time preferences for several reasons: (1) they are among the strongest individual difference predictors in the social sciences (Chabris et al 2008; Reimers et al 2009); (2) they are usually measured using minimal external information, often as binary intertemporal choices between a smaller outcome that is available sooner (e.g., $50 in today) and a larger outcome that is available later (e.g., $60 in 1 month); (3) there is a rapidly growing literature of both descriptive and normative models; and finally (4) they have become increasingly important in marketing, particularly in areas as diverse as food choice and consumer finance. We conclude by discussing general implications for preference measurement.

CONCEPTUAL MODEL

To study how repeated questioning may affect the reliability and external validity of preference measurement, we first develop a stylized model. In particular, we are interested in exploring conditions under which reliability and/or 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 the questions, the amount of variations in heuristics across decision makers and across sessions, and the speed with which respondents adopt heuristics within a session.

Our model reflects two countervailing forces that vary as the number of questions increase. First, more questions provide more data, which reduces the standard error of parameter estimates—i.e., estimates converges to some value. Second, more questions may increase the intensity of heuristic use, which may introduce bias. In other words, as more questions are added, our estimates become more precise (i.e., they have less standard error and converge to some value), but also potentially more biased (i.e., the value to which they converge may be further from the truth).

In this setting, we argue that a respondent, as they become fatigued or disinterested, might be tempted to adopt a heuristic. For example, for intertemporal choices, they may see if the difference in money is proportionally smaller or larger than the difference in time and chose the later option only if the payment seems to justify the wait, a heuristic that has recently gained significant empirical support (Marzilli et al 2015; Scholten and Read 2010; Scholten, Read, and Sanborn 2014). While such data will provide a clear estimate of the parameter in this estimation setting, the parameter that we estimate may not apply to situations when that heuristic is not used. For example, perhaps decisions to delay are made in advance so the smaller sooner is not available now. This removes any overweighting of the immediate reward, or present bias, from the decision, a factor that is confounded in the estimation procedure we have described. Alternately, in the actual savings decision, the time frames might be described in years or in terms of a specific date, and a different heuristic might be used in these circumstances (LeBoeuf 2006; Read et al 2005). If this is true, the increased accuracy that might be offered by asking more questions might be diminished or even offset by the increased use of heuristic processes.

Formally, we assume that $Y$ varies with true individual parameters, denoted by $X_i$, and allow for idiosyncratic variations, reflected by a normally-distributed error term, $e$;

$$Y_i = X_i + e_i \quad \text{where } e_i \sim N(0, \sigma)$$

We consider a researcher who tries to predict $Y$ based on an estimate of $X_i$ generated by asking respondents questions (e.g., in-
tertemporal choices), a process which is potentially repeated over multiple sessions. $\hat{X}_{i m q}$ is the estimate of $X_i$ for decision maker $i$ in session $w$ after question $q$. $\hat{X}_{i m q}$ reflects the two countervailing forces mentioned above: additional questions lead to convergence, but the quantity to which the estimate converges may become increasingly further from the truth due to the use of heuristics. In particular, as decision-makers answer questions, decision-makers may switch to a cognitively less demanding heuristic. This heuristic may use a subset of the information or make attribute-based comparisons rather than calculate expected utilities, and therefore using the heuristic produces estimates that may differ from $X_i$. If heuristic use is systematic, the result may be a biased estimate of $X_i$.

We denote the potential bias introduced by the use of heuristics as $\hat{b}_i$. The actual bias is likely to increase as the questionnaire progresses, and to converge to the potential bias as the number of questions increases. Therefore, we weigh $\hat{b}_i$ by a function $\sqrt{a(q)}$ which is non-decreasing in the number of questions.

With these assumptions, the combined effect of the true underlying value $X_i$ and the use of heuristics after question $q$ is: $X_i + \sqrt{a(q)} \hat{b}_i$. One the other hand, asking more questions reduces the standard error of the estimate—i.e., there is convergence with more questions. Respondents likely report their preferences with error, leading to imperfect estimates. However, as respondents answer more questions, estimates converge to the underlying construct, $X_i + \hat{b}_i$. We model this convergence using an additional normally-distributed error term, $\gamma(q)$. To reflect the idea that estimates converge within a session, we model the variance of $\eta_{iwq}$ as decreasing in $q$, that is, we assume: $\eta_{iwq} \sim N(0, \sigma \sqrt{\gamma(q)})$, where $\gamma(q)$ is a non-increasing function of $q$.

In sum, we assume that our estimate of $X_i$ for decision maker $i$ in session $w$ after question $q$, is given as follows:

$$\hat{X}_{iwq} = X_i + \sqrt{a(q)} \hat{b}_i + \eta_{iwq}$$

Where $\eta_{iwq}$ is a non-decreasing function of $q$, $\eta_{iwq} \sim N(0, \sigma \sqrt{\gamma(q)})$, and $\gamma(q)$ is a non-increasing function of $q$. Again, the second and third right-hand terms in Equation (2) capture the two countervailing forces mentioned above: $\sqrt{a(q)} \hat{b}_i$ leads to an increasing bias due to heuristic use as $q$ increases, and leads to a reduction in standard error as $q$ increase.

In interest of space, we skip the proofs and present the most interesting model predictions about reliability and external validity.

**Hypothesis 1:** The use of heuristics can give rise to an eventual decrease in reliability with more questions asked, leading to an inverted U-shaped pattern (i.e., a peak) in reliability.

**Hypothesis 2:** A decrease in reliability with more questions asked is more likely when heuristics vary more across sessions.

**Hypothesis 3:** A decrease in reliability with more questions asked is more likely when questions are asked in a more efficient manner, i.e., when convergence within a session is faster and/or when the use of heuristics increases at a slower rate.

**Hypothesis 4:** The use of heuristics can give rise to an eventual decrease in external validity, leading to an inverted-U shaped (i.e., a peak) pattern of external validity.

**Hypothesis 5:** An inverted-U shaped external validity pattern is more likely when the use of heuristics is more pronounced. Moreover, the peak happens earlier when the use of heuristics is more pronounced.

**Hypothesis 6:** An inverted-U shaped external validity pattern is more likely when questions are asked in a more efficient manner. Moreover, the peak happens earlier when questions are asked more efficiently.

---

**STUDY 1: CONCURRENT VALIDITY IN DEEP TIME**

Study 1 reanalyzed data from DEEP Time and the time preference external validity task collected by Toubia and colleagues (2013). This task consisted of 20 binary choices between smaller, sooner and large, later cash rewards. To assess external validity for estimated DEEP parameters, participants first completed a different time preference measure similar to the one used in Benhabib, Bisin, and Schotter (BBS, 2010). Specifically, each participant completed eight intertemporal matching questions, specifying the amount of money that would make them indifferent between a smaller, sooner amount and a larger, later amount. Half of the questions required filling out the smaller, sooner amount and half the larger, later amount.

After participants completed the BBS task, they were given the DEEP Time task, as described above. To make choices incentive compatible, 1 in every 100 participants was randomly selected to have one of their chosen DEEP Time alternatives paid for real money at the chosen time. All payments were made electronically.

**Results**

**Concurrent validity**

In order to investigate the effect of asking additional adaptive questions, we used the hierarchical Bayes approach to estimate $\beta$ and $\delta$ after each of the 20 DEEP questions. That is, we fit all participants’ responses to only the first question, to the first two questions, etc., up to all 20 questions, resulting in 20 sets of $\beta$’s and $\delta$’s for each participant.

To understand the effect of answering additional questions on external validity, we calculated the absolute deviation (in dollars) between participants’ actual responses on the BBS task with the predicted responses generated by their QTD model parameters. We then computed the median absolute deviation across the eight BBS questions for each participant, and took the median of each participant’s medians. We found similar trends when looking at the mean of each question’s median absolute deviation. We repeated this process using the QTD parameters estimated after each of the 20 DEEP time questions.

In order to demonstrate a statistical plateau in validity, we examined Pearson correlations (across the eight external validity questions) between each participant’s actual responses and their predicted responses based on QTD parameters after each successive DEEP time question. We then converted these correlations to $z$-scores using Fisher’s r-to-z transformation for ease of comparison, with clear evidence of an increase after the sixth question and no significant change after seven questions. To statistically test the trend, we analyzed a multilevel model on the r-to-z transformed correlations, with participant intercepts and fixed effects for each question order (20 separate effects), thus controlling for the correlated errors within each participant.
STUDY 2: RELIABILITY AND PREDICTIVE VALIDITY FOR REAL-WORLD TASKS.

In Study 2, in addition to measuring concurrent validity as a function of number of questions asked, we examined predictive validity by see how well DEEP Time parameters predicted self-reports of field intertemporal choice behaviors across health, financial, and personal domains (e.g., flossing, smoking, savings, and credit card repayment). We also assessed test-retest reliability by comparing participants’ responses on two sessions separated by 4 months.

Methods

We reanalyzed data obtained from Bartels & Li (2015), who recruited a large, broadly sampled set of 1308 participants (542 female, ranging in age from 18 to 86, with a mean age of 40.91) including 603 from Amazon Mechanical Turk and 705 from Pureprofile (a market research firm).

Data consisted of DEEP Time data as a part of a larger study on how time preferences relate to field intertemporal choices. In addition to DEEP Time, participants completed a static set of intertemporal choices designed using Item Response theory, which can be used for tests of concurrent validity. Finally, participants self-reported the degree to which they exhibited a variety of field behaviors related to intertemporal choice (e.g., going to the dentist, eating healthily, and saving for retirement). All measures were collected twice, with about 4 months delay between sessions and an 83% retention rate. No time preference, field behavior, or demographic variable predicted attrition from the study. For more details, see Bartels & Li (2015).

Similar to study 1, we used the hierarchical Bayesian method to estimate the DEEP Time estimates after each of 20 questions answered, separately for each session.

Results

Reliability

Test-retest reliability specifically requires that participants’ parameters elicited at different times are highly correlated. An important prediction of our model is a plateau or peak in test-retest reliability. This is, in our model, due to the heuristic process dominating the complex process. Test-retest reliability for the QTD beta and delta parameters estimated by DEEP time after each question plateaued after about six questions.

Concurrent Validity

We compared the DEEP estimates after each question to an external intertemporal titrator. Our model of choices predicts that external validity plateaus, or in some cases peaks, due to fact that the heuristic processes is elicitation specific and may not be measuring the “truth.”

Using the QTD parameters after each of the DEEP questions, we calculated the estimated indifference point for all of the 12 matching questions, which can be found in the appendix. Then for each question within each participant we calculated the Pearson’s correlation between the predicted and actual indifference points, meaning that we had 20 correlations for each of the 1308 participants. Due to skewness we then transformed these correlations via Fisher’s r-to-z transformation. Next we ran a multilevel model with fixed effects for each of the 20 DEEP questions and random effects for participant. Fixed effects showed that concurrent validity plateaued at eight questions.

Predictive validity for real-world behaviors

Further in Study 2 we had multiple real world measures of time discounting. We oriented all variables such that higher numbers indicated more patience. For instance, the more often you go to the dentist the higher your score. To determine if the items could be combined into one measure, we looked at Cronbach’s alpha. The raw alpha for all of the items combined was .57. Since alpha was sufficiently large, we z-scored all of the relevant items and combined them into one real world intertemporal behavior factor. Then we ran a multilevel model with fixed effects for each of the questions along with the beta parameters. Predictive validity peaked at six to eight questions for beta and five to six questions for delta.

Discussion

Study 2 expands on Study 1 by reproducing the concurrent validity plateau with a different time discounting instrument. Further Study 2 shows that test-retest reliability peaks at a similar point and that validity with real world intertemporal choices peaks at about 4 questions. We do not, however, see how information acquisition changes as more questions are asked. Study 3 analyzes process data to track the evolution of acquisition across contexts, how these contexts lead to varying choices, and how well parameters within and across contexts predict real world decisions.

STUDY 3: SHIFTING HEURISTICS

While studies 1 and 2 replicate a plateau in predictive validity, they do not indicate why that plateau exists. As the number of questions grows, respondents may, because of fatigue, adopt a simplified strategy that minimizes effort. One version of this fatigue argument might say that respondents respond randomly, at the extreme simply not looking at any information or a small but random subset of information.

If this were the case, we would expect predictive validity to decrease with the number of questions each additional question adds random error to the estimated. The second possibility is that decision-makers adopt a heuristic strategy that simplifies choices, but that is consistent with the preferences established in the first few choices. If this were the case, we would expect predictive validity to remain constant, since each choice simply reflects prior choices and carries very little additional information.

However, to the extent that such heuristics reflect the context used in the experiment, we may see parameters that reflect both true tastes and the context. This suggests a rather interesting hypothesis: That predictions will be better within a context than across contexts.

Methods

Participants

We recruited 533 participants from Amazon Mechanical Turk and used MouselabWeb (Willensen and Johnson, 2010), a process tracing technique to study information acquisition.

We constructed a set of 16 choices and randomly added or subtracted a small percentage of each of the amounts to make the amounts harder to process. We then generated a random presentation order for the 16 choices. Participants saw two blocks of these sixteen choices in the same random presentation order in a 2 (first block: day vs. hour) x 2 (second block: day vs. hour). The original 16 items were created with days as the time unit. To convert these to hour items we multiplied the number of days times 24. For both days and hours, we labeled today items “now.” This was done because we felt that 0 hours has a different connotation than 0 days.). The second block could be consistent with the first block (e.g. if the participant saw hour items in the first block they saw the same items again in the second block) or could be inconsistent (e.g., if the participant saw day items in the first block the saw hour items in the second block.)

We also measured the same self-reports of real-world intertemporal choices as in study 2.
Results

Choices

We ran a multilevel regression with choice as the dependent measure (smaller sooner = 0, larger later = 1) and random effects of participant. We included main effects of context (day vs hour), which block the choice was in (first or second), and whether or not the second block was consistent with the first block (consistent vs. inconsistent), as well as all possible interactions.

Participants made significantly more impatience choices in the hours condition as compared to the days condition. The effect of hour versus day was more pronounced in the first block of choices. This attenuation in the second block could be due to a persistent heuristic which is created in the first block and lasts until the second block.

One of the major advantages of Mouselab Web is that we acquire how participants search. We leverage this advantage to investigate if heuristics develop as more questions are asked. One metric which can be used to describe how people are searching is the Payne Index (PI). The PI is calculated for each choice by subtracting the number of attribute transitions from the number of alternative based transitions then dividing by the sum of attribute and alternative based transitions. A higher PI means that participants make more within option transitions (e.g. a transition between the larger later amount and the larger later time) than between option transitions (e.g. a transition between the larger later amount and the smaller sooner amount).

We also ran a model to see if there was a difference in PI between the first block and the second block. Since choices were shown in the exact same random order we subtracted the block 2 paPI from the block 1 PI to get a measure of how differently participants were acquiring data in the second half. We ran a multilevel model predicting for each choice the difference in PI with varying intercepts for each participant. We also included main effects of what was the first frame that the participant saw (days or hours), whether the second block had a consistent or inconsistent framing with the first block, and an index of the number of choices made, and all possible interactions. The results suggest that participants maintain their heuristic in the consistent condition, but alter it, especially early, when seeing items which have an inconsistent frame.

REFERENCES


