Optimizing Choice Architectures

Mark Schneider
University of Alabama

Cary Deck
University of Alabama

Mikhael Shor
University of Connecticut

Tibor Besedes
Georgia Tech

Sudipta Sarangi
Virginia Tech

May 4, 2018

Abstract

This paper investigates decision quality in large choice sets across several choice architectures in three studies. In the first controlled experiment, we manipulate two features of a choice architecture – the response mode (for ranking alternatives) and presentation mode (for presenting alternatives). Our design objectively ranks all sixteen choice options in each choice set, and makes it possible to observe decision quality directly, independent of attitudes toward risk. We find joint presentation outperforms separate presentation, and that choice response modes outperform 'happiness ratings' which outperform hypothetical monetary valuations. We also apply classical welfare criteria to assess the performance of the architectures. Our key finding is that low cognitive reflection subjects (as measured by the cognitive reflection test) perform better given a *large* choice set than given smaller sets collectively containing the same alternatives. This illustrates a basic tradeoff confronting choice architectures: For a fixed choice set, fewer options improve decision quality within that set, but require architectures to elicit multiple responses, increasing opportunities for errors. One follow-up study demonstrates the robustness of the response mode result in a comparison using the tournament presentation mode. A second follow-up study reveals that the impact of incentivizing monetary valuations depends on cognitive reflection.

Key Words: choice architecture, response mode, presentation mode, choice overload, experiments

JEL Codes: C91, D03

1. Introduction

Making good economic decisions is desirable for both individual welfare and achieving social goals. Sub-optimal or lower quality decisions, such as choosing a dominated option, can reduce the decision maker's welfare relative to what could have been achieved. Additionally, if the suboptimal decisions result in inappropriate health insurance plans or insufficient retirement savings, for example, the decisions also likely impose a cost on society. However, error prone decision making has long been recognized as a fact of life, embodied in Alexander Pope's famous statement, "To err is human." Although shunned by the classical economic models of the 1950's and 60's, this fact is widely recognized in behavioral economics today. This has led to an increased emphasis on choice architecture where the goal is to "nudge" participants towards decisions that are both individually and socially optimal by manipulating the decision-making environment. Yet, the properties that determine effective choice architectures are not fully understood.

Building on the design in Besedeš et al. (2015), this paper uses experiments to systematically test for the effectiveness of two general features of a choice architecture: (*i*) the response mode (how rankings over alternatives are expressed) and, (*ii*) the presentation mode (how information is presented). Our design enables us to consider a basic tradeoff between presentation complexity (number of alternatives presented at once) and response complexity (the number of discrete responses required by the architecture). While it is generally understood that decision making is better in smaller choice sets (see Besedeš et al 2012a, 2012b), decomposing a large choice set into a series of small ones may not necessarily result in improved decision making (see Besedeš et al 2015). Once we fix the size of a choice set, smaller presentation sets may improve average decision quality *per response*, but require architectures to elicit multiple responses, increasing the opportunity for error. This tradeoff suggests a novel implication: If error rates are sufficiently high, *smaller* presentation sets may actually *reduce* decision quality.

In our primary study, we consider three response modes – direct choice among a set of options, subjective happiness ratings of options, and monetary valuations of options. Under direct choice, decision makers select one of the available options from each choice set. For subjective happiness ratings, decision makers rate each option in the choice set on an emoticon scale, reflecting how happy each option makes

them. For monetary valuations, decision makers specify their maximum willingness to pay for each option in dollars. Of these response modes, direct choice is the one encountered most frequently in making decisions. Monetary valuation requires one to think very precisely about the value of every option and in this respect may be the most difficult and most time consuming. Response modes similar to happiness ratings include quality and satisfaction ratings of restaurants, books, movies, and other consumption experiences. Direct choices, ratings, and valuations are three of the classic response modes in the judgment and decision-making literature, although they have to date been primarily used to study the *consistency of preferences* across response modes, rather than the *optimality of decisions* across response modes. Our design enables us to utilize these classic response modes to go beyond studying consistency of preferences and study their possible role in making better decisions. In our design all choice options can be ranked according to stochastic dominance, enabling us to objectively rank the available choices, and *observe decision-making quality* in a way that is not contaminated by subjects' attitudes toward risk or other unobservable idiosyncratic properties of preferences.

Our choice of response modes was also motivated by the possibility that different response modes may induce different decision-making processes. For instance, it seems plausible that an emoticon scale (or subjective happiness rating) increases reliance on feelings, whereas a pricing task may increase reliance on calculation. If choosing by calculation is a superior decision-making strategy, especially when choices involve well-defined probabilities and monetary outcomes, one might predict that performance will be superior in the monetary valuation response mode. Alternatively, if relying on feeling and intuition (going with your gut) is a better strategy for decision making, especially when the choice set is large, one might predict better performance on the emoticon response mode. Finally, if one views decision makers as well-adapted to choice tasks, constantly facing discrete choices in the environment, and rarely providing explicit ratings or prices, one might predict superior performance under the choice response mode.

One feature of a choice architecture that varies across response modes is the degree to which the response mode constrains the possible responses for a choice set of a given size. A highly constrained response mode, such as a single direct choice admits only n possible response patterns for a choice set with

n alternatives. In contrast, an emoticon scale is less constrained in that each of the n options is given a rating on a scale with m possible ratings per option. The monetary valuation task is even less constrained than the rating scale, since each of the n options can be assigned any monetary value (between \$0.00 and \$20.00 in our experiment). One may also view the degree to which a response mode does not constrain responses as indexing the complexity of the response mode, with response modes requiring more responses per choice set (e.g., rating each item) and permitting a larger range of responses per item as being more complex.

We also consider two main presentation modes – joint presentation (all options are presented simultaneously) and separate presentation (each option is presented one at a time, in isolation). A variety of studies have documented that the order in which information such as risks and benefits is presented or acquired can significantly affect choices (Aimone et al., 2016; Arieli, et al., 2011; Bergus et al., 2002). However, it is not clear a priori whether choices presented jointly will produce higher quality decisions than separate presentation of alternatives. Under expected utility theory, any lottery has its own value, independent of other choice alternatives. If the only presentation mode effect at work is due to choice overload, one might predict that evaluating each option in isolation avoids the paralyzing effect on choice of seeing many complex options simultaneously. However, evaluating each option in isolation may also require one to remember how each previous alternative was valued. In this respect, decision quality in presentation modes may be affected by this tradeoff between the complexity of joint presentation and the memory required for separate presentation.

The previous literature on response mode and presentation mode effects has focused on *inconsistencies* (preference reversals) across response modes, and not the *decision quality* (selection of dominant vs. dominated choices) induced within response modes or presentation modes. For instance, the literature on response mode effects has mostly examined preference reversals across response modes (e.g., the pricing-choice reversals identified by Lichtenstein and Slovic (1971) and the pricing-rating reversals identified by Slovic et al. (2007)). Similarly, the literature on behavior under different presentation modes has also focused on the identification of preference reversals across presentation modes (the joint-separate

reversals identified by Hsee (1996), Hsee et al., (1999), and Hsee and Zhang (2010), and the comparative ignorance effect identified by Fox and Tversky (1995)). Two papers that do look at dominance in presentation modes are Hsee (1998) and List (2002) and both find that people are better able to value two options at once rather than one option at a time using a pricing response mode. However, these two studies both used a between-subjects design even within the separate presentation mode and thus under their design, no single subject could provide valuations that reveal a preference for the dominated option. In addition, these studies do not consider the large choice sets that are reflective of many economic situations.

Choice architecture may be seen as the 'engineering' branch of behavioral economics, perhaps analogous to how Roth (2002) envisioned mechanism design as the engineering branch of game theory. Whereas mechanism design analyzes how behavior changes in response to normatively *relevant* incentives (e.g., changes in monetary payoffs), choice architecture analyzes how behavior changes in response to normatively *irrelevant* features of the decision task (e.g., changes in framing, response mode, or presentation mode). Choice architecture has many practical applications such as increasing revenue through the presentation and organization of a grocery store (Reutskaja et al., 2011), designing healthcare plans or presenting healthcare information in a manner that helps people select the best plan for themselves (Peters et al., 2007), designing retirement pension plans to increase employee saving (Thaler and Benartzi, 2004), and designing the presentation of nutritional information to promote healthier food choices (Downs et al., 2009). These applications highlight basic questions about the principles underlying the selection of optimal or welfare-improving choices. Such basic questions as which response mode and which mode of presenting information lead to the most efficient welfare outcomes served as the motivation behind our study.

The options in our design comprise a choice set of sixteen lotteries with different expected payoffs whose outcomes are distributed over twelve possible states with pre-defined probabilities. These may be viewed as stylized versions of insurance plans, retirement plans, or financial investments. By varying the response mode, we can identify which method of eliciting rankings is most effective in producing high quality choices. By varying the presentation mode, we can identify whether providing complete information (presenting all sixteen options simultaneously) or incremental information (presenting one option at a time)

leads to better decisions for large choice sets. We also examine subject heterogeneity and test whether reflective thinkers perform better than intuitive thinkers across architectures using a version of the cognitive reflection test (Frederick, 2005; Toplak et al, 2014).

For the choice task, the joint presentation mode is compared to a tournament' architecture in which participants choose between four disjoint subsets of the overall choice set and then choose among their chosen options in a 'final four' round. We include this architecture since Besedeš et al. (2015) found this to be best among architectures using the choice response mode at helping individuals make optimal choices.

Across all subjects in our studies, we find the persistent ranking that a choice response mode outperforms a happiness rating response mode, which in turn outperforms a pricing mode and that joint presentation yields better performance than presenting each option sequentially. An important additional finding is that choice architectures need to account for individual differences in nuanced ways: when designing a choice architecture for a fixed choice set (in our case of size 16), there is a fundamental tradeoff between the number of options presented at once and the number of responses required by the decision maker. In particular, participants with low scores on the cognitive reflection test (CRT) performed best on the "choose one of sixteen" architecture, whereas moderate and high scorers on the CRT performed best overall on a tournament-style architecture.[1] This presents something of a puzzle. If the low CRT participants can perform fairly well in a sixteen-item choice set, one might suspect they would do even better in a tournament-style architecture where they choose among just four options at a time. We show that this puzzle is plausibly explained by a simple error rate model that assumes: (*i*) that low CRT participants have larger error rates than high CRT participants and (*ii*) that, for a fixed CRT level, error rates are larger from larger choice sets than from smaller choice sets. The key observation to note is that the tournament architecture requires multiple responses, increasing opportunities for error. If low CRT participants have sufficiently high error rates in four-item choice sets, when confronted with a series of such choice sets, they

[1] We refer to subjects with high, medium and low scores in the Cognitive Reflection Test as high CRT, medium CRT and low CRT respectively. The CRT does not necessarily imply intelligence, but rather reflects the approach a person naturally uses for problem-solving (intuitive versus reflective).

have more opportunities for error than in the 'choose one' architecture. However, as higher CRT subjects have lower error rates, they can benefit more from the tournament architecture. In the simple error rate model in Section 5 we solve for the unique error rates across CRT groups and choice set sizes. It turns out that assumptions (*i*) and (*ii*) noted above are necessary conditions for the simple error rate model to fit the observed data exactly.

In this first study, we also find that high CRT subjects perform better across all architectures than low CRT subjects. For both ratings and valuations, we find joint presentation to considerably outperform separate presentation, and we find less-constrained response modes to marginally outperform more constrained response modes.

We report two additional studies aimed at understanding the robustness of our initial findings. In our second study we use the happiness rating and pricing modes in a tournament structure similar to that used in the first study with the direct choice response mode. We show that the response mode ranking of choice, happiness, and payment is robust to this presentation mode. Our third study considers the effect of making the pricing response incentivized rather than hypothetical as in the first two studies. In this study subjects identify their maximum willingness to pay for an option, both in joint and separate presentation modes, using an incentive compatible mechanism which results in the subject purchasing an option. Ultimately, we find making payments incentivized does not impact decision quality among low CRT scores but may lead to better decision making for higher CRT individuals.

In terms of identifying the optimal architecture, considering the welfare of low CRT, medium CRT, and high CRT participants, there is no architecture that dominates across all CRT levels (no architecture is Pareto efficient) in our data. Instead, the optimal architecture depends on the social welfare criterion that a policy maker prefers to implement. In particular, for our experiment, John Rawls' (1971) maximin criterion, which helps the demographic with the lowest welfare, favors the architecture involving a single direct choice from a sixteen-item choice set, whereas Harsanyi's (1955) utilitarian welfare criterion, which

maximizes the average welfare of individuals in society, favors the tournament architecture involving five direct choices, each from a four-item choice set.

2. Identifying Optimal and Efficient Choice Architectures

We consider simple choice architectures that are defined to be a pair $(\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{P}})$ where $\boldsymbol{\mathcal{R}}$ is a response mode and $\boldsymbol{\mathcal{P}}$ is a presentation mode. We let $\boldsymbol{\mathcal{R}} \in \{r, v, c\}$, where r is a happiness *rating* task, v is a monetary *valuation* task, and c is a *choice* task. We let $\boldsymbol{\mathcal{P}} \in \{j, p, s\}$, where j is a 'joint' presentation mode (all options are presented simultaneously), p is a partial presentation mode (different subsets of the choice set are presented together), s is a 'separate' presentation mode (all options are presented individually, in isolation). Our design contains six choice architectures: $(\boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{P}}) = (r, j), (r, s), (v, j), (v, s), (c, j), (c, p)$.

Thus, the happiness rating and valuation tasks are shown both with all options at once and with each option presented separately. For one of the direct choice tasks, options were displayed all at once. The other choice task, (c, p), is the bench-mark best-performing choice architecture (the 'choice tournament' architecture from Besedeš et al., 2015) in which a large choice set is divided into a number of equally sized smaller choice sets with the decision maker selecting one option from each small choice set. The final decision is then made over all options selected in each of the smaller choices sets.

In our first study, we were primarily interested in comparing joint versus separate presentation modes, and choice, rating, and monetary valuation (pricing) response modes. We included the tournament architecture primarily because it performed best in the study of Besedeš et al. (2015) who considered only variations of choice response modes. Concerned about the possibility of subject fatigue if we included too many architectures, we did not consider variations in the tournament architecture in our base study.

In one follow-up study (Section 7), we do contrast the choice tournament with rating and pricing tournaments. This provides a robustness check and enables us to determine whether it is the tournament style architecture or whether it is primarily the choice response mode that drives performance in this architecture. In another follow-up study (Section 8), we consider joint versus separate presentation modes with incentivized monetary valuation tasks.

We are interested in whether some choice architectures lead to better normative behavior (i.e. a higher likelihood of choosing the optimal option) than others. In particular, we ask the following questions:

(i) Does separate presentation of alternatives improve decision making when the choice set is large, holding the response mode fixed? To be specific, does (r, s) perform better than (r, j) and does (v, s) perform better than (v, j)? Or do people perform better when having all options presented simultaneously even when the choice set is large?

(ii) Do more numerical or calculation-based response modes (such as a monetary valuation task) improve decision making relative to more qualitative or feeling-based response modes (such as a happiness rating task), holding the presentation mode fixed? More precisely, does (v, s) perform better than (r, s) and does (v, j) perform better than (r, j)?) Or do people perform better when making qualitative assessments than when specifying a precise willingness to pay?

(iii) Do more constrained response modes (those with fewer possible responses per option) perform better than less constrained response modes (those with more possible responses per option)?

(iv) Does heterogeneity in cognitive reflection account for heterogeneity in performance across architectures? Do participants who differ in cognitive reflection perform best on the same choice architectures?

(v) Which architectures perform best according to classical welfare criteria (such as Pareto efficiency, Rawls' maximin criterion, and Harsanyi's utilitarian criterion)?

Our design enables us to investigate each of these questions. The results inform whether the response mode and the presentation mode (and their interaction) can facilitate higher quality choices. To our knowledge, this issue has not been addressed in the literature.

We employ a design, building on Besedeš et al. (2015) in which choices can be objectively ranked across different configurations of response modes and presentation modes. Our design enables us to conduct a within-subjects experiment to study the optimality of response modes and presentation modes (whether some response modes or presentation modes systematically induce better decisions).

3. Experimental Design

We tested the performance of six different choice architectures, systematically varying the response mode (how subjects express their choices) and the presentation mode (how options were presented) across architectures using a within-subject design. In each case, the choice set involves 16 lotteries, each with 12 different mutually exclusive and exhaustive states. In each state, a given lottery either pays $0 or $20. A single state is randomly selected to determine payment. A unique feature of our design is that these lotteries can be ranked using stochastic dominance. In other words, we can objectively rank the 16 options, independent of subjects' attitudes toward risk. Consequently, we can directly evaluate the performance of different choice architectures, different response modes, and different presentation modes, based on how frequently they induce optimal choices, or based on how close subjects come to obtaining the best option in each task. The experiment can be accessed at http://gametheory.net/ms3/experiment.pl.

3.1 Options

The 12 possible states of the world were described as types of *Cards* and numbered 1-12. The likelihood of a particular state of the world was determined by the number of cards of that type that were present in a virtual deck of 100 cards. The number of cards of a particular type was referred to as the odds. The 16 lotteries were referred to as *Options* and were lettered A to P. Each option contained different cards (states of the world), with no two options containing the same set of cards. Table 1 shows the 16 options that were used in the experiment.

In Table 1, a check-mark appearing below a particular option indicates that the option contains that card. Selecting an option would result in the subject earning $20 if a card contained by the chosen option were drawn and $0 if a card not contained by the chosen option was drawn. Thus, Option X is a better choice than Option Y if Option X pays $20 for a greater percentage of the cards in the deck, i.e. has a greater probability of paying $20. For example, Option A in Table 1 is the best option, while Option P is the worst as it implies the lowest probability of payment. The order of options and cards was randomized for each subject for each choice architecture and relabeled sequentially from Option A to Option P and from Card 1 to Card 12.

Six different decks of cards were used in the experiment (one for each choice architecture) selected at random without replacement. Moreover, the decks were constructed in such a way that the probability an option would result in a payment of $20 was held fixed across the six decks of cards. To achieve this, we started with a master design with only six states (cards labeled I through VI) and sixteen options designed so that each option covers either three or four of the attributes. We then subdivided some of the six states (and their probabilities) into multiple new states, while preserving each lottery's coverage. That is, if an option contained the card in the six-state design, it also contained all subdivided states. The first column in Table 1 in each of the three panels presents the same six-attribute master design. In the first panel Cards I and V are each split into three new cards and Cards III and IV are split into two new cards. In the second panel Cards II and V are each split into three new cards and Card IV and VI are split into two new cards each. In the third panel Cards III and IV are split into three cards each and Cards I and II into two cards each. By varying how the probability assigned to a card in the six-state design is split among the new cards in the twelve-state design, we created two different probability distributions across the twelve states. We do so in each panel for a total of six different probability distribution functions (PDFs). The variety of PDFs conceals the similarity of options across tasks. Importantly, however, this design preserves the probability with which each option results in a payment. This is similar to how Besedeš et al. (2012a) add additional attributes in their experiment, but with the added advantage that option payoffs are held constant across the six PDFs.

Table 1. Options and Odds

6-state setup		12-state setup			Options															
Card	PDF	Card	PDF1	PDF2	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
I	19	1	10	7	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
		2	4	10	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
		3	5	2	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
II	15	4	15	15	✓						✓	✓	✓	✓		✓				✓
III	22	5	14	9	✓	✓	✓	✓		✓		✓		✓				✓		
		6	8	13	✓	✓	✓	✓		✓		✓		✓				✓		
IV	24	7	11	16	✓	✓	✓	✓	✓		✓		✓		✓		✓			
		8	13	8	✓	✓	✓	✓	✓		✓		✓		✓		✓			
V	12	9	4	3			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
		10	2	4			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
		11	6	5			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
VI	8	12	8	8		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓	✓
Card	PDF	Card	PDF3	PDF4	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
I	19	1	19	19	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
II	15	2	5	2	✓						✓	✓	✓	✓		✓				✓
		3	6	3	✓						✓	✓	✓	✓		✓				✓
		4	4	10	✓						✓	✓	✓	✓		✓				✓
III	22	5	22	22	✓	✓	✓	✓		✓		✓		✓				✓		
IV	24	6	10	6	✓	✓	✓	✓	✓		✓		✓		✓		✓			
		7	14	18	✓	✓	✓	✓	✓		✓		✓		✓		✓			
VI	12	8	3	5			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
		9	2	4			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
		10	7	3			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
VI	8	11	5	7		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓	✓
		12	3	1		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓	✓
Card	PDF	Card	PDF5	PDF6	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
I	19	1	4	13	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
		2	15	6	✓	✓		✓	✓	✓			✓	✓	✓	✓			✓	
II	15	3	9	11	✓						✓	✓	✓	✓		✓				✓
		4	6	4	✓						✓	✓	✓	✓		✓				✓
III	22	5	10	5	✓	✓	✓	✓		✓		✓		✓	✓			✓		
		6	5	7	✓	✓	✓	✓		✓		✓		✓				✓		
		7	7	10	✓	✓	✓	✓		✓		✓		✓				✓		
IV	24	8	13	14	✓	✓	✓	✓	✓		✓		✓		✓		✓			
		9	8	7	✓	✓	✓	✓	✓		✓		✓		✓		✓			
		10	3	3	✓	✓	✓	✓	✓		✓		✓		✓		✓			
V	12	11	12	12			✓		✓	✓	✓	✓				✓	✓	✓	✓	✓
VI	8	12	8	8		✓	✓		✓	✓	✓	✓			✓	✓	✓	✓	✓	✓
		Payoffs			80	73	66	65	63	61	59	57	58	56	51	54	44	42	39	35

3.2 Choice Architectures

Each subject provided responses to each of the following six choice architectures:

1. **Simple choice architecture** (c, j)**:** The *simple choice architecture* explicitly asked subjects to select their most preferred option and presented all sixteen options at once. Subjects only had to click the "Select" button beneath their chosen option and confirm their response. This task, shown in the top panel of Figure 1, involves a choice response mode and a joint presentation mode.
2. **Sequential tournament** (c, p)**:** The *sequential tournament architecture* decomposes the sixteen-option choice set into four choice sets, each with four options. Subjects are asked to choose from each of the four choice sets. A subject's chosen options from these sets are combined into a "final four" round where the subject chooses one of the four previously chosen options as the final choice.
3. **Rating all at once** (r, j)**:** The *rating all at once architecture* asks subjects to provide a happiness rating for each option on a seven-point scale with endpoints of happy and neutral emoticons.[2] Subjects are informed that if they rate one option higher than all others, that will be their selected option. Subjects are also informed that if there is a tie for the highest rated option, then each tied option is equally likely to be the selected option and one is randomly assigned. The rating all at once architecture is depicted in the middle panel of Figure 1 and involves a joint presentation mode.
4. **Rating one at a time** (r, s)**:** The *rating one at a time architecture* is identical to rating all at once, both in visual appearance and in the rules for determining payoffs, except that options are now presented sequentially (in random order), and each subject is asked to rate each option individually (viewing only one option at a time).
5. **Pricing all at once** (v, j)**:** The *pricing all at once architecture* asks subjects to record their maximum willingness to pay for each option. The willingness to pay is hypothetical and the subjects know there is no explicit cost for selecting an option. Subjects are informed that the option

[2] Sad emoticons were not used because each option involved a chance of winning $20 and there was no possibility of losing money. Employing the neutral emoticon instead of a sad emoticon also may encourage more use of the full seven-point scale.

with highest recorded price will be the one selected as their preferred option. Subjects are also informed that if there is a tie for the highest valued option, a random draw will determine the selected option. This architecture involves a pricing response mode and a joint presentation mode and requires the subject's valuation to be entered into the box at the bottom panel of Figure 1.

6. **Pricing one at a time** (v, s)**:** The *pricing one at a time architecture* is identical to pricing all at once, both in visual appearance and in the rules for determining payoffs, except that options are now presented sequentially (in random order), and each subject is asked to price each option individually (viewing only one option at a time).

3.3 Subjects

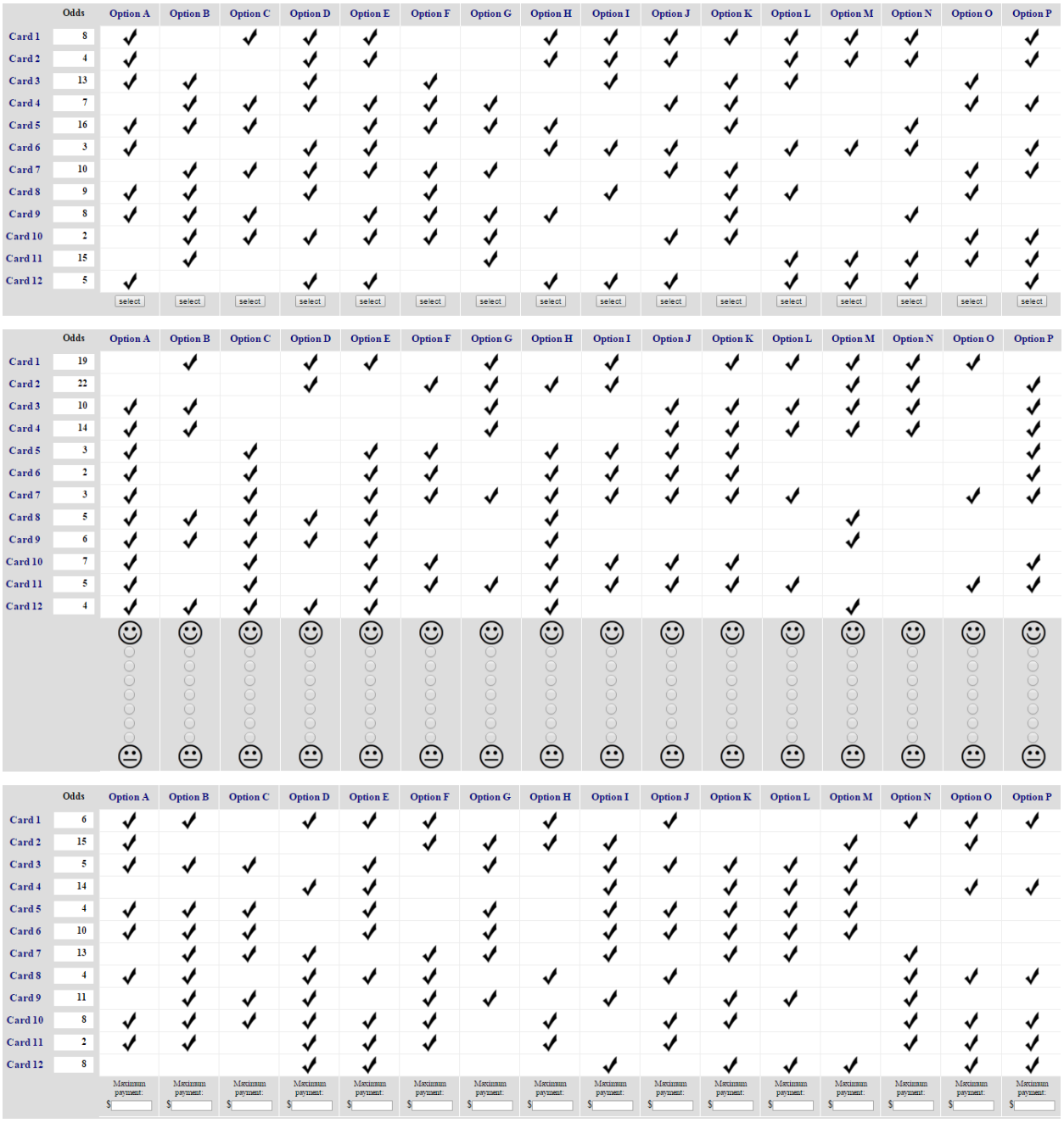
One hundred twenty undergraduate students at a private California university participated in the experiment. Subjects were recruited from a standing pool of volunteers who had not participated in any related study. Each session included twenty-four subjects. Subjects received $7 for participating in addition to any salient earnings.

3.4 Protocol

The six different architectures and six PDFs were presented in random order. The order in which options and cards were presented were randomized within each architecture although the first option was always labeled A and the first card was labeled Card 1. After making their choice(s) for each architecture subjects were presented with a deck of 100 cards each containing cards numbered 1 through 12, with the number of each card corresponding to the odds of the PDF used for that architecture. Subjects would then turn over the cards by clicking on them which would also trigger their shuffling. Once the deck was shuffled, the subject could turn over one card by clicking on any of the 100 cards. This card would determine whether subjects would earn a payment for that architecture: as long as the option they selected or rated as their most preferred one (either by assigning highest happiness rating or highest willingness to pay), they would receive $20 as payment for that architecture. In case of multiple most preferred options ties were broken by randomly assigning one of the tied options.

After completing all six tasks, subjects responded to a brief questionnaire including demographic questions and the seven-question cognitive reflection test or CRT (Toplak et al., 2014), which extends the three-question CRT from Frederick (2005). As explained to subjects in advance, a physical die was rolled by the experimenter once all subjects had completed all tasks and the questionnaire to determine which task would count for payment. Subjects were paid in cash with average salient earnings of $15.17 per subject.

Figure 1. Choice, Rating, and Pricing Architectures under Joint Presentation



4. Results

We focus on three measures of performance for comparing architectures: (i) The average rank of the selected option under that architecture (where an average rank of 1 would indicate that all subjects chose the best option, while an average rank of 2 would indicate that on average subject chose the second best option as their preferred choice), (ii) The percentage of subjects who chose the best option under that architecture and (iii) Money left on the table which measures the difference between the probability of receiving a payment under the optimal option and the probability of receiving the payment under the chosen option (multiplying this measure by the size of the monetary payment reveals how much money was forgone by choosing suboptimally). When ties occurred that contained the best option, for (i) we computed the average rank of all tied options, for (ii) we computed the probability of choosing the best option as $1/t$ where t is the number of tied options (since each of the tied options was equally likely to be randomly assigned), and for (iii) we computed the average money left on table across all chosen options.

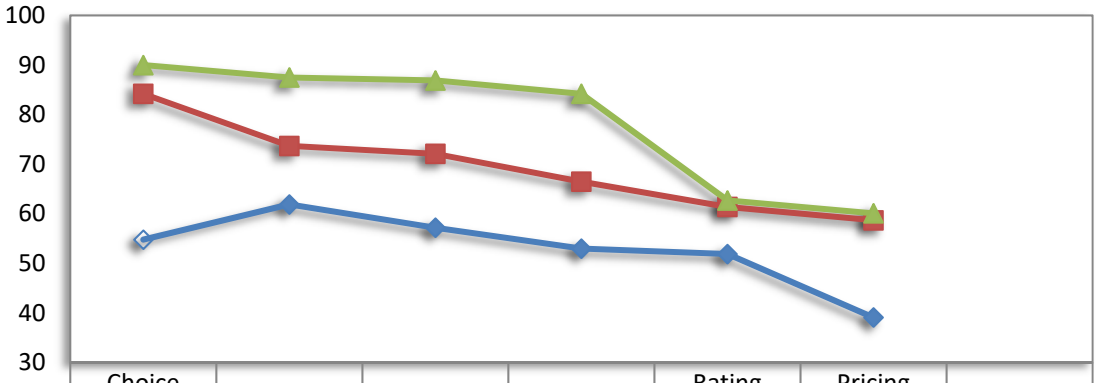
We will refer to the first metric as welfare ranking or 'efficiency'[3] and the second as optimality. We refer to a task as the efficient architecture for a group of subjects if it assigns those subjects the best average ranked option among the 16 options, relative to the other architectures. The average rank of assigned lotteries is shown for all six architectures in Table 2, with the results broken down by the CRT score of the subjects. We will refer to a task as the optimal architecture for a group of subjects if it maximizes the average probability of selecting the best option for those subjects. A similar breakdown for optimal architectures and for money left on the table is provided in Figures 2 and 3. Subjects were grouped according to how many of the CRT questions they correctly answered, and classified into roughly equal-sized categories of low CRT, medium CRT, and high CRT, based on whether they correctly answered two or fewer questions, three or four questions, or five or more questions correctly, respectively.

[3] While the 'welfare ranking' differs from the typical metric of efficiency in economics, the two are perfectly correlated in this study. This is a consequence of our design which holds the probability of every option constant while allowing for the probability of each state to be different. Thus, calculating any of the typical measures of efficiency (such as the ratio of the probability of the chosen option and highest option probability) would differ from our results by only a constant scalar.

Table 2. Efficiency: Average Rank of Assigned Option across All Six Choice Architectures

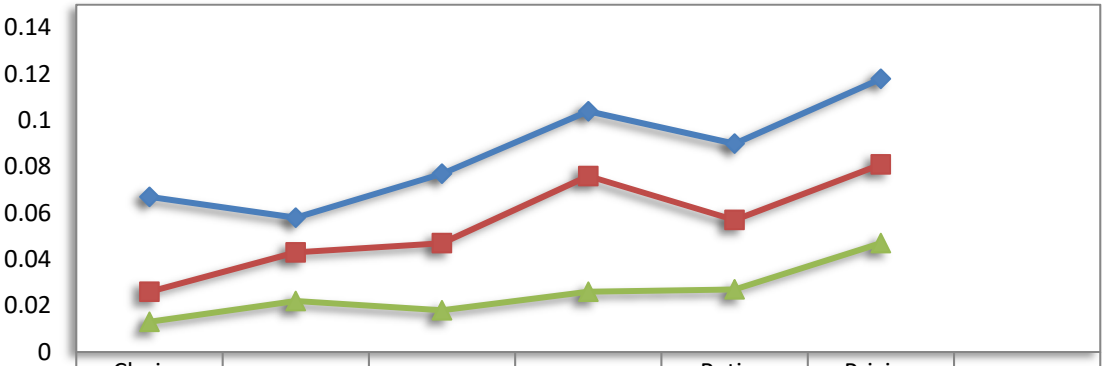
CRT group	Number of Subjects	Choice Tournament	Choice Joint	Rating Joint	Pricing Joint	Rating Separate	Pricing Separate	Average Rank
Low	42	3.190	**2.310**	3.441	3.976	3.406	4.658	3.497
Medium	38	**1.632**	2.737	2.488	3.278	2.721	3.177	2.672
High	40	**1.275**	1.575	1.475	1.934	3.399	3.521	2.196
All	120	2.058	2.200	2.484	3.074	3.187	3.810	2.802

Figure 2. Optimality: Percentage of Optimal Assignments across All Six Choice Architectures



	Choice Tournament	Choice Joint	Rating Joint	Pricing Joint	Rating Separate	Pricing Separate	Average
Low CRT	54.8	61.9	57.2	53.0	51.9	39.1	53.0
Medium CRT	84.2	73.7	72.1	66.5	61.4	58.7	69.4
High CRT	90.0	87.5	86.9	84.2	62.7	60.1	78.6
Average	75.8	74.2	71.8	67.7	58.5	52.3	66.7

Figure 3. Money Left on Table as measured by Loss in Probability of Winning across Architectures



	Choice Tournament	Choice Joint	Rating Joint	Pricing Joint	Rating Separate	Pricing Separate	Average
Low CRT	0.067	0.058	0.077	0.104	0.09	0.118	0.086
Medium CRT	0.026	0.043	0.047	0.076	0.057	0.081	0.055
High CRT	0.013	0.022	0.018	0.026	0.027	0.047	0.025
Average	0.036	0.041	0.048	0.069	0.059	0.082	0.056

4.1 Optimality of Presentation Modes

As noted in the introduction, previous literature has tested for *consistency* of preferences across response modes and presentation modes (e.g., Lichtenstein and Slovic, 1971; Slovic et al., 2007; Hsee, 1996). But to our knowledge, the *optimality* of response modes and presentation modes (whether some response modes or presentation modes systematically induce better decisions) has not been investigated.

In our experiment, the rating and pricing response modes enable us to compare joint presentation of alternatives (displaying all 16 options at once) with separate presentation (displaying only one option at a time). For a fixed response mode, we find that joint presentation performs significantly better in inducing a higher welfare ranking than separate presentation. For instance, rating all at once achieves a 0.703 better average ranking than rating separately ($p < 0.05$, two-tailed Wilcoxon signed rank test) and pricing all at once achieves a 0.736 better average ranking than pricing separately ($p < 0.10$, two-tailed Wilcoxon signed rank test). However, when computing welfare for each CRT group, this difference is only significant for the high CRT group[4] (for high CRT subjects, rating options jointly outperforms rating separately ($p < 0.002$), and pricing jointly outperforms pricing separately ($p < 0.02$), two-tailed Wilcoxon signed rank tests). Similarly, joint presentation produces more optimal choices than separate presentation mode. This finding holds for both pricing and rating response modes. In particular, for the rating response mode, joint presentation induced 71.8% optimal responses and separate presentation induced 58.5% optimal responses. This difference is highly significant across all subjects ($p < 0.02$, two-tailed Wilcoxon signed rank test). Similarly, for the pricing response mode, joint presentation induced 67.7% optimal responses and separate presentation induced 52.3% optimal responses. This difference is also significant across all subjects ($p < 0.02$ two-tailed Wilcoxon signed rank test). However, when computing the optimality for each CRT group, the advantage of joint over separate presentation is only significant for the high CRT group[5] (for high CRT

[4] For the average rank metric, for the low CRT group, the p-value for rating jointly vs. rating separately is 0.881 and the p-value for pricing jointly vs. pricing separately is 0.453. For the medium CRT group, the p-value for rating jointly vs. rating separately is 0.734 and the p-value of pricing jointly vs. pricing separately is 0.8415.

[5] For the percentage of optimal choices, for the low CRT group, the p-value for rating jointly vs. rating separately is 0.576 and the p-value for pricing jointly vs. pricing separately is 0.153. For the medium CRT group, the p-value for rating jointly vs. rating separately is 0.204 and the p-value for pricing jointly vs. pricing separately is 0.358.

subjects, rating options jointly outperforms rating separately ($p < 0.01$), and pricing jointly outperforms pricing separately ($p < 0.02$), two-tailed Wilcoxon signed rank tests). The high CRT group experiences a large loss in performance under separate presentation, while the medium and low CRT groups only perform a little worse (but not significantly so).

The Wilcoxon signed rank test results for average welfare ranking also extend to the "money left on the table" measure which has the same ordinal ranking as the average welfare measure. In terms of effect size, high CRT subjects leave roughly 50% more money on the table under separate presentation than under joint presentation for both the rating and pricing response modes. In aggregate, the worst-performing architecture under the money-left-on-the-table metric (the pricing architecture under separate presentation) leaves twice as much money on the table as the two-best performing architectures for this metric (the choice tournament and the simple choice architecture). Thus, we find that for all 120 subjects taken collectively (and for the subset of 40 high CRT subjects), joint presentation performs better than separate presentation across the three welfare criteria in Table 2 and in Figures 2 and 3. An alternative way to read this result is that all subjects, regardless of CRT group, performed similarly poorly in the separate presentation mode. In the joint presentation mode, low and medium CRT subjects do not significantly improve, while high CRT subjects are better able to evaluate the large 16-item choice set and improve significantly when they can compare all options simultaneously. Indeed, a plausible explanation for the superior performance of joint presentation is that it makes cross-option comparisons easier. With separate presentation modes, the subject must recall previous options. This is a difficult task. While strategies for doing this may exist, it is not clear subjects realize that they should or could be applying them.

4.2 Optimality of Response Modes

For a fixed presentation mode, we can also consider whether one response mode performs systematically better than the others. Fixing the presentation mode at joint presentation of alternatives we observe that choice performed significantly better than pricing in terms of efficiency (two-tailed Wilcoxon signed rank test, $p < 0.01$). Neither the difference in efficiency between choice and rating or between rating and pricing

was significant. Within CRT groups, only the differences in efficiency for the low CRT group were significant: Choice outperformed rating ($p < 0.05$, two-tailed Wilcoxon signed rank test) and choice outperformed pricing ($p < 0.01$, two-tailed Wilcoxon signed rank test). The size of the effect is also large with the low CRT group performing more than one full rank better in the choice response mode than in either the rating or pricing response modes, as can be seen in Table 2.

None of the differences in percentage of optimal assignments were significant for different response modes. However, the pattern that choice performs better than rating and that rating performs better than pricing is persistent in our data, even though these differences are usually small and not significant. For instance, for all 120 subjects, we consistently observe that under joint presentation, choice outperforms rating and rating outperforms pricing for each of the three metrics we use. The money-left-on-the-table metric reveals that, fixing the presentation mode at joint presentation, the simple choice architecture leaves less money on the table than the pricing architecture ($p < 0.01$, two-tailed Wilcoxon signed rank test).

Under separate presentation, we again find that rating outperforms pricing for all 120 subjects for each of the three metrics. Moreover, in our study of choice, rating, and pricing tournament architectures with a different group of subjects (see Section 7), we observe the same ranking across response modes for each of the three metrics. Surprisingly, in every comparison involving all subjects, for each of our performance measures, we find the ranking that choice performs better than rating which performs better than pricing.

The persistent ranking of response modes we observed is consistent with the possibility that people perform better on more constrained response modes than those that permit a larger range of possible responses for each option. For joint presentation, choice requires only one discrete response, which is more constrained than rating which permits seven possible ratings per option, which in turn is more constrained than pricing which permits any response between \$0.00 and \$20.00 per option. Indeed, for the joint presentation mode, we can rank all three response modes and observe overall behavior consistent with this ranking of 'constrained responses' for both the optimality and efficiency metrics (with choice outperforming rating and rating outperforming pricing). One reason more constrained response modes induce better performance might be that they reduce the complexity of the choice architecture. For rating

and pricing, subjects are confronted with complex options (each contingent on 12 possible states[6]), a complex choice set (16 options), and a complex response mode (seven possible ratings or many more possible valuations). The choice response mode minimizes this added layer of complexity, so subjects need only deal with the complexities of the options and the choice set.

4.3 Heterogeneity in Cognitive Reflection and Performance

From Table 2 and Figures 2 and 3, we see that the choice tournament performed best in terms of efficiency optimality, and money left on the table, validating its performance in Besedeš et al. (2015). However, this difference is not significant when comparing the choice tournament to the choice joint architecture or to the rating joint architecture. Moreover, the low CRT group generally performed best in the task of making one choice from all 16 options simultaneously, although this performance was not significantly better than the performance on the choice tournament. We also observe remarkable predictive power of the CRT in sorting out subject performance (see also, for example, Table 2 and Figures 2 and 3), regardless of the architecture. High CRT subjects achieved 78.6% optimal choices averaged across all architectures, but this reduces to 69.4% for the moderate CRT group and to 53% for the low CRT group (The difference in optimal choices between high CRT and low CRT subjects is significant ($p < 0.02$; two-tailed difference in proportions test), but the difference between either of these groups and the moderate CRT group is not). The money left on the table, as shown in Table 5, is particularly large for the low CRT group (resulting in more than an 8 percentage point loss in the probability of winning, on average across all architectures relative to choosing optimally). While in our experiment this is not a large amount of money (a little more than $1.50), an 8% loss in wealth due to poor financial investing, or due to suboptimal selection of one's healthcare plan or insurance policy could be a significant amount of money over time. In contrast, the high CRT group leaves very little on the table.

[6] We varied the complexity of response and presentation modes and kept the complexity of the options fixed. Huck and Weizsacker (1999) analyze deviations from expected value maximization by varying the complexity of lotteries.

4.4 Response Time

The experimental software recorded the response time for each subject and for each architecture. The median response times across choice architectures and CRT groups are presented in Table 3. Across all 120 subjects, we see that the joint pricing task had the longest median response time, requiring slightly more than three minutes for the median subject. In contrast, the joint choice task had the smallest median response time, requiring less than one and a half minutes for the median subject. These differences are highly statistically significant: The joint pricing architecture has a significantly longer distribution of response times than either rating architecture or the simple choice architecture (all $p < 0.001$, two-tailed Wilcoxon signed rank test). The difference in response times between the two pricing architectures and between the joint pricing task and the tournament architecture were not significant.

Table 3. Median Response Times (in seconds) across CRT Groups and Choice Architectures

CRT Group	Tournament	Choice	Rating All	Pricing All	Rating separate	Pricing separate
Low	159.77	78.96	153.75	158.70	140.68	169.02
Medium	167.79	79.05	131.02	177.19	125.09	128.72
High	185.12	78.47	149.98	206.51	139.29	185.09
All subjects	172.33	78.96	149.98	180.29	135.19	160.60

The simple choice architecture had significantly faster response times than any of the other architectures (all $p < 0.001$, two-tailed Wilcoxon signed rank test). It is not surprising that the choice architecture was fastest given that it required only a single response, whereas the other architectures required multiple ratings, prices, or choices. However, it is surprising how well the choice architecture performs given it is more efficient on the time dimension.

One might evaluate choice architectures on multiple dimensions such as the average rank assigned by the architecture (efficiency in terms of outcomes) and the time required by the architecture (efficiency in terms of time). For architectures evaluated on these two dimensions, we propose that a choice architecture A, *strictly dominates* another architecture B, if A provides a more efficient welfare ranking in a shorter amount of time. Comparing the median response times across all subjects in Table 3 and the average rank of assigned options in Table 2, we see that the simple choice architecture strictly dominates all rating and

pricing architectures (it assigns both a lower average rank and requires a shorter median response time). Moreover, when accounting for the time dimension, the simple choice architecture is not dominated by any other architecture since the only architecture with a lower average rank (the choice tournament architecture) had more than twice as long a median response time as the simple choice architecture.

4.5 Pricing Strategies

While the previous section discussed heterogeneity in performance due to cognitive reflection, this section considers heterogeneity in strategies (e.g., for pricing options in the choice set) due to cognitive reflection. Figures 4, 5, and 6, display the strategies each subject used for the low, medium, and high CRT groups, respectively, in pricing all 16 options in both the all-at-once task (solid red lines) and the one-at-a-time task (dotted blue lines). Each figure displays the prices (between $0.00 and $20.00) provided by each subject for each option sorted in descending order from best to worst (A through P). For the pricing all-at-once task, a clear pattern that emerges with some regularity is a strategy that prices the best option higher than all of the others and assigns all other options the same price. This strategy may conserve cognitive resources as once the best option is identified, there is no need to deliberate how to assign prices to each of the inferior options. We refer to this strategy as an "L" strategy since it visually appears in the shape of the letter L. From Figure 4, we see that three out of 42 low CRT subjects used a perfect L strategy. Three of 38 medium CRT subjects used a perfect L strategy as can be seen from Figure 5, while eight of 40 high CRT subjects used a perfect L strategy in Figure 6. There are also several additional subjects who used a strategy that could be termed a "noisy L" in which there is some oscillation in the horizontal portion of the L. For the pricing one-at-a-time task, an L strategy is unlikely to appear since prices are provided sequentially and a subject could not know if a better option was forthcoming. Indeed, the pattern that emerges from the pricing one-at-a-time task is one that more closely mimics that which would arise from pricing each option at its expected value, although for some subjects the pricing-one-at-a-time strategy closely parallels their pricing-all-at-once strategy. We show in Section 8 that, as suggested by Figures 4, 5, and 6, correlations between prices and expected values are higher for the 'price one-at-a-time' task than for the 'price all-at-once' task in our primary study.

Figure 4. Pricing Strategies in Joint (Solid) and Separate (Dotted) tasks for Low CRT Subjects (Option (A – P) on Horizontal Axis; Price ($0 - $20) on Vertical Axis)

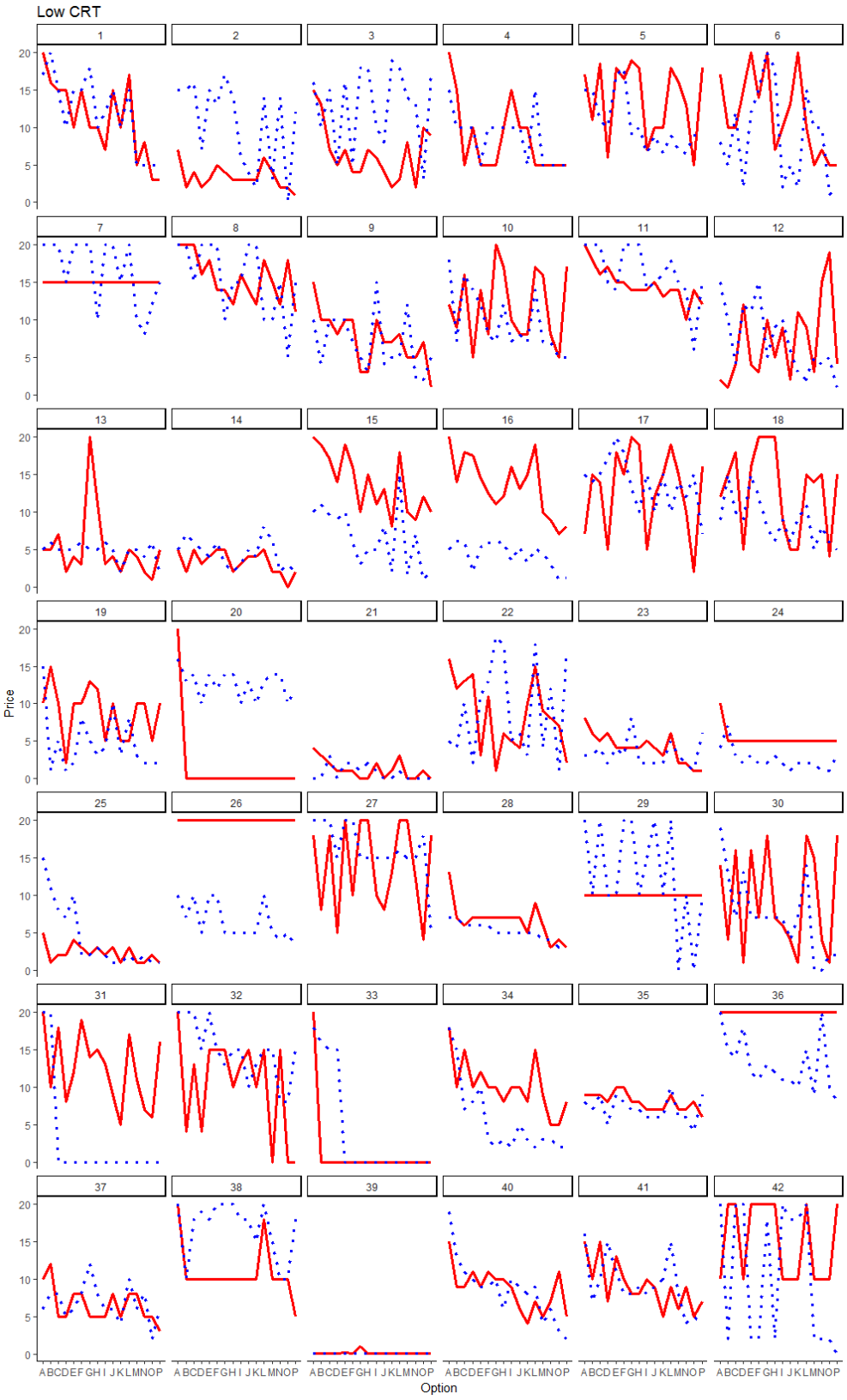


Figure 5. Pricing Strategies in Joint (Solid) and Separate (Dotted) tasks for Medium CRT Subjects (Option (A – P) on Horizontal Axis; Price ($0 - $20) on Vertical Axis)

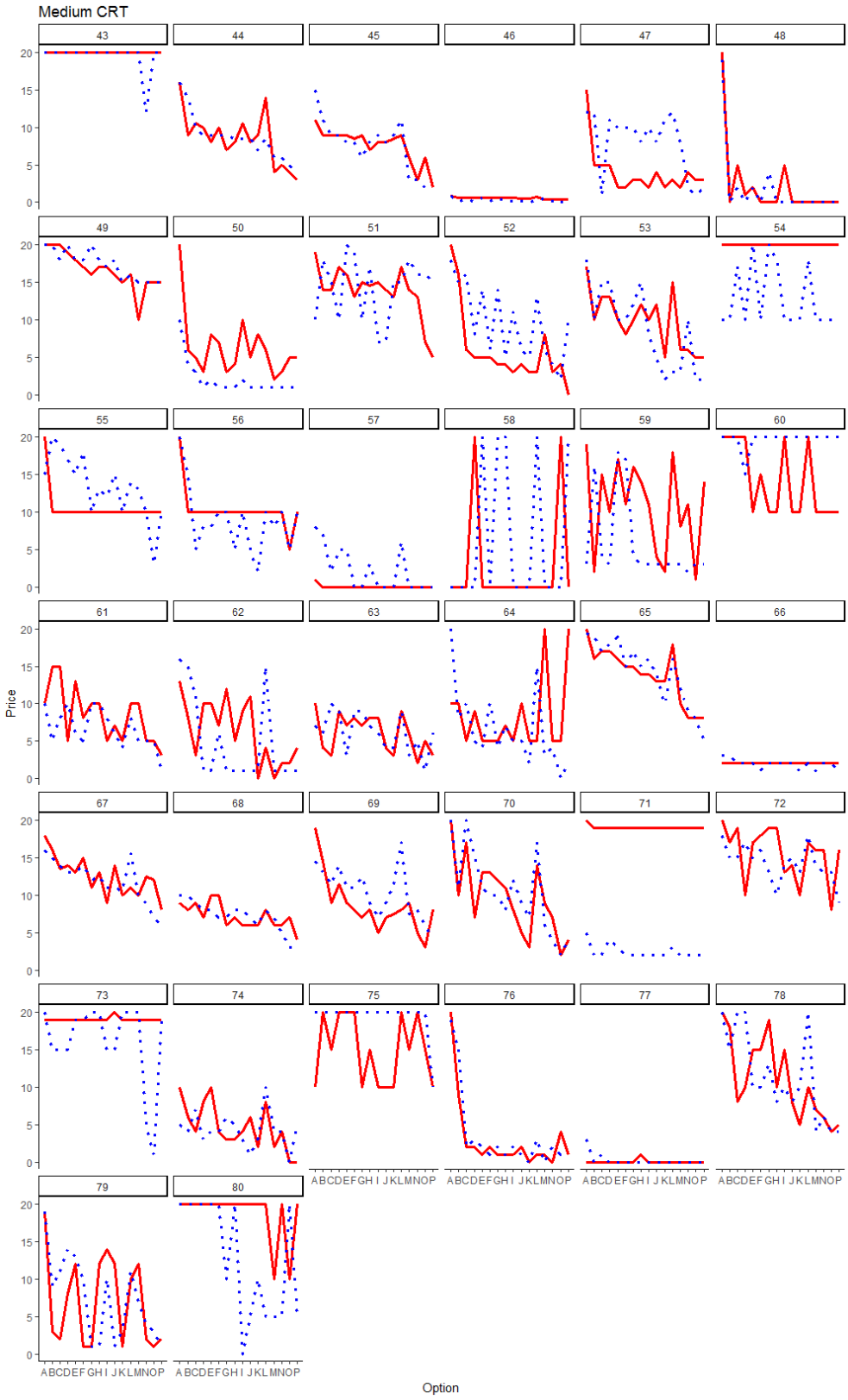
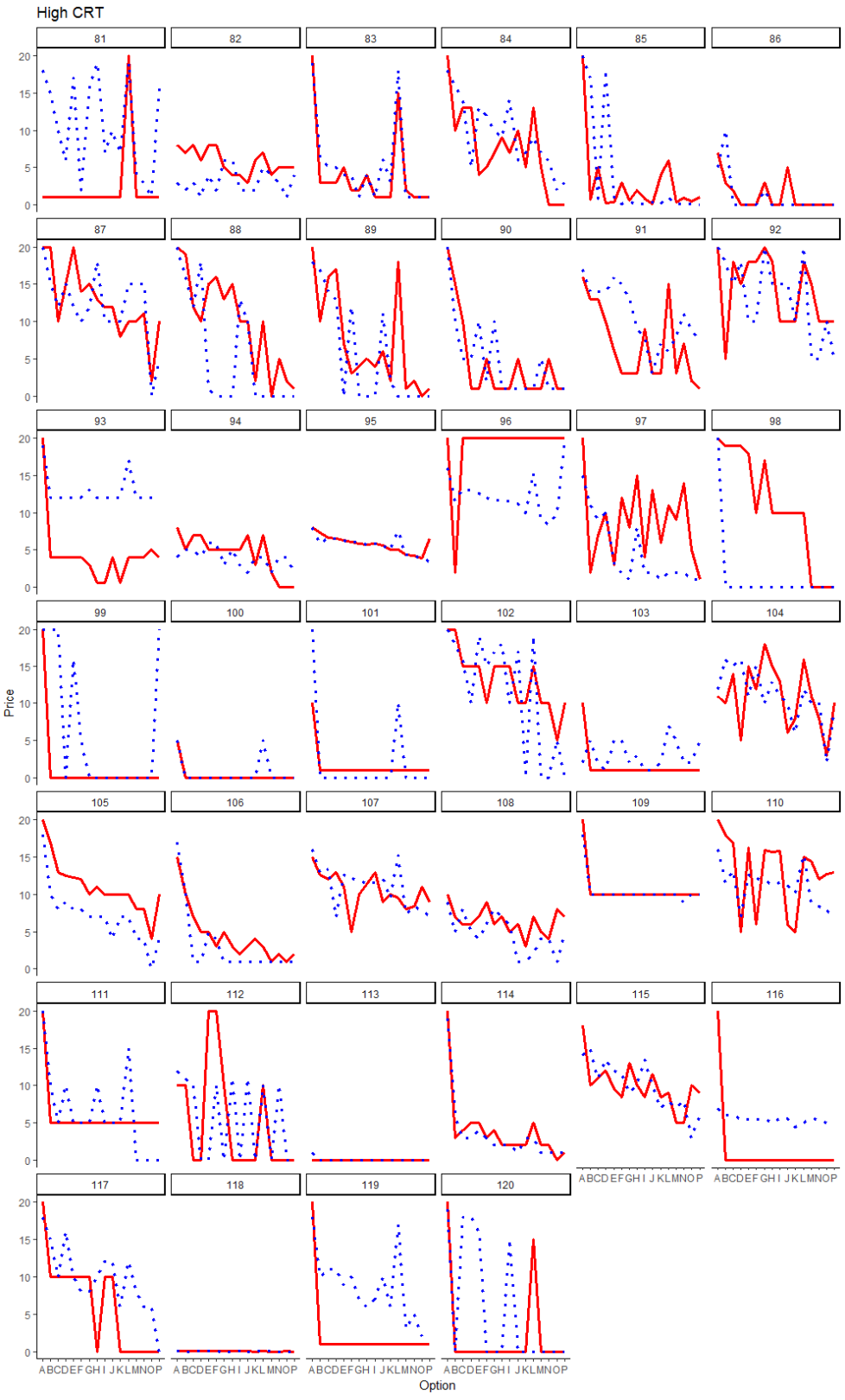


Figure 6. Pricing Strategies in Joint (Solid) and Separate (Dotted) tasks for High CRT Subjects (Option (A – P) on Horizontal Axis; Price ($0 - $20) on Vertical Axis)



4.6 Happiness Rating Strategies

While the previous section discussed heterogeneity in pricing strategies, this section considers heterogeneity in rating strategies (e.g., for rating options in the choice set) due to cognitive reflection. Figures 7, 8, and 9, display the strategies each subject used for the low, medium, and high CRT groups, respectively, in rating all 16 options in both the all-at-once task (solid red lines) and the one-at-a-time task (dotted blue lines). Each figure displays the ratings (between 1 and 7) provided by each subject for each option which are ordered in a descending order from best to worst (A through P). For the rating all-at-once task, the L strategy and 'noisy' L strategies again emerge with some regularity, in which a subject rates the best option higher than all of the others and assigns all other options the same (or approximately the same) rating. As noted, this strategy may conserve cognitive resources as once the best option is identified, there is no need to deliberate how to assign ratings to each of the inferior options. That such L strategies emerge in both the pricing and rating (all-at-once) tasks may suggest that a common decision process guided behavior in both tasks, despite the difference in response modes. As before, L-strategies and noisy L strategies appear concentrated among high CRT subjects. From Figure 7, we see that two of 42 low CRT subjects used a perfect L strategy, none of 38 medium CRT subjects used a perfect L strategy as can be seen from Figure 8, while five of 40 high CRT subjects used a perfect L strategy in Figure 9. In addition, many other high CRT subjects used a noisy L strategy. For the rating one-at-a-time task, the pattern that emerges more closely mimics that which would arise from assigning ratings that were monotonically increasing in the expected value of an option.

Figure 7. Rating Strategies in Joint (Solid) and Separate (Dotted) tasks for Low CRT Subjects (Option (A – P) on Horizontal Axis; Rating (1 - 7) on Vertical Axis)

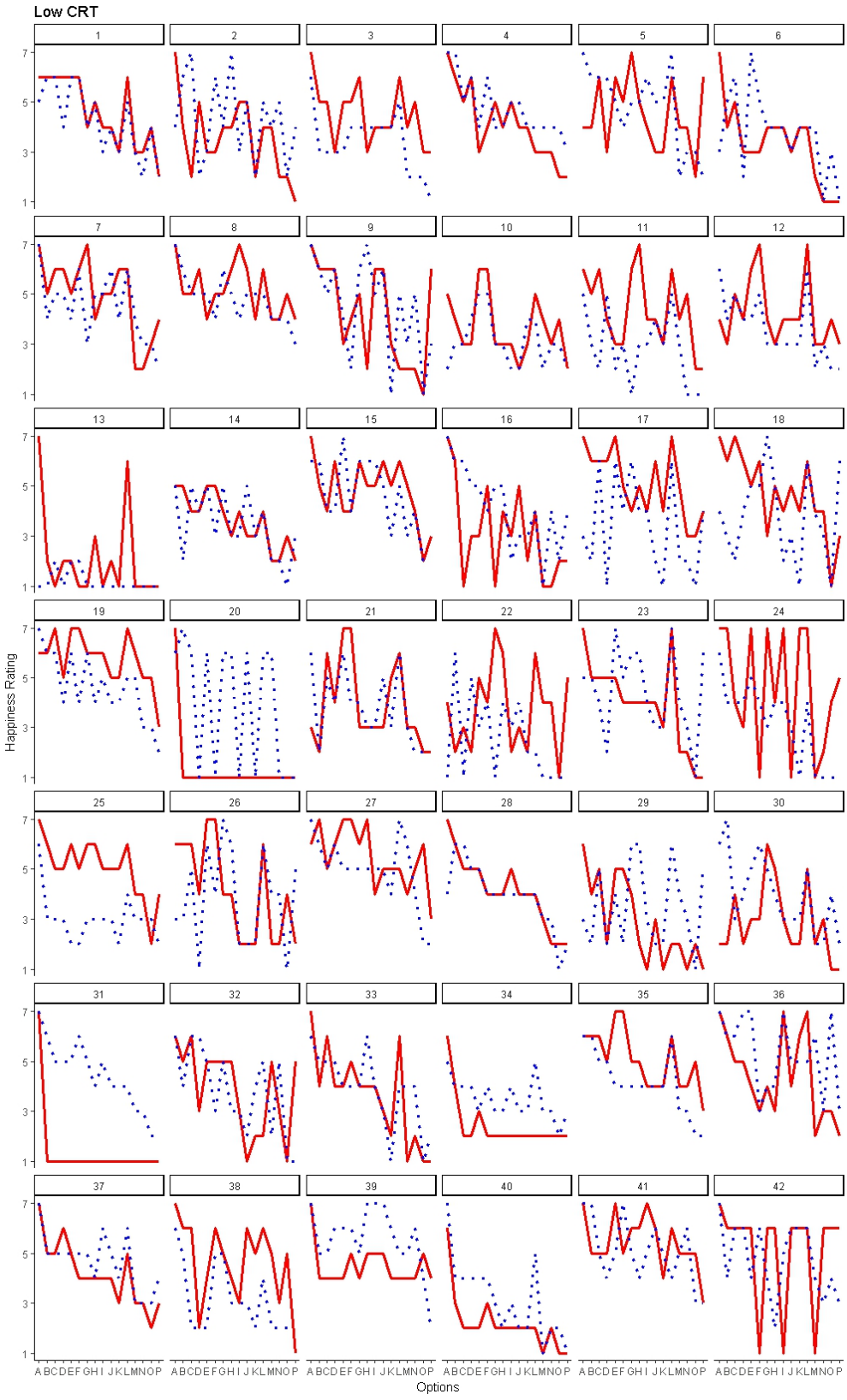


Figure 8. Rating Strategies in Joint (Solid) and Separate (Dotted) tasks for Medium CRT Subjects (Option (A – P) on Horizontal Axis; Rating (1 - 7) on Vertical Axis)

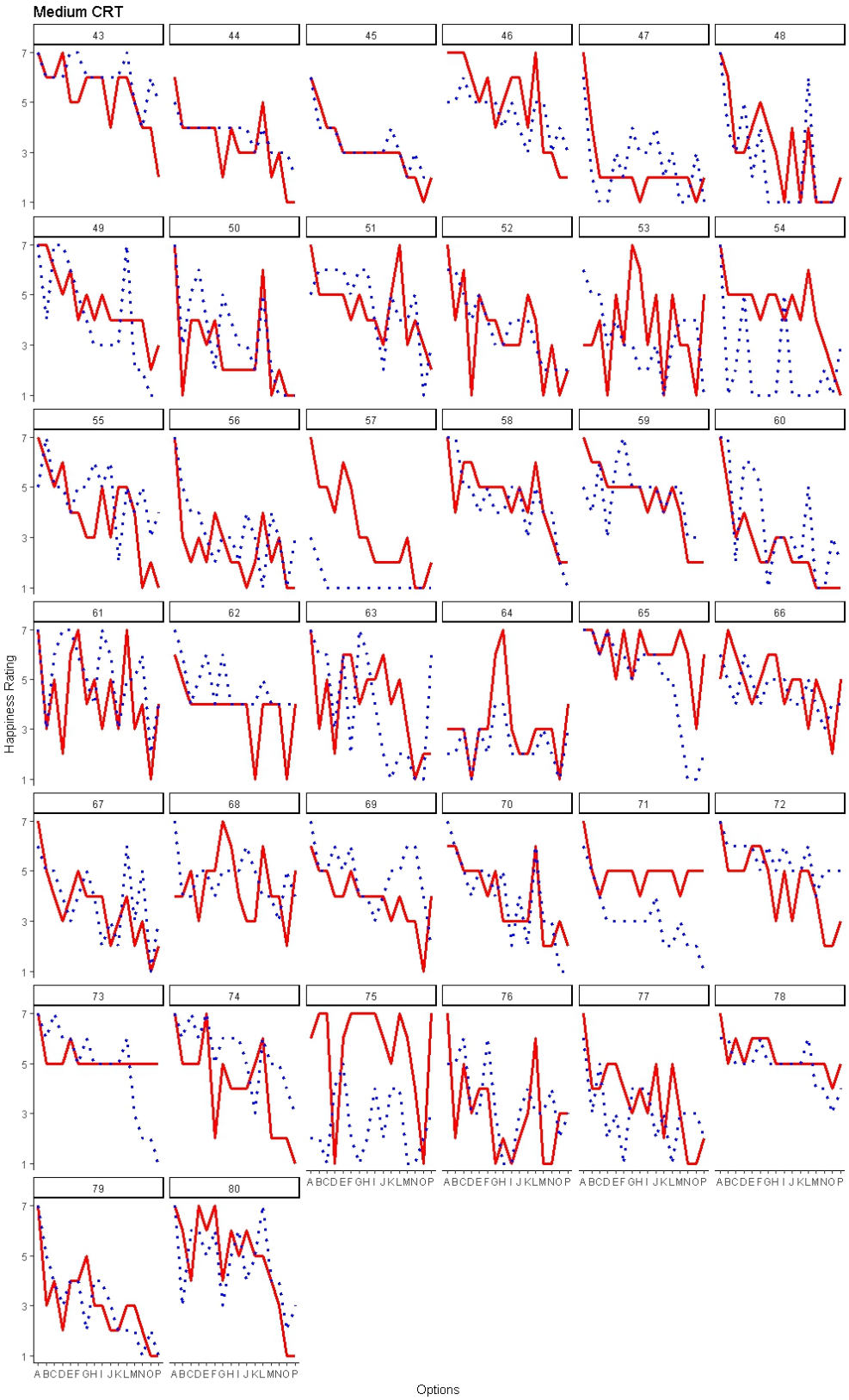
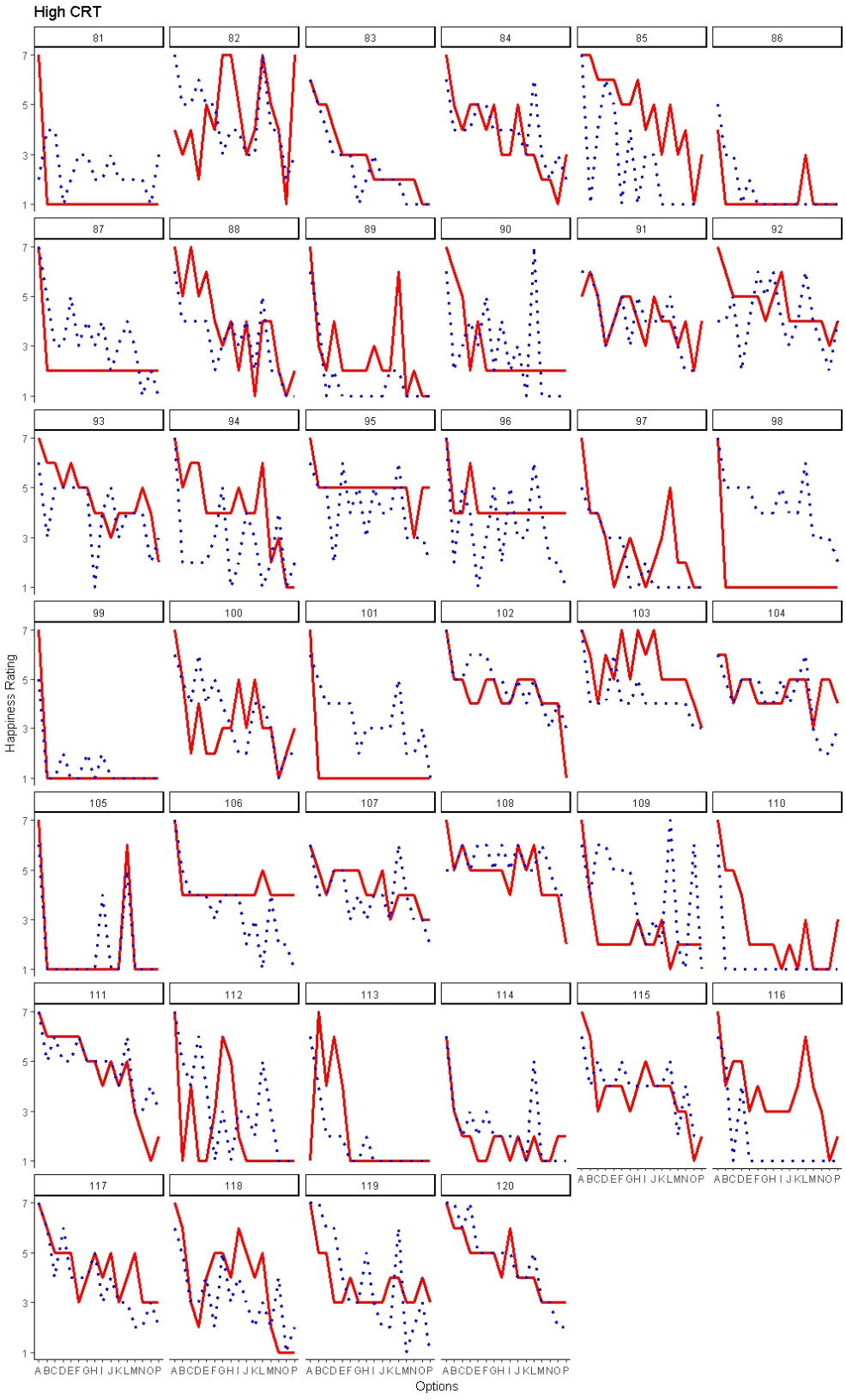


Figure 9. Rating Strategies in Joint (Solid) and Separate (Dotted) tasks for High CRT Subjects (Option (A – P) on Horizontal Axis; Rating (1 - 7) on Vertical Axis)



5. A Simple Error Rate Model of Decision Quality

We next consider in more depth the optimal (and efficient) architectures for low, moderate, and high CRT groups. For the low CRT group, the simple choice architecture has the highest percentage of optimal assignments, the lowest average ranking (highest efficiency), and the smallest amount of money left on the table among all architectures. (However, the differences in performance relative to the tournament architecture are not statistically significant).[7] In terms of efficiency, the simple choice architecture was the only architecture for the low CRT group to assign better than the third best of the 16 options, on average. Moreover, it assigned nearly a full rank better than the tournament architecture which had the second-best efficiency ranking for the low CRT group. In particular, as shown in Table 2, simultaneous choice produced an average rank of 2.31 out of 16 options (assigning low CRT subjects nearly the second-best option overall, on average). In contrast, the tournament produced an average rank of 3.19, suggesting that choosing among all 16 options at once was better for the low CRT group. In terms of optimality, choosing from all 16 options at once was the only architecture to produce over 60% optimal assignments for the low CRT group, approximately a 7 percentage point improvement over the tournament architecture. We consider the implications of these differences in performance for the low CRT group to illustrate how complexity and cognitive reflection might lead different architectures to be optimal for different populations, with the caveat that the observed differences between the two choice-based architectures are not significant.

As can be seen from Table 2 and Figures 2 and 3, the medium and high CRT groups both performed best on the tournament architecture. In terms of optimality, tournament offered roughly a 10 percentage point improvement over the simple choice architecture for the moderate CRT group, but produced less than a 3 percentage point improvement over that architecture for the high CRT group. In terms of efficiency, the tournament architecture achieved nearly a full rank better, on average, than the next best architecture for

[7] In contrast, the tournament architecture performs significantly better than the simple choice architecture for the medium CRT group in terms of efficiency ($p < 0.05$, two-tailed Wilcoxon signed rank test), but not in terms of the percentage of optimal choices. At the aggregate level, the performance of the tournament architecture is not significantly different from the performance of the simple choice architecture.

the moderate CRT group (two-tailed Wilcoxon signed rank test, p = 0.048), but produced only a 0.20 better rank, on average, compared to the next best architecture for the high CRT group.

What can explain the higher percentage of optimal assignments and higher efficiency for the simple choice architecture for the low CRT group and the better performance of the tournament architecture for the higher CRT groups? If the true effect is not entirely due to chance, the better performance of the simple architecture for the low CRT group may seem particularly puzzling. Both the simple choice architecture and the tournament involve the choice response mode, but the tournament presents only four options at once, whereas the simple choice architecture presents all 16 options. Surely if one can choose the best option from a set of 16, that person could choose the same best option from a set of 4. Since we have fixed the response mode, should we not expect better performance on the tournament architecture since it presents the decision maker with fewer options to choose from at each stage?

To address these questions we construct a simple error rate model as an interpretive framework for understanding the differences in performance between the simple choice and tournament architectures for the low, moderate, and high CRT groups. To begin, consider a world with three types of agents – those with low, moderate, and high degrees of cognitive reflection (labeled types $l, m,$ and h, respectively). Suppose that for each choice, agents of type $\theta \in \{l, m, h\}$ have an error rate of $\epsilon_\theta(n)$, for choice sets of size n. It seems natural to make the following two predictions:

(i) $\epsilon_l(n) > \epsilon_m(n) > \epsilon_h(n)$ for any n.

(ii) $\epsilon_\theta(n) > \epsilon_\theta(k)$ for all $n > k$, and for all $\theta \in \{l, m, h\}$.

Prediction (i) is a monotonicity condition on agents. It predicts that agents with higher cognitive reflection have smaller error rates. Prediction (ii) is a separate monotonicity condition on the size of the choice set. It predicts that for a given level of cognitive reflection, larger choice sets induce larger error rates than smaller choice sets. This latter condition may be further augmented to control for the quality of the best option relative to the alternatives (admitting the possibility that some choices are inherently easy and others are inherently difficult), but given that the same options were used in the large and small choice

sets for the tournament and simple choice architectures, this seems unlikely to be an important dimension for our experiment.

We take the empirically observed percentages of optimal choices in the simple choice architecture as providing empirical estimates of the error rates from a 16-option choice set (i.e., as empirical estimates of $\epsilon_\theta(16)$). That is, for the data on the simple choice architecture from Figure 2, we have $1-\epsilon_l(16) = 0.619; 1-\epsilon_m(16) = 0.737; 1-\epsilon_h(16) = 0.875$. Next, we seek to estimate error rates for a 4-item choice set. Note that error rates for a four-item choice set cannot be determined in the same way as for the 16-item choice set, since in the tournament architecture, each subject made five discrete responses (one for each of the initial four-option choice sets and one 'final four' round). Performance in the three of the four choice sets not containing the optimal option is irrelevant to the selection of the optimal alternative. Subjects of type θ thus have error probability $\epsilon_\theta(4)$ in the first round of the tournament, and if they select the optimal option in the round in which it initially appears, they face a second choice among the final four options, again with error rate $\epsilon_\theta(4)$ (assuming a constant error rate for a given cognitive type and given choice set size for simplicity). The overall probability of choosing optimally in the tournament architecture for a subject of type θ is then $[1-\epsilon_\theta(4)]^2$.

Taking the empirically observed percentages of optimal assignments in the tournament architecture as providing empirical estimates of the quantity $[1-\epsilon_\theta(4)]^2$, we have $[1-\epsilon_l(4)]^2 = 0.548$; $[1-\epsilon_m(4)]^2 = 0.842$; $[1-\epsilon_h(4)]^2 = 0.900$. Solving for $\epsilon_\theta(4)$, we can identify the unique error rates for each type which fits the observed data exactly[8]. Note that this simple model would not fit the observed data exactly if $\epsilon_\theta(4) > \epsilon_\theta(16)$, or if it is not the case that $\epsilon_l > \epsilon_m > \epsilon_h$ for either the four-option or sixteen-option choice set. That is, conditions (i) and (ii) are necessary for this simple error rate model to fit the observed data exactly. The implied error rates are displayed in Table 4, from which it is clear that conditions (i) and (ii) both hold.

[8] Alternatively, we could observe the error rates for four-option choice sets containing the optimal option directly. Doing so yields $(\epsilon_l(4), \epsilon_m(4), \epsilon_h(4)) = (0.260, 0.069, 0.051)$ which are close to the values in Table 3. That these error rates are not identical to those in Table 3 suggests that error rates between rounds may not be independent.

Table 4. Error Rates Inferred from Data

Type	$\epsilon_\theta(4)$	$\epsilon_\theta(16)$
$\theta = l$	0.260	0.381
$\theta = m$	0.082	0.263
$\theta = h$	0.051	0.125

Using this simple error rate model to interpret the optimal architectures for agents who differ in cognitive reflection, we see that low CRT subjects perform best on the simple choice architecture because it provides the smallest number of opportunities for error (it requires just one discrete choice response). The effect of choice overload, which may be quantified by the difference $\epsilon_l(16) - \epsilon_l(4) = 0.121$ (a 12 percentage point increase in error probability for low CRT subjects when moving from the four-option to the sixteen-option choice set) is weaker than the decline in performance due to making multiple choices which are prone to error. This latter effect can be quantified by the difference $1 - \epsilon_l(4) - [1 - \epsilon_l(4)]^2 = 0.192$ (a 19 percentage point reduction in success probability due to the two-stage tournament procedure). Since implied error rates in a four-option choice set are considerably lower for moderate and high CRT subjects, choice overload has a comparatively greater effect, making the tournament architecture superior to the simple choice architecture for these subjects.

6. Identifying Optimal Choice Architectures

We briefly consider how one might identify the optimal architecture in our experiment, given the heterogeneity in performance by CRT scores. Viewing the three different CRT groups as representing different populations, we consider three welfare criteria to see if they provide a consensus on which architecture should be used by a social planner or policy maker who wants to benefit society.

The strongest welfare criterion is Pareto efficiency, which would advocate an architecture that makes at least some CRT groups better off without making any group worse off. As we can see from Table 2 and Figures 2 and 3, none of the architectures is Pareto efficient. The tournament architecture maximizes the welfare (in terms of both optimality, average rank assigned, and smallest amount of money left on the table) for the medium and high CRT groups, but does so at the expense of the low CRT group whose

chances of choosing the best option are reduced by approximately 7 percentage points and whose average rank of assigned option worsens by nearly a full rank, relative to the simple choice architecture.

The maximin criterion proposed by political philosopher John Rawls (1971) advocates maximizing the welfare of the population in society that is least well-off. In our experiment, this criterion would recommend the architecture in which subjects make one direct choice from the large choice set as it improves the welfare of the low CRT group who have the lowest welfare in terms of probability of optimality, average rank of assigned option, and money left on the table, relative to the other CRT groups.

The utilitarian criterion of Harsanyi (1955) would recommend choosing the architecture that maximizes the average social welfare. In many cases it is difficult to clearly infer which policy is optimal in this regard since it requires knowledge of the utilities of all members in society which are difficult to observe. However, in our experiment, since we can rank all options by stochastic dominance, Harsanyi's criterion makes the unambiguous recommendation for the tournament-style architecture as it optimizes both the average probability of choosing the optimal option and the average rank assigned, when taking into account all participants in the experiment.

The Kaldor-Hicks criterion (Hicks, 1939; Kaldor, 1939) views an outcome to be efficient if the group that benefits could in principle compensate the group that is made worse off to produce a Pareto improvement, even if such compensation does not actually occur. If welfare was evaluated based on money left on the table, then the tournament architecture is Kaldor-Hicks efficient since the medium CRT group and the high CRT group could each compensate the low CRT group such that everyone is better off under the tournament architecture. For instance, the medium CRT group could pay one percentage point in the probability of winning to the low CRT group in which case money left on the table would be 0.057 for the low CRT group, 0.036 for the medium CRT group, and 0.013 for the high CRT group. In contrast, the low CRT group cannot compensate the others to make the simple choice architecture Kaldor-Hicks efficient.

Finally, note that if one is able to engage in 'architecture differentiation' by providing different architectures to the different CRT groups, the low CRT group would be assigned the simple choice architecture, and the medium and high CRT groups would both be assigned the tournament architecture.

This assignment holds regardless of whether we seek an architecture that maximizes the probability of choosing the optimal option, or that maximizes efficiency, or that minimizes the amount of money left on the table. Under this assignment of architectures, the average rank of the option assigned across all members of society (or at least across all subjects in our experiment) would improve by 15% from 2.05 to 1.75.

7. Study with Choice, Rating, and Pricing Tournaments

In our primary study described in the preceding sections, a 'tournament' style architecture was conducted only for the choice response mode. This was done for two reasons. First, the main objective of that study was to focus on ways different response modes are most often presented in practice. The choice tournament was included due to its previous success in lab experiments as a means to measure the success of other architectures should they have outperformed choice all – at – once. Second, we worried about fatigue or boredom impacting subjects if they were asked to do too many tasks. However, given the success of the tournament presentation mode with the choice response mode, it is worth directly investigating the performance of other response modes in a tournament structure. In this section we report the results of an additional study we ran for this purpose. In particular, we conducted tournaments for choice, rating, and pricing response modes.

7.1 Experimental Design for Choice, Rating, and Pricing Tournaments

Using a within-subjects design, we conducted a study with three choice architectures – a choice tournament, identical to the one used in the primary study described above, as well as a 'happiness rating' tournament and a 'pricing' tournament architecture. In all three tournament architectures, subjects were presented with four options (out of sixteen) on each screen, and they were asked to either choose one of the options or rate each option on a happiness scale or assign a price to each option. The option selected or assigned the highest rating or the highest price for each four-option set was sent to a 'final four' round, where subjects were then asked to choose or rate or price each of the options in the final four round. Ties (which could occur in rating and pricing tournaments but not in the choice tournament) were broken randomly.

The sixteen options in this study were the same as those used in the primary experiment, but only the first three PDF's (from Table 1) were used in this study. The payment protocol was the same as in the

primary experiment with subjects receiving either $0 or $20 depending on how much they earned in the one task that was randomly selected for payment. After all three tasks were completed, subjects completed a survey, received payment in private, and were dismissed from the study. Sixty new subjects were recruited for this study.

7.2 Results for Choice, Rating, and Pricing Tournaments

Summary statistics for the study of choice, rating, and pricing tournaments are given in Table 5. In addition to the different CRT distribution relative to our primary study, this follow-up study had slightly younger subjects (average age of 19.2 years in the primary study and 18.6 years in this tournament study), and fewer male subjects (there were 61 males of 120 subjects in the primary study and 19 males of 60 subjects in this tournament study).

Table 5. Results from Choice, Rating, and Pricing Tournaments

	Average Rank of Assigned Option			
CRT Group	Number of subjects	Select Optimal	Rate Optimal	Price Optimal
Low	32	3.344	4.625	3.969
Medium	12	3.333	3.333	3.583
High	16	2.563	2.813	4.500
Overall	60	3.133	3.883	4.033
	Percentage of Optimal Assignments			
CRT Group	Number of subjects	Select Optimal	Rate Optimal	Price Optimal
Low	32	0.469	0.375	0.375
Medium	12	0.583	0.583	0.500
High	16	0.813	0.750	0.375
Overall	60	0.583	0.517	0.400
	Money Left on the Table			
CRT Group	Number of subjects	Select Optimal	Rate Optimal	Price Optimal
Low	32	0.094	0.138	0.112
Medium	12	0.085	0.085	0.098
High	16	0.061	0.072	0.129
Overall	60	0.083	0.110	0.114

We make two observations: First, for the tournaments, choice performs better than rating which performs better than pricing. This is the same ranking we observed across all three response modes in the all-at-once

tasks from the primary study and it is also the ranking we observed between rating and pricing in the one-at-a-time tasks from the primary study. Thus, although the differences are not always large, we consistently observe that the choice response mode performs better than ratings which performs better than prices. As noted in Section 4.2, this might be due to choice response modes being simpler in that there are fewer possible responses that the decision maker can provide, whereas rating and pricing response modes have a larger 'message space.' Second, we find that high CRT subjects do much better than low CRT subjects in the choice and rating tournaments, but not in the pricing tournament, although with only 16 high CRT subjects in this study it is difficult to draw strong conclusions. While these trends replicate our qualitative findings from the primary study, we also observe that performance was lower in the choice tournament in this new study, with only 58.3% of subjects selecting the optimal option. Part of this drop was due to the larger proportion of low CRT subjects who comprised approximately one third of the primary study but constituted over half of the subjects in this follow-up study. In addition, we observe little variation in the amount of 'money left on the table' across tournament architectures. The higher amount of money left on the table for the choice tournament, relative to our primary study, may also partly reflect the differences in the demographics of our subjects between studies as noted above.

8. Incentivized Pricing Study

The numerical values of the price responses discussed in the previous sections of this paper were not incentivized. That is, although the subject had to bear the consequences of the option selected based upon the subject's stated prices, the subject did not actually have to pay the stated price. That is, those prices gave an ordinal but not a cardinal ranking. It is plausible that incentivized pricing may lead people to think more earnestly about how much a given option is worth and thus lead to optimal decision making. In this section we describe an additional study that was conducted to examine this issue directly.

8.1 Experimental Design for Incentivized Pricing Study

This study used a within-subject design that involved two presentation modes: one-at-a-time and all-at-once. In both cases, the subject is required to enter 16 prices that each represent the maximum amount the subject is willing to pay for one of the respective 16 options. The two tasks were presented in random

order and the two pdfs that were observed were drawn randomly without replacement from the 6 pdfs described in Table 1 for the primary study. The option labels and card numbers were randomized for each subject as in the other studies discussed above. After both tasks were completed, one task was randomly selected to determine the subject's actual payoff. The subject then completed the survey, received payment in private and was dismissed from the study. Sixty new subjects were recruited for this study.

After a subject enters a price for each of the 16 options in the task, a random price was drawn independently for each option. The computer automatically determined which option yielded the greatest revealed surplus to the subject and this is the option that was selected for the subject. The subject was required to pay the random price associated with the selected option, shuffled the deck of cards, and then drew a card from the deck to determine if the subject earned the additional payment. Formally, let WTP_i and P_i denote a subject's stated willingness to pay and the random price for Option i. Option i^* was selected for the subject where $i^* \equiv \underset{i}{\text{argmax}}(WTP_i - P_i)$. This procedure is such that it is incentive compatible for subjects to truthfully reveal their willingness to pay for each option, but it also means that subjects are unlikely to end up actually purchasing the option that they indicated is their most preferred (i.e. the one they priced the highest).

Because a subject had to be able to actually pay for the selected option, for each task a subject was given an endowment. To keep the total stakes comparable to those in the other tasks presented in this paper, the additional payment from an option was reduced to \$10. Thus, the prices were randomly drawn from the uniform distribution from \$0.00 to \$10.00. For this reason the endowment was set to \$10. Hence, the earnings of a subject in a given tasks were \$10 - P_{i^*} if the subject failed to select a card contained in the selected option and \$10 - P_{i^*} + \$10 if the subject selected a card that was contained in the selected option.

8.2 Results for Incentivized Pricing Study

Due to the smaller sample sizes for each CRT group relative to our baseline study (with no CRT group having more than 30 subjects), we refrain from attempting to identify statistically significant differences for each group in this section. Across all subjects, we find that the differences in average rank

and in the percentage of optimal choices induced by the incentivized pricing task are not significantly different for the joint and separate presentation modes.

Relative to the primary experiment, we do find that the elicited prices in the incentivized pricing study were more highly correlated with the expected values of the lotteries. Table 6 displays the median correlation coefficient between the pricing tasks and the expected values of the lotteries as well as between the two pricing tasks for each CRT group and across all subjects. This data is provided for both the primary study and the incentivized pricing study. For the primary study, the median correlation coefficient for each CRT group is higher in the pricing one-at-a-time task than in pricing the options all at once, indicating that subjects were more likely to price lotteries in a ranking consistent with their expected values when evaluated in isolation. In addition, the correlation within subjects between pricing tasks is higher for the high CRT group (0.556) than for the medium CRT group (0.480) and the low CRT group (0.262) in the primary study, suggesting that subjects with higher CRT scores were more consistent in their pricing strategies across the one-at-a-time and all-at-once tasks.

For the incentivized pricing study, in the pricing 'one-at-a-time' task, the median correlation coefficient for correlating the expected values of the lotteries to subjects' elicited prices increases from 0.664 for the low CRT group to 0.914 for the high CRT group. For the pricing 'all-at-once' task, the median correlation coefficient increases from 0.568 for the low CRT group to 0.860 for the high CRT group. These results are shown in Table 6 along with the median correlation coefficient for elicited prices between tasks. Table 6 also reveals that the low CRT group was largely unaffected by incentives with similar correlations in the primary study and the incentivized pricing study described in this section. The incentivized pricing study did however produce more consistent rankings for the low CRT group with correlations within subjects and between tasks increasing from 0.262 in the primary study to 0.473 in the incentivized pricing study. Incentives also increased the correlation between elicited prices and expected values for both the medium and high CRT groups. From Table 6, we also see that for the incentivized pricing study, the median correlation coefficient for correlating the joint and separate pricing tasks is much higher for the high CRT group than for the other groups, which was also observed in the primary study.

In addition to the different CRT distribution relative to our primary study, this follow-up study had slightly younger subjects (average age of 19.2 years in the primary study and 18.4 years in this pricing study), and fewer male subjects (there were 61 males of 120 subjects in the primary study and 25 males of 60 subjects in this pricing study).

Table 6. Median Correlations between Pricing Tasks and Expected Values

Median Correlations between Pricing Tasks and Expected Values in Primary Study

CRT Group	Number of Subjects	EV vs. Pricing One at a Time	EV vs. Pricing All at Once	Pricing One a Time vs. Pricing All at Once
Low	38	0.625	0.515	0.262
Medium	35	0.640	0.622	0.480
High	40	0.630	0.517	0.556
Overall	113	0.627	0.563	0.401

Median Correlations between Pricing Tasks and Expected Values in Incentivized Study

CRT Group	Number of Subjects	EV vs. Pricing One at a Time	EV vs. Pricing All at Once	Pricing One a Time vs. Pricing All at Once
Low	30	0.664	0.568	0.473
Medium	22	0.771	0.740	0.441
High	5	0.914	0.860	0.872
Overall[9]	57	0.750	0.642	0.506

Metrics of performance on the incentivized pricing task are shown in Table 7. The option that is actually selected in this study depends on 16 random prices and is not actually informative with regards to the quality of decision making. What one really cares about is the option with the highest stated willingness to pay. For ease of exposition and to facilitate comparison to the primary study, we refer to the option with the highest stated price as the assigned option.[10]

Between subjects, we find that incentives do not affect the percentage of optimal assignments for the low CRT group (0.391 from the primary study vs. 0.392 with all options incentivized) for pricing one

[9] Seven subjects in the primary study and one subject in the incentivized pricing study had correlation coefficients that were undefined (they assigned the same price to all options) and so are not included in the statistics in Table 6. Two subjects in one session did not enter their ID number into the survey and so we could not link their CRT scores to their pricing data. The software for conducting this study differed from the software for the other two which automated the link between the task and the survey.

[10] In expectation, the option with the highest stated price is the most likely to be assigned since prices are drawn from the same distribution for each option.

at a time and (0.530 from the primary study vs. 0.507 with all options incentivized) for pricing all at once. The average rank for pricing one at a time for the low CRT group also is very similar to that from the primary experiment (4.658 with optimal option incentivized vs. 4.675 with all options incentivized). Incentives appear to improve the average rank for the low CRT group when considering all options at once (3.976 with optimal option incentivized vs. 2.926 with all options incentivized).

For the medium CRT groups, the percentage of optimal assignments in the all-at-once pricing task was higher in the primary study (0.665) than in the incentivized study (0.411) and the average rank was also better in the primary study (3.278) than in the incentivized study (3.582). However, the percentage of optimal assignments in the one-at-a-time pricing task was higher in the incentivized study (0.633) than in the primary study (0.587). The average rank for the one-at-a-time task was also better in the incentivized study (2.659) than in the primary study (3.177).

Incentives appear more effective for the high CRT group, although it is difficult to draw conclusions with so few subjects. The data at least suggests that incentives can improve the performance of high CRT subjects on the one-at-a-time task (60.1% optimal assignments with optimal option incentivized vs. 70% with all options incentivized). Similarly, the average rank of assigned options is much better for the one-at-a-time pricing task with all options incentivized than it is with only the optimal option incentivized (average ranks of 1.30 vs. 3.52, respectively). Indeed, the average rank of 1.30, if it holds for a larger group of high CRT subjects, could make the incentivized pricing one-at-a-time task competitive with the choice tournament. In contrast, the percent of optimal assignments is lower for the high CRT group in the all-at-once task than for the low CRT group, but that is also likely due to small sample size.

Table 7. Average Rank and Percent of Optimal Assignments for Incentivized Pricing Tasks

		Average Rank of Assigned Options		Percent of Optimal Assignments	
CRT Group	Subjects	One at a Time	All at Once	One at a Time	All at Once
Low	30	4.675	2.926	0.392	0.507
Medium	23	2.659	3.582	0.633	0.411
High	5	1.300	2.250	0.700	0.250
Overall	58	3.585	3.127	0.514	0.447

9. Discussion

In Section 2, we laid out five questions that our study was intended to answer. With regard to question (i), our results show people perform better when presented with all options simultaneously rather than in isolation. With regard to (ii), we find some evidence that more numerical or calculation-based response modes (such as a monetary valuation task) do not improve decision making relative to more qualitative or feeling-based response modes (such as a happiness rating task). With regard to (iii), we find that, in general, more constrained response modes (those with fewer possible responses per option) perform better than less constrained response modes. With regard to (iv), we find that heterogeneity in cognitive reflection is remarkably effective in sorting out heterogeneity in performance across architectures. When using percentage of optimal choices and money left on the table as metrics, those with higher levels of cognitive reflection consistently perform better than those with lower levels of cognitive reflection across all six architectures in our experiment. Also participants who differ in cognitive reflection do not perform best on the same architectures. We find that the simple choice architecture (where all options are presented simultaneously) induces the best overall decisions for subjects with low cognitive reflection, and that the sequential tournament performed best for subjects with moderate to high cognitive reflection. The rankings observed for the above results are generally consistent across our three performance metrics (proportion of optimal responses, average welfare ranking, and money left on the table). With regard to (v), we find that no architecture is Pareto efficient and the optimal architecture depends on one's preferred welfare criterion. In particular, Rawls' maximin criterion recommends implementing the simple choice architecture, whereas Harsanyi's utilitarian criterion recommends implementing the sequential tournament architecture.

More broadly, there has been relatively little research assessing the optimality of different features of a choice architecture. The possible components of a choice architecture that are behaviorally relevant are well known and include, for instance, the frame of a decision, and the presence of a default option, in addition to the response mode and presentation mode. Thus far the literature on framing and default options has largely focused on inconsistencies or preference reversals (Thaler 1980, Tversky and Kahneman, 1981) rather than on which frames or default options might induce better decisions. There have been a few

exceptions: see Read et al. (2005) on temporal frames which induce more patient behavior; Thaler and Benartzi (2004) on improving savings decisions by systematically modifying the default option; and Johnson and Goldstein (2003) on how default options can save lives through increased organ donations.

As noted earlier, previous research has studied preference reversals across response modes (e.g., Lichtenstein and Slovic, 1971; Grether and Plott 1979; Tversky et al., 1990; Slovic et al., 2007), and preference reversals across presentation modes (e.g., Hsee 1996; Hsee et al., 1999; Hsee and Zhang, 2010). However, little research has focused on the optimality of response modes or presentation modes[11]. Our design enables us to test both of these features of a choice architecture in individual decisions, within subjects, providing evidence that choice outperforms happiness ratings, which outperforms pricing, holding the presentation mode fixed, and that joint presentation yields better performance than separate presentation, holding the response mode fixed.

Surprisingly, we also found that for subjects with low cognitive reflection, choosing among *more* options at a time produced *higher* decision quality than choosing directly from a smaller choice set. This observation illustrates a fundamental tradeoff between presentation complexity and response complexity: For a fixed set of alternatives, a choice architecture designed with smaller presentation sets must also be designed to elicit multiple responses. If error rates are sufficiently high, as they appear to be for the low cognitive reflection subjects in our experiment, then smaller presentation sets can reduce decision quality because they present multiple opportunities for error.

Fernandes et al. (2014) observe, "Public policy tools drawn from economics point to three broad classes of interventions to help consumers make better decisions: offering more choices; providing better information to consumers about options they might consider; and providing incentives for consumers or sellers to change their behavior." Our results suggest that a fourth class of interventions – those which manipulate the structure of the decision task (e.g., the presentation mode or response mode) or the flexibility

[11] But see Hsee (1998) and List (2002) on between-subject valuations for a dominant and a dominated option, and see Bohnet et al. (2015) on using joint presentation mode to improve profit maximizing performance evaluation procedures by corporations.

of the decision task (e.g., the range of possible responses permitted by the architecture) can also be effective in improving decisions. In addition, our findings suggest that performance is best on choice tasks, perhaps because they are simple, familiar, and intuitive. It is also important to know the population for whom the architecture is being designed, as our results pertaining to heterogeneity in cognitive reflection suggest that there is not one architecture which consistently optimizes performance for everyone. Knowing the population the choice architect is trying to help may lead to improved tradeoffs among design features when engineering choice architectures to optimize decision quality.

References

Aimone, J., Ball, S., King-Casas, B. 2016. 'Nudging' risky decision-making: The causal influence of information order. *Economics Letters*. **149**: 161-163.

Arieli, A., Ben-Ami, Y., & Rubinstein, A. 2011. Tracking Decision Makers under Uncertainty. *American Economic Journal: Microeconomics*, **3**(4): 68–76

Bergus, G. R., Levin, I. P., & Elstein, A. S. 2002. Presenting Risks and Benefits to Patients. *Journal of General Internal Medicine*, **17:** 612–617.

Besedeš, T., Deck, C., Sarangi, S., & Shor, M. 2012a. Age Effects and Heuristics in Decision Making, *Review of Economics and Statistics*, **94**(2):580-595.

Besedeš, T., Deck, C., Sarangi, S., & Shor, M. 2012b. Decision-making Strategies and Performance among Seniors, *Journal of Economic Behavior and Organization*, **81**(2):524-533.

Besedeš, T., Deck, C., Sarangi, S., & Shor, M. 2015. Reducing Choice Overload without Reducing Choices. *Review of Economics and Statistics*, **97**(4)**:**793-802.

Bohnet, I., van Green, A., & M. Bazerman. 2015. When performance trumps gender bias: Joint vs. separate evaluation. *Management Science*, **62**:1225-1234.

Downs, J.S., Loewenstein, G., & Wisdom, J. 2009. Strategies for promoting healthier food choices. *The American Economic Review*, **99**(2): 159-164.

Fernandes, D., Lynch, J.G., & Netemeyer, R.G. 2014. Financial literacy, financial education, and downstream financial behaviors. *Management Science*, **60**:1861-1883.

Frederick, S. 2005. Cognitive reflection and decision making. *The Journal of Economic Perspectives*, **19**:25-42.

Grether, D. M., & Plott, C. R. 1979. Economic theory of choice and the preference reversal phenomenon. *The American Economic Review*, **69**:623-638.

Harsanyi, J.C. 1955. Cardinal welfare, individualistic ethics, and interpersonal comparisons of utility. *Journal of Political Economy*, **63**: 309-321.

Hicks, J. 1939. The foundations of welfare economics. *Economic Journal*, **49**: 696-712.

Hsee, C. K. 1996. The evaluability hypothesis: An explanation for preference reversals between joint and separate evaluations of alternatives. *Organizational behavior and human decision processes*, **67**:247-257.

Hsee, C.K. 1998. "Less is better: When low-value options are valued more highly than high-valued options." *Journal of Behavioral Decision Making*, **11**: 107-121.

Hsee, C. K., Loewenstein, G. F., Blount, S., & Bazerman, M. H. 1999. Preference reversals Between joint and separate evaluations of options: A review and theoretical analysis. *Psychological bulletin*, **125**:576 – 590.

Hsee, C. K., & Zhang, J. 2010. General evaluability theory. *Perspectives on Psychological Science*, **5**:343-355.

Huck, S. & Weizäcker, G. 1999. Risk, Complexity and Deviations from Expected-Value Maximization: Results of a Lottery Choice Experiment. *Journal of Economic Psychology*, **20**: 699-715.

Kaldor, N. 1939. Welfare propositions in economics and interpersonal comparisons of utility. *Economic Journal*, **49**: 549-552.

Lichtenstein, S., & Slovic, P. 1971. Reversals of preference between bids and choices in gambling decisions. *Journal of experimental psychology*, **89**:46 – 55.

List, J.A. 2002. Preference reversals of a different kind: The "more is less" phenomenon. *American Economic Review*, **92**:1636-1643

Peters, E., Hibbard, J., Slovic, P. & Dieckmann, N. 2007. Numeracy Skill and the Communication, Comprehension and Use of Risk-Benefit Information. *Health Affairs*, **26**(3): 741-748.

Rawls, J. 1971. *A Theory of Justice*. Harvard University Press. Cambridge, MA.

Read, D., Frederick, S., Orsel, B., & Rahaman, J. 2005. Four score and seven years from now: the date/delay effect in temporal discounting. *Management Science*, **51**:1326-1335.

Reutskaja, E., Nagel, R., Camerer, C. F., & Rangel, A. 2011. Search Dynamics in Consumer Choice Under Time Pressure: An Eye-Tracking Study. *The American Economic Review*, **101**(2): 900-926.

Roth, A.E. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics, *Econometrica*, **70**(4): 1341-1378.

Slovic, P., Finucane, M. L., Peters, E., & MacGregor, D. G. 2007. The affect heuristic. *European Journal of Operational Research*, **177**:1333-1352.

Thaler, R. 1980. Toward a positive theory of consumer choice. *Journal of Economic Behavior & Organization*, **1**:39-60.

Thaler, R., & Benartzi, S. 2004. Save More Tomorrow: Using Behavioral Economics to Increase Employee Saving. *Journal of Political Economy*, **112:** S164- S187.

Toplak, M. E., West, R. F., & Stanovich, K. E. 2014. Assessing miserly information processing: An expansion of the Cognitive Reflection Test. *Thinking & Reasoning*, **20**:147-168.

Tversky, A., & Kahneman, D. 1981. The framing of decisions and the psychology of choice. *Science*, **211**:(4481), 453-458.

Tversky, A., Slovic, P., & Kahneman, D. 1990. The causes of preference reversal. *The American Economic Review*, **80**:204-217.