DISCUSSION PAPER SERIES

No. 10457

# CONSISTENCY IN SIMPLE VS. COMPLEX CHOICES OVER THE LIFE CYCLE

Isabelle Brocas, Juan D Carrillo, T. Dalton Combs
and Niree Kodaverdian

*PUBLIC ECONOMICS*


Centre for Economic Policy Research

*ISSN 0265-8003*

# CONSISTENCY IN SIMPLE VS. COMPLEX CHOICES OVER THE LIFE CYCLE

*Isabelle Brocas, Juan D Carrillo, T. Dalton Combs and Niree Kodaverdian*

Discussion Paper No. 10457
March 2015
Submitted 21 February 2015

Centre for Economic Policy Research
77 Bastwick Street, London EC1V 3PZ, UK
Tel: (44 20) 7183 8801
www.cepr.org

# CONSISTENCY IN SIMPLE VS. COMPLEX CHOICES OVER THE LIFE CYCLE[†]


## Abstract

Employing a variant of GARP, we study consistency in aging by comparing the choices of younger adults (YA) and older adults (OA) in a 'simple', two-good and a `complex' three-good condition. We find that OA perform worse than YA in the complex condition but similar in the simple condition. Working memory scores correlate significantly with consistency levels. Finally, OA are more prone to use simple heuristics than YA, and this helps them behave consistently in the simple condition. Our findings suggest that the age-related deterioration of neural faculties responsible for working memory is an obstacle for consistent decision-making.



JEL Classification: C91, D11 and D12
Keywords: aging, complexity, laboratory experiments and revealed preferences



Isabelle Brocas brocas@usc.edu
*University of Southern California and CEPR*

Juan D Carrillo juandc@usc.edu
*University of Southern California and CEPR*

T. Dalton Combs
*University of Southern California*

Niree Kodaverdian kodaverd@usc.edu
*University of Southern California*



[†] We are grateful to members of the Los Angeles Behavioral Economics Laboratory (LABEL) for their insights and comments in the various phases of the project. We also thank Cary Deck, Mara Mather, John Monterosso, and participants at the 2014 LABEL Experimental Economics Conference (University of Southern California), and at the 2013 Social Neuroscience retreat (Catalina island, USC) for useful comments. All remaining errors are ours. The study was conducted with the University of Southern California IRB approval UP-08-00052. We acknowledge the financial support of the National Science Foundation grant SES-1425062.

# 1 Introduction

As most day-to-day decisions involve comparing items and making trade-offs between them, attributing value is a fundamental element of decision-making. Economics builds theories under the assumption that individuals have unambiguous values for items and maintain stable preferences. These in turn imply consistency of choice, which can be tested empirically. Experimental studies have shown that choice consistency is prevalent, at least for younger adults (Andreoni and Miller, 2002; Andreoni and Harbaugh, 2009; Choi et al., 2014). However, little is known about choice consistency in older adults. Understanding the effect of age on consistency can provide a foundation for refined economic models.

Recent field and experimental evidence has shown that older adults (OA) make significantly different choices as compared to younger adults (YA) in a variety of domains (Harrison et al., 2002; Fehr et al., 2003; Holm and Nystedt, 2005; Ameriks et al., 2007; Bellemare et al., 2008; Engel, 2011; Albert and Duffy, 2012; Castle et al., 2012).[1] Such differences can potentially be due to two very different factors: preferences that *change* with age or choices that become more *erratic* with age, and there is indirect evidence for both. Indeed, aging brings dramatic changes in our motivations, which is believed to cause changes in the decisions we make (Carstensen and Mikels, 2005; Mather and Carstensen, 2005). In parallel, the aging process affects many brain structures and brain mechanisms, hindering the ability to evaluate alternatives and select among them (Mohr, et al., 2010; Nielsen and Mather, 2011), especially when they become complex (Brand and Markowitsch, 2010; Besedes et al., 2012a, 2012b). Disentangling between preference changes and mistakes is essential for policy-making purposes (Bernheim and Rangel, 2009) as well as for purposes of cost avoidance on the part of the decision maker (Lichtenstein and Slovic, 1973). In this paper, we propose to use the Generalized Axiom of Revealed Preference (GARP) to test the internal consistency of subjects' preferences by offering repeated choices between bundles of goods. Our goal is to understand the effect of *age* and task *complexity* on choice consistency.

The study builds on three strands of the literature. First, laboratory experiments have used GARP to assess the degree of consistency of subjects in different domains. In the risk domain (bundles of quantities and probabilities), studies find that YA are generally consistent with revealed preference (Choi et al., 2007; Andreoni and Harbaugh, 2009; Choi

[1] However, there is also evidence that OA and YA make similar choices in some of the same domains (Dror et al., 1998; Kovalchik et al., 2005; Sutter and Kocher, 2007; Charness and Villeval, 2009; Chao et al., 2009). Yet others find curvilinear age effects (Harrison et al., 2002; Read and Read, 2004). Some studies offer to resolve these mixed findings arguing that results are highly sensitivity to differences in task characteristics (Mata et al., 2011), task complexity (Zamarian et al., 2008; Brand and Markowitsch, 2010) and contents of choice sets (Mather et al., 2012).

et al., 2014), although violations of GARP are stronger in the loss as compared to the gain sub-domain.[2] In the social domain (bundles of money for oneself and another party), YA are found to be largely consistent with revealed preference (Andreoni and Miller, 2002; Fisman et al., 2007). While experiments in the goods domain (bundles with positive quantities of two or more desirable goods) have reported a moderate range of percentages of consistent subjects, once severity of violations are taken into account, the studies concur on the general consistency of YA (Sippel, 1997; Mattei, 2000; Fevrier and Visser, 2004).[3]

Second, individual decision-making experiments with different age groups have concurred in the finding that young children are less consistent than YA. Bradbury and Nelson (1974) report that inconsistent choices decrease with age in a task of pairwise comparisons of preferred colors whereas Harbaugh et al. (2001) find a significant increase in choice consistency between 8 and 12 years old children and no differences between 12 years old children and undergraduates. These findings suggest that after a certain age, one's consistency culminates and thereafter stabilizes. By contrast, the full trajectory across the lifetime has not been established, as results have been mixed: healthy OA are more consistent than YA according to some studies (Tentori et al., 2001; Kim and Hasher, 2005) while healthy OA are less consistent as compared to YA according to others (Finucane et al., 2002; Finucane et al., 2005). These disparate findings may be partly due to two methodological choices. First and contrary to standard practices in experimental economics, decisions in those studies are not incentivized. Second, they use different domains (health, extra credit, grocery coupons, nutrition, finance). This introduces confounding factors since different age groups have varying degrees of domain-specific expertise (for example, OA are more likely to be familiar, if not experienced, with grocery coupons).

In more recent studies, the findings are mixed yet again. Using a panel of household consumption data, Echenique et al. (2011) find that older, poorer, and less educated households make more severe GARP violations than younger, richer, and more educated households. In a large scale field experiment, Choi et al. (2014) also find that OA are less consistent than YA. By contrast, Dean and Martin (2013) find that older households are more rational as compared to younger households. The need exists for an incentivized, controlled laboratory study comparing choice consistency by YA and OA in a decision context where neither group holds substantially greater domain-specific expertise.

[2]Studies also report GARP consistent behavior in the context of criminal behavior (Visser et al., 2006) and by inebriated (Burghart et al., 2013) or sleepy (Castillo et al., 2014) subjects. In a cross cultural study, Tanzanian YA are found to commit more GARP violations as compared to YA from the United States (Cappelen et al., 2014).

[3]In a multi-domain study (bundles of consumption goods, labor hours, and token money) with female mental hospital patients, Battalio et al. (1973) find some inconsistencies, but when a subsequent work (Cox, 1997) studies the same data taking into account severity of violations, all but one of the subjects is deemed inconsistent.

Third, studies have demonstrated that task complexity imposes demands on working memory. Since working memory is responsible for the short-term mental maintenance and manipulation of information and this process is less efficient in OA, varying levels of task complexity can account for differences in choices between OA and YA. Indeed, activation studies have shown that the working memory circuitry (involving lateral regions of the prefrontal cortex) is recruited during tasks that are deemed difficult, such as task-switching (Dove et al., 2000) or while producing random strings of digits or key presses (Frith et al., 1991; Jahanshahi et al., 2000). Crucially, the region is also differentially recruited as tasks become more complex (Demb et al., 1995; Baker et al., 1996; Braver et al., 1997; Cohen et al., 1997; Carlson et al., 1998). This relationship extends to tasks requiring the explicit representation and manipulation of knowledge, where the ability to reason relationally is essential (Kroger et al., 2002) or when the number of dimensions to be considered simultaneously are increased (Christoff et al., 2001). Interestingly, it has been shown that older adults perform worse on such tasks (Grady et al., 2006; Zamarian et al., 2008; Brand and Merkowitsch, 2010; Henninger et al., 2010) and that a main factor for such decline is the age-related atrophy of regions involved in working memory (Resnick et al., 2003; Raz et al., 2005).

The fundamentals of our GARP experiment is a $2 \times 2$ design (younger/older adults, simple/complex domain) directed to study the effect of aging on consistency as a function of the difficulty of the situation. In the simple domain, subjects choose between two bundles each composed of different quantities of the same two goods (e.g., pistachios and cheese). In the complex domain, subjects choose between two bundles each composed of different quantities also of two goods, but now with only one common good (e.g., pistachios and cheese vs. pistachios and crackers). To understand the determinants of consistency, our subjects also perform working memory and IQ tests. We obtain three main findings.

First, both OA and YA are reasonably (and roughly equally) consistent in the simple domain whereas OA are significantly more inconsistent than YA in the complex domain. Consistency violations in the complex domain are also more severe for OA under several metrics.

Second, differences in violations in the complex domain are associated with differences in performance in the working memory test. Since YA score significantly higher in that test compared to OA, most of the difference in performance across ages is captured through the working memory effect. Our findings thus indicate that the working memory system is heavily recruited in the complex task but not so much in the simple one. The result echoes the aforementioned studies which show this precise relationship between complexity and working memory demands. Interestingly, the result also extends to IQ, although working memory and IQ are strongly correlated for our subjects.

Third, both YA and OA populations are heterogenous. We implement a model-based clustering method to group our subjects in subpopulations. We estimate a random utility model and use the number of misclassifications of the model as the clustering criterion. We find three distinct clusters. The first group of subjects is very inconsistent in both the simple and complex domains. These subjects have the lowest working memory and IQ scores. The other two groups are relatively consistent in both domains. One of these two groups is composed of subjects who commit almost no violations in the simple domain. Interestingly, we show that this near perfect consistency occurs because these subjects resort to heuristics. Their behavior becomes significantly more inconsistent in the complex domain, possibly because heuristics are less intuitive to implement. Many subjects in this cluster are OA. Subjects in the remaining group do not seem to be using heuristics in either domain. They are slightly less consistent than the previous group in the simple domain but significantly more consistent in the complex domain. These subjects have higher working memory scores and they are mostly YA. Taken together, the results suggest that subjects drift from the more consistent cluster to the less consistent ones over the life cycle and that consistency at the old age is achieved whenever heuristics are easily available.

The article is organized as follows. he theoretical framework is presented in section 2. The experimental setting is described in section 3. The aggregate analysis is conducted in section 4 and the individual analysis can be found in section 5. Concluding remarks are gathered in section 6.

# 2 Theory

Consider a subject making choices between pairs of bundles with two goods that are desirable, in the sense that more of each good is strictly preferred to less. A choice between a pair of bundles is called a "trial". Call $a_{xy} := (q_x^a, q_y^a)$ the bundle $a_{xy}$ that has positive quantities $q_x^a$ and $q_y^a$ of goods $x$ and $y$ respectively.

## 2.1 Bundles with identical goods

Suppose first that bundles are composed of the same two goods ($x, y \in \{1, 2\}$ with $x \neq y$) and consider trials with bundles $a_{12}$ and $a'_{12}$ so that each bundle has more quantity of exactly one good ($q_x^a > q_x^{a'} \Leftrightarrow q_y^a < q_y^{a'}$). In the experimental section, this is called treatment **S** (for simple). When a trial is considered in isolation, any choice between pairs of bundles with the aforementioned properties is perfectly valid and depends exclusively on the taste of the subject. However, some combinations of choices may constitute a violation

of revealed preferences (which we call $\mathcal{D}_{\mathcal{S}}$, for direct violation in the simple treatment).[4] Here is why. Consider the example in Figure 1 and suppose that $a_{12}$ is chosen over $a'_{12}$ and $b_{12}$ is chosen over $b'_{12}$. Since $q_x^{a'} > q_x^b$ for all $x$, we have $a_{12} \succ a'_{12} \succ b_{12}$. Since $q_x^{b'} > q_x^a$ for all $x$, we have $b_{12} \succ b'_{12} \succ a_{12}$. This forms a contradiction.

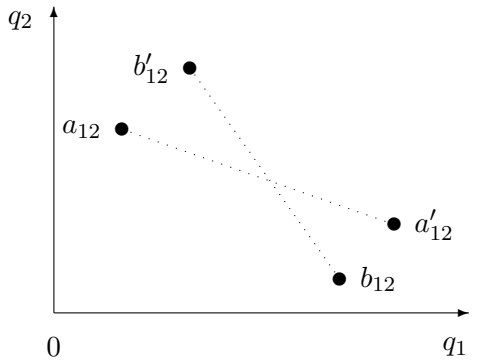


**Figure 1:** Trials $a_{12}$ vs. $a'_{12}$ and $b_{12}$ vs. $b'_{12}$

**Definition 1** *Direct violation in simple treatment ($\mathcal{D}_{\mathcal{S}}$).*

*(i) Trials $a_{12}$ vs. $a'_{12}$ and $b_{12}$ vs. $b'_{12}$ may involve a $\mathcal{D}_{\mathcal{S}}$-violation if and only if $q_x^{a'} > q_x^b$ and $q_x^{b'} > q_x^a$ for all $x$.*

*(ii) A $\mathcal{D}_{\mathcal{S}}$-violation occurs when $a_{12}$ is chosen over $a'_{12}$ and $b_{12}$ is chosen over $b'_{12}$.*

The logic of the argument is very similar to the standard revealed preferences argument made in earlier GARP studies (Sippel (1997), Harbaugh et al. (2001), Choi et al. (2007)). The only difference is that, in our case, the set of choices is dramatically reduced. Therefore a choice in one trial only reveals that the bundle selected is preferred to the only other bundle proposed rather than to any bundle on the "budget line." Notice that Definition 1 is made of two distinct parts. Part (i) provides conditions such that choices in two trials *may* result in a violation. Intuitively, the requirement is that for each trial one bundle dominates (i.e., has more quantity of both goods than) a bundle in the other trial whereas the other bundle is dominated by (i.e., has less quantity of both goods than) the remaining bundle in the other trial. Naturally, some pairs of trials will fail to satisfy this condition, in which case a $\mathcal{D}_{\mathcal{S}}$-violation will not be possible. Given a pair of trials such that a $\mathcal{D}_{\mathcal{S}}$-violation is possible, part (ii) provides conditions such that the violation indeed occurs. Again intuitively, the requirement is that in each trial the subject selects the bundle that

[4]The seminal work on revealed preference theory is due to Samuelson (1938). It was subsequently extended by Houthakker (1950), Afriat (1967) and Varian (1982) among others.

is dominated by a bundle in the other trial. In our example, bundles $a_{12}$ and $b_{12}$. Hence, only one out of the four possible choice combinations will result in a $\mathcal{D}_S$-violation.

Given $n$ trials, there are $n(n-1)/2$ different pairs of trials. By looking at all pairs of trials and checking whether the condition in Definition 1(i) is satisfied, we can identify all possible violations. Then, the actual violations are determined simply by checking whether the bundles selected by the subject on the pairs of trials in which a violation is possible satisfy the condition in Definition 1(ii).[5]

## 2.2 Bundles with different goods

Assume now that there are three possible goods ($x, y, z \in \{3, 4, 5\}$ with $x \neq y \neq z$) and consider trials between pairs of two-good bundles that have exactly one good in common, that is, between bundle $a_{xy}$ and bundle $a'_{xz}$. In the experimental section, this is called treatment **C** (for complex).[6] By definition, each bundle has now more quantity of at least one good (only $a_{xy}$ has a positive quantity of good $y$ and only $a'_{xz}$ as a positive quantity of good $z$). Once again, when a trial is considered in isolation any choice between pairs of bundles is perfectly valid. Definition 2 identifies conditions for a direct violation in two trials to occur. These are similar to the conditions described in Definition 1.

**Definition 2** *Direct violation in complex treatment ($\mathcal{D}_C$).*

*(i) Trials $a_{xy}$ vs. $a'_{xz}$ and $b_{xz}$ vs. $b'_{xy}$ may involve a $\mathcal{D}_C$-violation if and only if $q_x^{a'} > q_x^b$, $q_z^{a'} > q_z^b$, $q_x^{b'} > q_x^a$ and $q_y^{b'} > q_y^a$.*

*(ii) A $\mathcal{D}_C$-violation occurs when $a_{xy}$ is chosen over $a'_{xz}$ and $b_{xz}$ is chosen over $b'_{xy}$.*

In treatment **C**, there is also the possibility of incurring in an indirect (or cyclical) violation $\mathcal{I}_C$. An $\mathcal{I}_C$-violation involves choices in *three* trials each with a different common good. Definition 3 describes an indirect violation.

**Definition 3** *Indirect violation in complex treatment ($\mathcal{I}_C$).*

*(i) Trials $a_{xy}$ vs. $a'_{xz}$, $b_{xz}$ vs. $b'_{yz}$ and $c_{yz}$ vs. $c'_{xy}$ may involve an $\mathcal{I}_C$-violation if and only if $q_x^{a'} > q_x^b$, $q_z^{a'} > q_z^b$, $q_y^{b'} > q_y^c$, $q_z^{b'} > q_z^c$, $q_x^{c'} > q_x^a$ and $q_y^{c'} > q_y^a$.*

[5] Importantly, however, it is not trivial to determine the maximum number of violations that a subject can *effectively* incur. Indeed, when a subject makes a choice that induces a violation it may preclude violations between other pairs of trials. To see this, consider the example in Figure 1 and suppose there is a third trial between bundles $c_{12}$ and $c'_{12}$ such that $q_x^c < q_x^a$ and $q_x^{c'} > q_x^{a'}$ for all $x$. By choosing $a_{12}$ over $a'_{12}$ and $b_{12}$ over $b'_{12}$ the subject incurs in a violation. However, by choosing $a_{12}$ over $a'_{12}$ the subject precludes any possible violation between the pair of trials $a_{12}$ vs. $a'_{12}$ and $c_{12}$ vs. $c'_{12}$ (even though a violation would have occurred had the subject chosen $a'_{12}$ over $a_{12}$ and $c_{12}$ over $c'_{12}$).

[6] Since the choice problem involves more goods, the decision is arguably more complicated. Since a trial has still two bundles and each bundle still has positive quantities of exactly two goods, the two treatments are still comparable.

*(ii) An $\mathcal{I}_{\mathcal{C}}$-violation occurs when $a_{xy}$ is chosen over $a'_{xz}$, $b_{xz}$ over $b'_{yz}$ and $c_{yz}$ over $c'_{xy}$.*

Although the argument is slightly more sophisticated, the idea behind indirect violations is similar to that behind direct violations. An $\mathcal{I}_{\mathcal{C}}$-violation may occur if in each trial, one bundle dominates the bundle composed of the same goods in another trial and the other bundle is dominated by the bundle composed of the same goods in the remaining trial. In Definition 3(i) and given that more quantity is always desirable, we have $a' \succ b$, $b' \succ c$ and $c' \succ a$. If this condition is satisfied, only the combination involving the choice of the three dominated bundles $a$, $b$ and $c$ results in a violation.

# 3 Experiment

## 3.1 Design and procedures

To study choice consistency in younger adults (YA) an older adults (OA) we conducted an experiment based on the setup described in the theory section using an extension of Psychtoolbox. We ran 10 sessions with OA and 7 sessions with YA. Each session had between 5 and 8 subjects and lasted between 1.5 and 2 hours. OA sessions were conducted at two OASIS senior centers in Los Angeles, OASIS Baldwin Hills and OASIS West Los Angeles. A total of 51 OA (age 61-84) were recruited through the OASIS activities catalogue.[7] We omitted from the analysis four subjects due to software malfunctioning and one who explicit expressed miscomprehension of the task halfway through the experiment. There was only one male subject in the entire pool so we excluded him as well. We therefore retained 45 female OA for the analysis.[8] YA sessions were conducted in the Los Angeles Behavioral Economics Laboratory (LABEL) in the department of Economics at the University of Southern California. Subjects were recruited from the LABEL pool of USC undergraduate students.[9] In order to match gender, we recruited 50 YA female USC undergraduate students, age 18-34. A potential concern in this type of studies is differences in education level across populations, since the pool of YA is college students and therefore non-representative of the US population. In fact, if anything, our OA subjects are more educated than the YA (average years of education is 15.9 for OA and 13.4 for

[7] OASIS is a non-profit organization active in 25 states. Its mission is to promote successful aging by disseminating knowledge and offering classes and volunteering opportunities to its members (age 50 and older). More information can be found at `http://www.oasisnet.org`.

[8] Such extreme gender selection is striking. Our understanding is that women tend to be more involved in activities at OASIS and more generous with their time Besedes et al. (2012b) also report a larger fraction of female participation (75%), although the difference is not as extreme as ours. This should be taken into account when interpreting the results.

[9] For information about the laboratory, see `http://dornsife.usc.edu/label`.

YA). All subjects were compensated with a fixed amount of $20 plus an incentive payment (described below).

*GARP task.* Each subject participated in 140 core trials with five goods $(1, 2, 3, 4, 5)$. In each core trial, subjects chose between two bundles with two goods each, and were not allowed to express indifference. There were 35 core trials of the simple treatment **S**, where the same two goods $(1, 2)$ appeared on both bundles ($a_{12}$ vs. $a'_{12}$). There were also three sets of 35 core trials of the complex treatment **C**, where each bundle had one common good and one unique good for a total of three goods $(3, 4, 5)$ in each trial. These three sets of 35 trials were identical up to a permutation of the identity of the common good: good 3 ($a_{34}$ vs. $a'_{35}$), good 4 ($a_{34}$ vs. $a'_{45}$) and good 5 ($a_{35}$ vs. $a'_{45}$). Quantities in each bundle were chosen to maximize the chances to satisfy condition (i) in Definitions 1, 2 and 3: for each trial, we chose one bundle that dominated a bundle in as many other trials as possible and the other bundle that was dominated by a bundle also in as many other trials as possible. The reason was to give enough chances to observe $\mathcal{D}_{\mathcal{S}}$-, $\mathcal{D}_{\mathcal{C}}$- and $\mathcal{I}_{\mathcal{C}}$-violations if the subjects were inconsistent (the list of all 35 trials can be found in Appendix A1). Finally, we added 10 trivial trials to check for the attentiveness of subjects, treatment **A**. In these trials, subjects chose between different quantities of the same good ($q_x$ vs. $q'_x$). These trivial choices are typical in psychology experiments (under the misleading terminology "catch trials") but less common in economics, which assumes that incentive payments ensure attentiveness. For subjects who failed to choose the higher quantity option in treatment **A**, our design is not intended to (and therefore cannot) distinguish between inattention, satiation, disliking or miscomprehension of the task, although our procedures were intended to minimize all four possibilities as described below. Either way, such violations would call into question the reliability and interpretability of choices in treatments **S** and **C**. All subjects faced the 150 trials which were presented in a randomized and counterbalanced order. Table 1 summarizes the information.

| **Treatment** | **Goods** | **# of trials** |
|---|---|---|
| **S** | (1,2) vs. (1,2) | 35 |
| **C** | (3,4) vs. (3,5) | 35 |
| **C** | (3,4) vs. (4,5) | 35 |
| **C** | (3,5) vs. (4,5) | 35 |
| **A** | (1) vs. (1) | 10 |
| Total | | 150 |

**Table 1:** Summary of treatments

A major concern in experiments on revealed preferences is the choice of goods. Following some of the recent literature on revealed preferences and value elicitation (Harbaugh et al. (2001); Hare et al. (2009); Rangel and Clithero (2013)), we opted for food items. We presented subjects with 21 salty and sweet snacks and asked them to pick five of them for consumption: two were then randomly used in treatment **S** and the other three in treatment **C**. Each portion was small (for example, one portion consisted of "two pistachios") ensuring that the maximum quantity offered of each good was substantially below satiation level.[10] Subjects were instructed not to eat or drink anything except for water for a period of at least three hours prior to the experiment and all sessions were conducted between 10am and 2:30pm to ensure that subjects were hungry. Figure 2 presents a sample screenshot of a trial in treatment **C**. In this example, the subject had to choose between 5 portions of crackers plus 1 portion of almonds and 4 portions of crackers plus 2 portions of pretzels. At the end of the experiment, one trial was randomly selected and the subject had to consume in the experimental room the choice made in that trial. Every subject complied with the procedure. One advantage of using food is that subjects cannot trade goods at the end of the experiment.

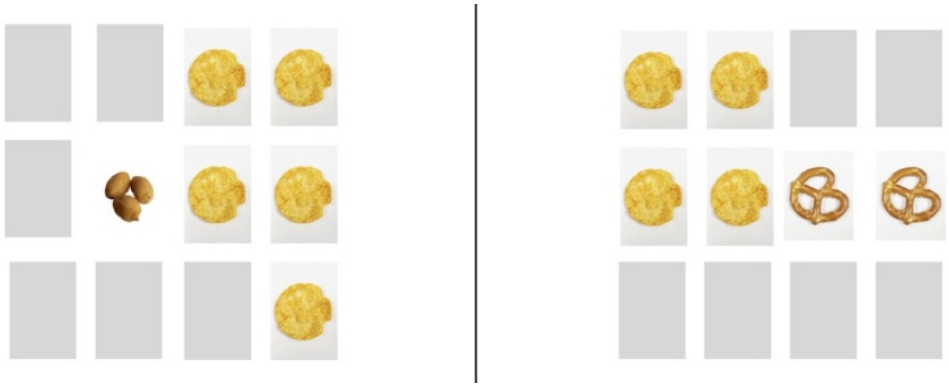

**Figure 2:** Screenshot of one trial in treatment **C**

*Working memory and Raven's IQ tests.* At the end of the GARP task, subjects performed a spatial working memory test and an IQ test. For working memory, we used the computerized Spatial Working Memory test (WM) developed by Lewandowsky et al. (2010). This test measures the capacity of individuals to store and retrieve information in short term memory. It runs as follows. The individual observes a $10 \times 10$ grid. A trial

[10] We made sure that all five selected items were desirable. To address the issue of complementarity or substitutability of goods, we also made sure that subjects understood they would have to consume a combination of two items at the end of the experiment. Appendix A2 presents the list of food items and portions used in the experiment.

consists of a sequence of 2 to 6 dots that appear in different cells of the grid for 0.9 seconds with 0.1 seconds between dots. The objective of the individual is to tap in any order the cells where the dots appeared. Score decreases with the distance between the correct and the selected cells. The entire test consists of 30 such trials. No feedback is given between trials or upon completion of the test. For IQ, we used the short version of the Raven's IQ test, namely Set I of the Raven's Advanced Progressive Matrices (APM) as developed by Raven et al. (1998). This set consists of 12 non-verbal multiple choice questions that become progressively more difficult. For each question, there is a pattern with a missing element. From the eight choices below the pattern, the subject is to identify the piece that will complete the pattern. As Set I is typically used as a screening tool for Set II of the APM, the test provides a rough measure of IQ. Instructions for the test were read directly from the script provided with the test. The test was administered in the intended format (paper) and was not timed. Subjects were made familiar with the format of the test and method of thought required through two practice problems preceding the test. During this time, they were allowed to ask questions from the experimenters. The test started only after all subjects had affirmed their understanding of the instructions.

*Questionnaire.* Following completion of the tests, subjects were asked to complete a questionnaire, adapted from the one used by the Emotion and Cognition Lab at USC. It included questions about their highest diploma, occupation, income, ethnicity, various stress rankings and health levels, as well as information relative to current medications.

*Summary.* A sample copy of the instructions can be found in the Appendix. From a design viewpoint, there are two new elements relative to the existing experimental tests of revealed preferences: we study choice across ages and across task complexity but most importantly, the interaction between the two. From a methodological viewpoint, there are also two main novelties. First, we add trivial tasks. This allows us to differentiate between subjects who violate consistency because they violate one of the premises of the model (inattention, satiation, disliking, miscomprehension) from those who violate consistency even though they satisfy all those premises. Second, each trial has only two possible choices. This is obviously less rich than the traditional setting, where a large number (or even a continuum) of options are presented. However, it allows us to focus on a simpler choice problem with an easy graphical depiction so that we can conduct a large number of trials in a relatively short period of time.[11]

[11] Our choice design contrasts with some recent experimental literature in other domains (risk, time) where it is shown that convexifying the budget set helps obtaining accurate estimates (Andreoni and Sprenger (2012a,b)). Note also that Choi et al. (2007) also perform many trials thanks to their ingenious software presentation. As explained in their paper, it is important to make sure that consistency is not accidental. This is ensured with a large number of decisions so that the test of Bronars (1987) is sufficiently

### 3.2 Hypotheses

As discussed in the introduction, the existing literature provides mixed evidence regarding consistency over the life cycle. The goal of our $2 \times 2$ design (YA/OA, simple/complex) is to evaluate each population in two different environments in order to test whether differences in consistency are related to the difficulty of the choice. In general, it is difficult to compare levels of difficulty across choices (for example, is choosing between two houses more or less difficult than choosing between two mortgages?). Our design addresses that concern by proposing bundles of choices that differ exclusively in the number of goods involved. We make the following hypothesis.

**Hypothesis 1** *OA are less consistent than YA in complex decisions but equally consistent in simple decisions.*

We conjecture that the ability to choose consistently between alternatives does not deteriorate *uniformly* with age. Instead, it crucially depends on the type of situation. If the decision is sufficiently simple, both YA and OA will perform well and differences across ages (both in terms of amount and severity of violations) will be small to non-existent. By contrast, for more complicated decisions, the ability to think through finer aspects of the choice process becomes relatively more important for optimal decision-making and this ability may be somewhat impaired in older populations. Our second hypothesis deals with the determinants of choice inconsistencies.

**Hypothesis 2** *The decrease in consistency by OA is explained in part by compromised working memory and fluid intelligence.*

There might be multiple reasons why OA perform worse than YA in complex situations. Working memory is well known to deteriorate with age and we conjecture that it may play a critical role in treatment **C**. Indeed, even though subjects always compare bundles composed of exactly two goods, a total of three goods need to be evaluated in conjunction in treatment **C**. This means that subjects need to store and retrieve consistently from working memory more information regarding value in **C** rather than **S**. We hypothesize that the decrease in performance by OA in the complex treatment may be related to their inability to carry over all the information pertaining to the three goods. We also conjecture that GARP consistency is related to logical reasoning over values of goods and bundles. As such we hypothesize that fluid intelligence, the reasoning an problem solving capacity of the individual, should play a role in choice consistency. This means that subjects with

powerful. In our view, for our OA population the simplicity of a straightforward repeated 2-options problem is worth the sacrifice in accuracy of estimates and sophistication of software design.

higher scores in IQ tests should behave more consistently, independently of their age. Since fluid intelligence declines with age, part of the decrease in performance by OA may be attributed to that decline. Our last hypothesis deals with the strategies employed by subjects.

**Hypothesis 3** *OA are more likely to resort to heuristics than YA and this will be reflected in the level of consistency.*

Subjects who are aware of limitations in their capacity to store information and perform trade-offs are likely to resort to simple heuristics. This can potentially mitigate the adverse effect of impaired problem solving capacity since, in equilibrium, those subjects will appear among the most consistent of all. We have two alternative hypothesis of the situations where heuristics will be most employed. One possibility is that treatment **S** is complex enough that some subjects will benefit from simple rules. Another possibility is that all subjects are able to perform trade-offs in **S** and that heuristics will be useful as a guideline to succeed in **C**. Either way, we believe that the disparity in difficulty between the simple and complex treatments is sufficiently important that a heuristic shortcut will be applicable to at most one situation. Since the existing literature suggests that OA are more likely to rely on heuristics as opposed to deliberative processing (Thornton and Dumke, 2005; Mata et al., 2007; Peters et al., 2007; Besedes et al., 2012a) but that they do not give up when the choice is complex (Besedes et al., 2012b), we also expect heuristic users to be more prevalent in the OA population of our sample.

# 4 Aggregate analysis

## 4.1 Frequency of violations

Our first and central objective is to assess choice consistency across populations and treatments. Comparisons across treatments are only possible for direct violations since the metric is radically different between direct and indirect violations. To give an idea, for each set of 35 trials there are $\frac{35\times 34}{2} = 595$ pairs of trials, of which 170 can result in direct violations (therefore a total of 170 possible $\mathcal{D}_{\mathcal{S}}$-violations and 510 possible $\mathcal{D}_{\mathcal{C}}$-violations). By contrast, of the $35^3 = 42,875$ triplets of trials in treatment **C**, only 188 can result in a $\mathcal{I}_{\mathcal{C}}$-violation.[12] This means that at most 28.6% of choices can result in direct violations but only 0.4% can result in indirect violations. A better measure to assess the extent of violations across treatments is to compare them with the number of violations incurred

[12]Recall also the caveat in footnote 5 that choices which induce some violations may preclude some others, so 170 and 188 are upper-bounds on the number of effectively feasible violations.

by a subject who chooses randomly between bundles. Figure 3 presents the cumulative distribution function (c.d.f.) of the number of realized $\mathcal{D}_S$-violations in each population (OA, YA) for treatment **S** (left) and the total number of realized $\mathcal{D}_C$- and $\mathcal{I}_C$-violations in each population (OA, YA) for treatment **C** (right). It also presents the c.d.f. of violations when the decisions in each set of 35 trials are simulated 100,000 times using a random choice rule.

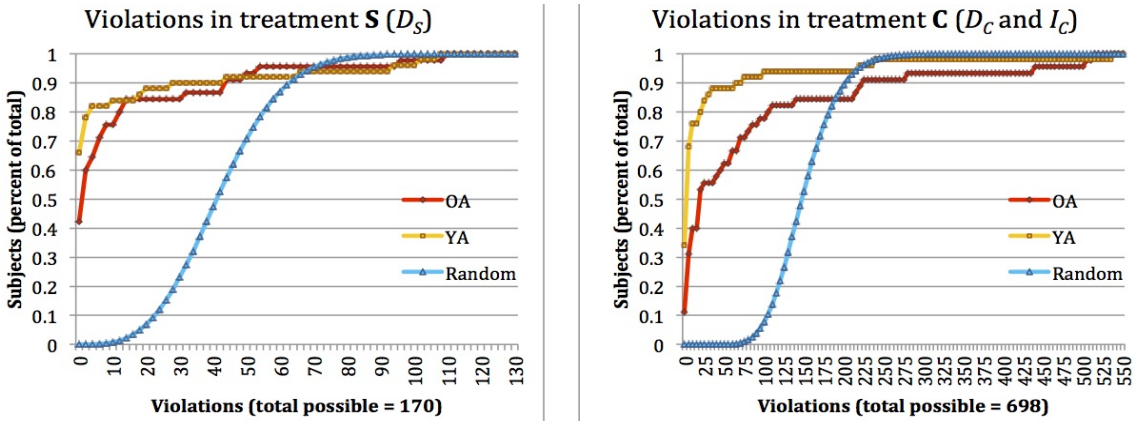


**Figure 3:** Number of violations in treatments **S** (left) and **C** (right)

In treatment **S**, a significant fraction of subjects have no violations (66% of YA and 42% of OA). This fraction shrinks substantially in treatment **C** (34% of YA and 11% of OA). To quantify the extent of violations, we can use the random choice distribution. According to our simulation, there is a 10% chance that a subject choosing randomly will incur less than 23 violations in treatment **S** and 105 in treatment **C**. Using these numbers as a benchmark, we get instead that 88% of our YA and 84% of our OA incur less than 23 violations in treatment **S** and 94% of our YA and 80% of our OA incur less than 105 violations in treatment **C**. Therefore, in line with previous studies (Battalio et al. (1973), Cox (1997), Sippel (1997), Harbaugh et al. (2001), Choi et al. (2007) and others), the majority of our subjects incur relatively few violations.

Perhaps more interestingly, we can compare violations across age groups. We find that YA incur in less violations than OA and differences are more pronounced in treatment **C** than in treatment **S**. More precisely, non-parametric Kolmogorov-Smirnoff (KS) and Wilcoxon Rank Sum (WRS) tests of comparisons of c.d.f. establish marginal differences of distributions in treatment **S** (p-value = .110 and .043 respectively) and strong differences of distributions in treatment **C** (p-value = .001 and .000 respectively).[13] As we can see

[13] As it is well-known, KS is sensitive to any difference on distributions (shape, spread, median, etc.)

from the graph, the difference in treatment **S** is mostly driven by the higher fraction of subjects with 0 violations in the YA population. Figure 3 also highlights the usefulness of the random choice benchmark: even if in both treatments the empirical distributions of violations by YA and OA are significantly smaller than if they were generated by a random choice process, the difference between empirical (YA or OA) and random distributions is more pronounced in **S** than in **C** for both populations. This is consistent with the hypothesis that treatment **C** is more difficult to comprehend and therefore likely to generate *relatively* more mistakes than treatment **S**. In this respect, it is particularly interesting to notice that the 16% of OA who commit the most mistakes in treatment **C** perform worse than the 16% of subjects who would commit the most mistakes if they all behaved randomly. As we will see later on, these are subjects who are likely to violate the premises of the model. Overall Hypothesis 1 is supported by the data. Treatment **C** is more difficult than treatment **S** and generates relatively more mistakes in both populations. OA perform (weakly) worse in both treatments than YA but the difference is most significant for the complex situation than for the simple one.

It is also interesting to distinguish between direct and indirect violations in treatment **C**, especially since $\mathcal{D}_{\mathcal{C}}$-violations are of similar (though not identical) nature to the $\mathcal{D}_{\mathcal{S}}$-violations presented in the left graph of Figure 3. Figure 4 separates violations in treatment **C** into direct ($\mathcal{D}_{\mathcal{C}}$) and indirect ($\mathcal{I}_{\mathcal{C}}$) for each population. As before, it also presents the results of a random choice rule.

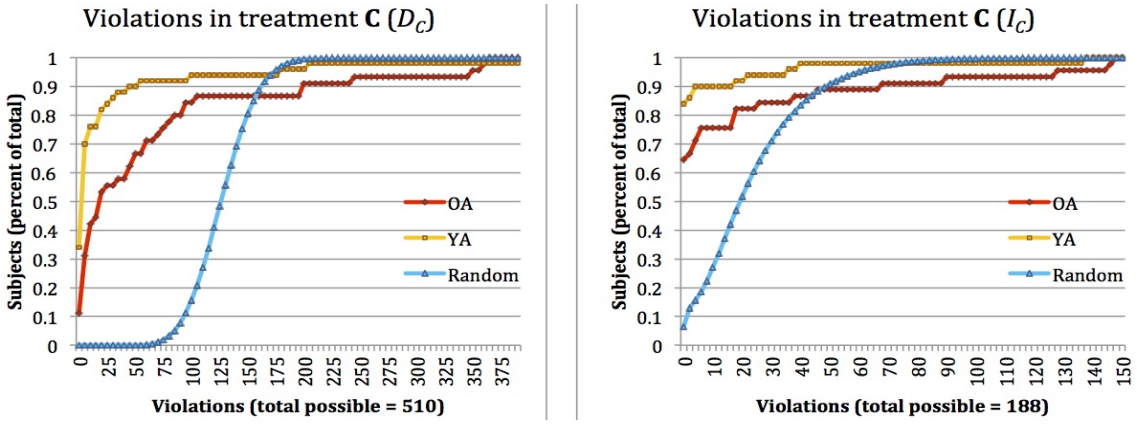


**Figure 4:** Direct (left) and Indirect (right) violations in treatment **C**

According to KS and WRS tests, in treatment **C** differences in the distributions be-

whereas WRS is mostly sensitive to changes in the median. In an attempt to remain agnostic about which test is more adequate in our sample, we will report results for both tests in all our comparisons of distributions.

tween YA and OA are substantially more pronounced for direct violations (p-value = .000 and .000 respectively) than for indirect violations (p-value .187 and .022). The difference is mainly driven by the fact that a relatively high fraction of subjects in both populations (84% of YA and 64% of OA) do not incur any indirect violation. It also suggests that treatment **C** is cognitively more demanding even when we look only at direct violations. Hence, it is the difficulty of having to compute and keep track of the value of a third good which makes the comparison of two-good bundles more challenging and not so much the added possibility of a different type of intransitivity through indirect violations.

### 4.2 Severity of violations

So far we have focus on number of violations. However, not all violations are equally important. Indeed, as emphasized by Afriat's (1967) efficiency index and further developed by Varian (1990) and more recently by Echenique et al. (2011) and Dean and Martin (2013) among others, one should also take into account the *severity* of violations. Populations may differ in frequency of violations but not on severity and viceversa.

There are several ways to study severity. One possibility is to consider a heuristic severity index applied to the case of paired comparisons, which intuitively runs as follows.[14] For each pair (or triplet) of choices involved in a violation, we measure the euclidean distance between the quantities in the choices made by the subject and the quantities in the alternative choices. We then take the minimum of these distances, which we call **d**. This value captures the minimum amount we should change one of the choices of the individual in order to remove the violation. To illustrate the concept, consider the case of a $\mathcal{D}_{\mathcal{S}}$-violation described in Figure 1 (the same procedure applies to $\mathcal{D}_{\mathcal{C}}$ an $\mathcal{I}_{\mathcal{C}}$). If the individual commits a violation (that is, selects $a_{12}$ and $b_{12}$), the severity is given by $\mathbf{d} \equiv \min\left\{d(a_{12}, b'_{12}), d(b_{12}, a'_{12})\right\}$. Intuitively, if $a_{12}$ is very close to $b'_{12}$, it means that the "mistake" is small and reversing two very similar choices would remove the violation.

Including all subjects in the analysis would exacerbate differences in severity between OA and YA since we know from section 4.1 that the fraction of perfectly consistent subjects (for whom $\mathbf{d} = 0$) is larger in the younger population. To avoid this, we include in the analysis only subjects with a positive number of violations and count the average severity of the choices that are inconsistent for that subject (not of all choices). Figure 5 presents the c.d.f of this severity index by population and treatment.

Given the bundles proposed in the experiment, the range of **d** is relatively small: between 1.0 and 3.0 in treatment **S** and between 1.0 and 2.0 in treatment **C**. If anything, this will bias the results against finding differences across treatments. With this in mind,

[14] Contrary to the previously mentioned papers, our goal here is not to develop a new measure of severity in violations but, instead, to use a simple heuristic to quantify their extent.

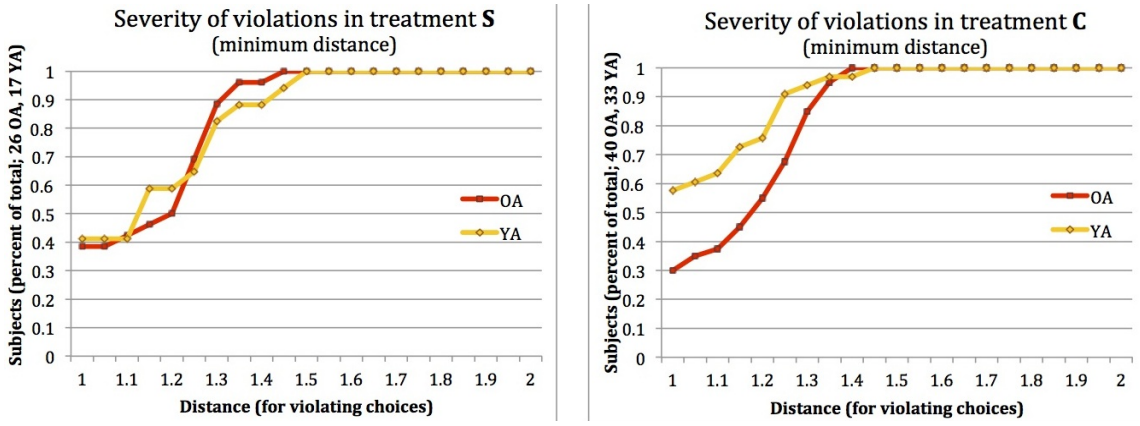


**Figure 5:** Severity of violations in treatments **S** (left) and **C** (right)

we can see from the graph that some subjects commit only the minimal possible violations ($\mathbf{d} = 1.0$) whereas others incur more severe ones ($\mathbf{d} = 1.5$ on average). Perhaps more importantly, in treatment **S** the distribution of severity of violations appears to be equal in both populations (p-value .881 and .858 for KS and WRS tests). By contrast, in treatment **C** violations are significantly more severe for OA than for YA (p-value .049 and .012 for KS and WRS tests). The result thus provides further support to Hypothesis 1: in treatment **C** not only OA commit more violations than YA, but the violations are on average also more severe.[15]

An alternative measure of severity of violations consists in finding for each individual the minimum number of trials that need to be removed in order to suppress all violations for that individual. Obviously, subjects with more violations are likely to necessitate the elimination of more trials to achieve consistency. At the same time, if a subject makes one outlier choice, he may exhibit many inconsistencies that are "cleaned up" when that single trial is removed.[16] As before, we exclude the individuals with no violations to avoid exacerbating differences between OA and YA. This means that the minimum number of trials to be removed is 1. Figure 6 presents the c.d.f. of the number of choices to be

[15] We performed the exact same analysis with the average amount we should change the choices of the individual in order to remove the violation. So, for example, in Figure 1 that would be $\mathbf{d}' \equiv \big(d(a_{12}, b'_{12}) + d(b_{12}, a'_{12})\big)/2$. The results were very similar (sharper difference, if anything): still no significant difference between OA and YA in treatment **S** (p-value .720 and .820 for KS and WRS tests) and significantly more severe violations for OA than YA in treatment **C** (p-value .023 and .002 for KS and WRS tests). The graphs are omitted for brevity.

[16] This is similar to the Houtman-Maks index (Houtman and Maks, 1985), a measure of severity often cited in the literature (e.g., in Choi et al. (2007) and Burghart et al. (2013)), and which is defined as the largest subset of all observed choices that do not include any cycles.

removed for perfect consistency by population and treatment.

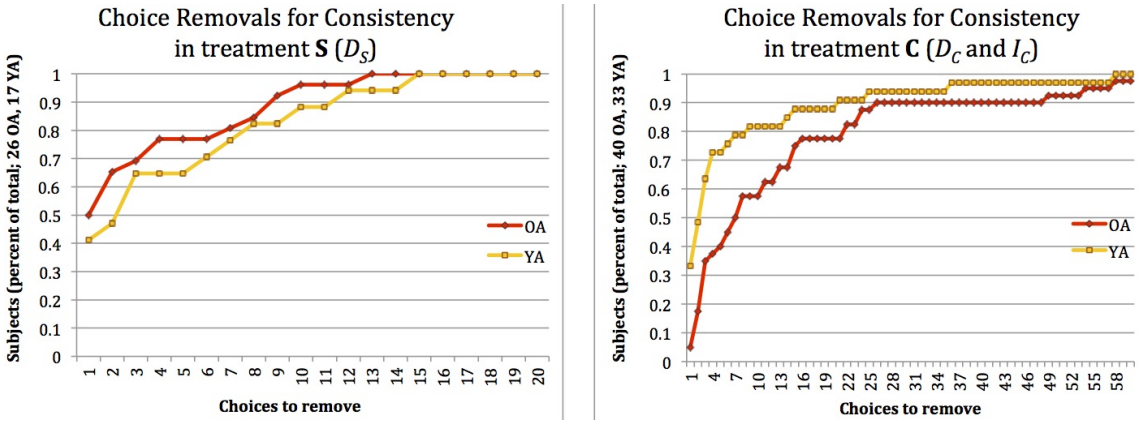


**Figure 6:** Choices to remove for consistency in treatments **S** (left) and **C** (right)

This severity measure yields similar results to the previous one. Indeed, the distribution of the number of choices that need to be removed to achieve consistency is not statistically different for both populations in treatment **S** (p-value .807 and .440 for KS and WRS tests) but it is highly significant in treatment **C** (p-value .016 and .002 for KS and WRS tests). For instance, in order to achieve consistency for two-thirds of the YA in treatment **C**, we only need to remove 3 trials whereas to achieve consistency for the same fraction of older adults we need to remove 14 trials. Overall, the results in this section reinforce our previous findings and the support for Hypothesis 1: OA are less consistent than YA both in the number of violations and in the severity of the violations, and this difference is significantly more pronounced in the complex treatment than in the simple treatment.

### 4.3 Trivial trials

We next analyze the behavior in treatment **A** to see if the premises of our analysis– that subjects are attentive, understand the task, like the good and always prefer more to less– are satisfied. Figure 7 presents the number of violations incurred by the YA and OA in the 10 trivial trials.

The results are highly surprising; we expected some mistakes but not that many. In both populations there is a significant fraction of subjects who violate at least one trivial trial (28% of YA and 62% of OA). There are even 3 subjects who violate all 10 trials. Violations are much stronger in OA than in YA: both KS and WRS tests reject that samples are drawn from the same cumulative distribution functions (p-value = .002 and .001 respectively). This is a severe problem and suggests that at least some of our subjects

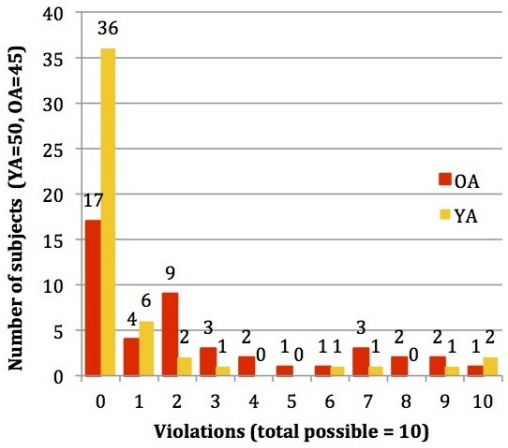


**Figure 7:** Number of violations in treatment **A** (trivial trials)

do not satisfy the assumptions of the model. Subjects who fail 9 or 10 trivial trials are very likely expressing a preference for less rather than more quantity, even though our protocol put the strongest possible emphasis into having hungry subjects, desirable goods and small portions.[17] For subjects who fail 4 or 5 trivial trials, it is more difficult to disentangle between inattentiveness, miscomprehension or interior optimal quantity. Either way, it calls into question the reliability and interpretability of the results on choice consistency. More generally, our results raise a red flag on choice consistency experiments and strongly suggests the importance in these studies of including trivial trials to test whether the premises of the model are satisfied.

A natural step is to conduct the same study as in section 4.1 but only with the subset of the populations that *we think* satisfy the premises of our model. This substantially reduces the sample size, and asymmetrically for YA and OA. Below, we present the results when we restrict attention to subjects who fail at most two trivial trials. We choose that number to remove the subjects who unquestionably violate the premises of the model but, at the same time, to permit some mistakes and keep a reasonable sample size (44 YA and 30 OA). The choice of allowing two errors, however, is admittedly ad-hoc. Figure 8 is the analogue Figure 3 for those individuals.

As expected, violations are reduced when we consider only the subjects with two or less errors in the trivial trials, most notably in treatment **C**. This suggests that a non-negligible fraction of violations may be attributed to factors outside the objective of the study. On the other hand, the basic results of the previous analysis remain unaltered. As

[17] One subject with 101 violations in **S** and 536 violations in **C** explicitly stated during the debriefing that she tried to minimize the quantity to consume.

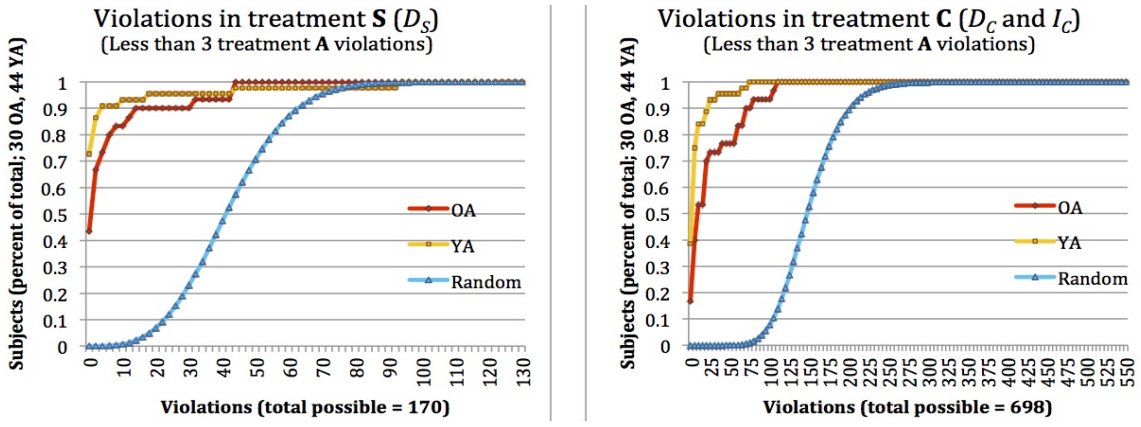


**Figure 8:** Choice violations by subjects with two or less treatment **A** violations

before, there are more violations by OA than by YA and the difference is more significant in the complex treatment than in the simple treatment. Formally, KS and WRS tests show marginal differences in distributions in treatment **S** (p-value = .072 and .016 respectively) and highly significant differences in treatment **C** (p-value = .004 and .001 respectively).[18]

## 4.4 Understanding violations

An intuitive reason why a subject might commit violations is that his preferences do not satisfy the main GARP assumption, namely that more quantity is preferred to less. This is obviously the case for a subject who is not hungry or dislikes the goods and strictly prefers to minimize the total quantity consumed. We compute the choices that we should observe if a given subject use such strategy and we compare the predicted choices with the actual choices of the subject. We then classify a subject as a "minimizer" in a given treatment if more then 80% of his choices are consistent with the minimizing strategy. We find that few subjects could be classified in that way. Only 2 OA fall in that category in both treatments and 5 other subjects minimize quantity in one treatment (5 YA in **S** and 2 YA and 3 OA in **C**). Obviously, these subjects have a large corresponding number of violations. However, not a single subject minimizes quantity in all trials and none of them is therefore perfectly classified as a minimizer. Therefore, it remains unclear whether

[18] Due to the ad-hoc nature of allowing two errors in treatment **A**, we also performed the same analysis with the most conservative possible measure, which is to include only subjects with no errors in trivial trials. Violations decrease substantially and the sample size is dramatically reduced to 17 OA and 36 YA so the statistical power is limited. However, the treatment effect is similar than with the entire population: KS and WRS tests show no significant differences in distributions in treatment **S** (p-value = .534 and .305 respectively) and significant differences in treatment **C** (p-value = .060 and .022 respectively). Again, the graphs are omitted for brevity but available upon request.

the subjects are actually using that strategy so we decided to not exclude them from the analysis.[19]

A main hypothesis of our experiment is that OA will commit more violations than YA due in part to the cognitive difficulty to store information regarding the attributes of the goods. If information storage is at the origin of non-consistent choices, increases in GARP violations are therefore expected to be more pronounced in treatment **C** –where more information has to be carried over– than in treatment **S**. To investigate this hypothesis, we study the scores in the spatial working memory test performed in the experiment.

Not surprisingly, performance in the working memory test is higher for YA (mean = 202, st. error = 3) than for OA (mean = 152, st. error = 1.73), the difference being highly significant (p-value = .000). A simple regression between the working memory score and a group dummy shows that the two are highly correlated (p-value = .000, Adj. $R^2 = 0.71$). The distribution of performances within each population is also similar when we include all subjects or only those with no violation in treatment **A**.

We find no relationship between the number of violations in treatment **S** and working memory performance. To study the relationship between consistency in treatment **C** and working memory, we find practical to collapse violations in a single measure, namely total violations $\mathcal{D}_{\mathcal{C}} + \mathcal{I}_{\mathcal{C}}$.[20] We find that total violations are negatively correlated with working memory scores (Pearson correlation = -.23, Spearman correlation = -.33, p-value = .001).

Another candidate to explain differences in consistency across age groups is IQ. General intelligence has two main components: fluid intelligence, our reasoning and problem solving ability, and crystallized intelligence, our ability to use skills, knowledge and experience. Intuitively, when a subject is asked to choose between two bundles, his objective is to represent accurately his true preferences and act accordingly. This task requires a certain level of logical reasoning about true values, which may rely on fluid intelligence. To test for this hypothesis, we analyze the Raven's IQ answers that we collected from our subjects. This test is designed to measure fluid intelligence. We obtain two main findings.

First, the performance in the Raven's IQ test is higher for YA (mean = 11.44, st. error = 0.16) than for OA (mean = 8.16, st. error = 0.4), the difference being highly significant (p-value =.000). This is not surprising. Indeed, the consensus is that fluid intelligence declines with age after early adulthood while crystallized intelligence remains intact (Horn and Cattell (1967); Kaufman and Horn (1996)). Given that the Raven's test measures fluid intelligence, OA are expected to perform worse. Second, we find a negative relationship

[19] Importantly, however, the results reported below hold if we exclude those subjects from the analysis.

[20] We also conducted the analysis for direct and indirect violations separately. The results are unchanged for $\mathcal{D}_{\mathcal{C}}$. They are qualitatively similar but slightly less significant for $\mathcal{I}_{\mathcal{C}}$. It is also worth noting that $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{I}_{\mathcal{C}}$ are strongly correlated (Pearson correlation = 0.84).

between IQ and the number of violations only in treatment **C** (Pearson correlation = -.22, Spearman correlation = -.26, p-value = .001). This suggests that fluid intelligence is involved in choice processing for the most complex treatment.

To further investigate the relationship between violations in the complex treatment and performance in working memory and IQ tests, we conduct three OLS regressions where the dependent variable is the number of violations in treatment **C**. In all regressions, we include violations in treatment **S** (*Viol-S*) as a control variable. The other variable of the regression is working memory score for the first regression (*WM*), IQ score for the second one (*IQ*) and a Younger Adult dummy for the third one (*YA-d*). The results are presented in Table 2.

| *Const.* | *Viol-S* | *WM* | *IQ* | *YA-d* | Adj. $R^2$ |
|---|---|---|---|---|---|
| 142** | 2.44*** | -0.65* | | | 0.35 |
| (57) | (0.37) | (0.31) | | | |
| 118** | 2.47*** | | -9.3* | | 0.35 |
| (37) | (0.37) | | (3.6) | | |
| 50*** | 2.45*** | | | -46* | 0.35 |
| (14) | (0.37) | | | (18.5) | |

standard errors in parentheses

*, **, *** = significant at 5%, 1% and 0.1% level

**Table 2:** OLS Regression of number of violations in treatment **C**

After controlling for violations in treatment **S**, all three variables have a significant explanatory power to understand consistency in treatment **C**. The similarities between the regressions with working memory and IQ scores are striking. As in previous studies (Engle et al. (1999)), performance in those tests are highly correlated (Pearson correlation = .68, Spearman correlation = .75, p-value = .000), reflecting the fact that both working memory and fluid intelligence can be traced to the same brain systems (Prabhakaran et al. (1997), Kane and Engle (2002), Gray et al. (2003), Olesen et al. (2004), Geary (2005), Jaeggi et al. (2008)). A principal component analysis on the two variables suggests that working memory data contains the largest fraction of the relevant information: the first component is mostly driven by the working memory score and explains 70% of the data. Finally, the regression with the age group dummy delivers similar results, which is not surprising given that age group is a strong predictor of performance in the working memory and IQ tests.

Overall, the findings provide support for the idea that choice inconsistencies in the complex treatment are due to the difficulty to assign value to bundles, store that value in

working memory for subsequent comparisons, and perform logical reasoning over bundles. These difficulties are easily overcome by YA but it prevents OA from choosing consistently across trials.

Next, we examine the responses obtained in our questionnaire. We find that OA self-report a higher stress level compared to YA (p-value = .001) and that reported stress correlates with working memory score (Pearson correlation = .29, Spearman correlation = .29, p-value = .006). Interestingly, self-reported health rankings are similar across groups and uncorrelated to any relevant element of our analysis. We last check for differences across ethnic groups. We first note that our OA population is mostly composed of White and African American subjects while our YA population is composed of White and Asian subjects. Working memory scores, IQ scores and violation counts across White OA and African American OA are not statistically different. The same applies for the comparison between White YA and Asian YA. Ethnicity is therefore not a determinant of choice consistency in our sample.

The results of this section taken together provide strong support for Hypothesis 2. GARP consistency is mediated by the processes involved in working memory and general (fluid) intelligence, both of which are affected by aging. When the environment is simple, the cognitive demands are limited so subjects with a low working memory and fluid intelligence (typically, but not exclusively, OA) can still perform the necessary reasoning. By contrast, when the environment is more complex, the capacity of a subject to store an retrieve information as well as to perform logical reasoning is reflected in the consistency of his choices.

# 5 Individual analysis

The aggregate results suggest that complexity affects the ability to make consistent choices differentially across individuals. Effects are stronger among OA and may be attributable to the decline in working memory. Yet, behavior is heterogeneous even in the OA group indicating that aging is either not affecting all subjects similarly or that some subjects are capable of developing strategies to remain consistent.

## 5.1 A simple random utility model

In order to better understand individual differences, we estimate a random utility model (RUM) for each subject in each treatment. Specifically, in treatment **S**, each subject $i$ in each trial $o$ chooses between a bundle on the left ($l$) of the screen, denoted by $\text{BU}^l$, and a bundle on the right ($r$) of the screen, denoted by $\text{BU}^r$. A decision is obtained by comparing the utility derived by each option. We assume that the utility depends linearly on the

observable quantities of the goods 1 and 2 and a stochastic unobserved error component $\epsilon_i^k$, $k = l, r$. Formally:

$$u_{io}(\text{BU}^l) = \beta_{i1} q_{1o}^l + \beta_{i2} q_{2o}^l + \epsilon_i^l \qquad \text{and} \qquad u_{io}(\text{BU}^r) = \beta_{i1} q_{1o}^r + \beta_{i2} q_{2o}^r + \epsilon_i^r$$

where $q_{jo}^k$ is the quantity of good $j = \{1, 2\}$ in bundle $k = \{l, r\}$ of trial $o$. The probability of individual $i$ choosing option $\text{BU}^l$ is therefore:

$$\begin{aligned} P_{io}^l &= \Pr\left[\beta_{i1} q_{1o}^l + \beta_{i2} q_{2o}^l + \epsilon_i^l > \beta_{i1} q_{1o}^r + \beta_{i2} q_{2o}^r + \epsilon_i^r\right] \\ &= \Pr\left[\epsilon_i^r - \epsilon_i^l < \beta_{i1}(q_{1o}^l - q_{1o}^r) + \beta_{i2}(q_{2o}^l - q_{2o}^r)\right] \end{aligned}$$

Assume that error terms are i.i.d. and follow an extreme value distribution: the cumulate distribution function of the error term is $F_i(\epsilon_i^k) = exp(-e^{-\epsilon_i^k})$. Therefore, the probability that subject $i$ chooses option $\text{BU}^l$ is the logistic function:

$$P_{io}^l(q_{1o}^l - q_{1o}^r, q_{2o}^l - q_{2o}^r) = \frac{1}{1 + e^{-\left(\beta_{i1}(q_{1o}^l - q_{1o}^r) + \beta_{i2}(q_{2o}^l - q_{2o}^r)\right)}}.$$

For each individual $i$ the parameters to estimate are $\beta_{i1}$ and $\beta_{i2}$ which we estimate by maximum likelihood.[21]

A similar model is estimated in treatment **C**. The bundle on the left is made of goods $s$ and $w$ while the bundle on the right is made of goods $p$ and $s$. The utilities are now:

$$u_{io}(\text{BU}^l) = \beta_{is} q_{so}^l + \beta_{iw} q_{wo}^l + \epsilon_i^l \qquad \text{and} \qquad u_{io}(\text{BU}^r) = \beta_{ip} q_{po}^r + \beta_{is} q_{so}^r + \epsilon_i^r$$

and the probability that subject $i$ chooses option $\text{BU}^l$ is the logistic function:[22]

$$P_{io}^l(q_{wo}^l, q_{so}^l - q_{so}^r, q_{po}^r) = \frac{1}{1 + e^{-\left(\beta_{iw} q_{wo}^l + \beta_{is}(q_{so}^l - q_{so}^r) - \beta_{ip} q_{po}^r\right)}}.$$

We estimate the parameters for each individual in each treatment. We then predict the choice in each trial given the estimated parameters and we count the number of

[21] We obtain $O$ observations. The log-likelihood is therefore:

$$\log L_{ik} = \sum_o \log\left[P_{io}^l(q_{1o}^l - q_{1o}^r, q_{2o}^l - q_{2o}^r)\mathbf{1}_l + [1 - P_{io}^l(q_{1o}^l - q_{1o}^r, q_{2o}^l - q_{2o}^r)][1 - \mathbf{1}_l]\right]$$

where $\mathbf{1}_l = 1$ if $\text{BU}^l$ is chosen and $\mathbf{1}_l = 0$ if $\text{BU}^r$ is chosen.

[22] The log-likelihood is now:

$$\log L_{ik} = \sum_o \log\left[P_{io}^l(q_{wo}^l, q_{so}^l - q_{so}^r, q_{po}^r)\mathbf{1}_l + [1 - P_{io}^l(q_{wo}^l, q_{so}^l - q_{so}^r, q_{po}^r)][1 - \mathbf{1}_l]\right]$$

misclassified trials. We find that the misclassification rate in each treatment is strongly correlated with the number of violations (Pearson coefficient = .79 in treatment **S** and .77 in treatment **C**).[23] This suggests that the classification level of RUM is a reliable proxy for GARP consistency: subjects who are not well predicted by the model are inconsistent.[24]

### 5.2 Clustering

In this section we use RUM misclassification data to group individuals with the objective of finding common patterns of behavior. For each individual, we compute the percentage of misclassified trials given the maximum likelihood estimation of the RUM model in treatments **S** and **C** respectively. Contrary to violation counts, these two percentages are comparable between treatments. They provide two interpretable measures related to (but not based on) violations that we can use to cluster our subjects. We consider a model-based clustering method to identify the clusters present in our population. We retain two measures: the % of RUM misclassifications in **S** and the difference between the % of RUM misclassifications in **C** and the % of RUM misclassifications in **S**. We opt for this second measure (rather than simply % of RUM misclassifications in **C**) because of the importance of the treatment effect between simple and complex choices. A wide array of heuristic clustering methods are commonly used but they usually require the number of clusters and the clustering criterion to be set ex-ante rather than endogenously optimized. Mixture models, on the other hand, treat each cluster as a component probability distribution. Thus, the choice between numbers of clusters and models can be made using Bayesian statistical methods (Fraley and Raftery, 2002). We implement our model-based clustering analysis with the Mclust package in R (Fraley and Raftery, 2006). We consider ten different models with a maximum of nine clusters each, and determine the combination that yields the maximum Bayesian Information Criterion (BIC).[25] For our data, the ellipsoidal, equal shape model that endogenously yields *three* clusters maximizes the BIC.

Table 3 provides summary statistics of the three clusters. The first two rows display

[23] We obtain the same results if we split GARP violations between $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{I}_{\mathcal{C}}$.

[24] Note that RUM presupposes more errors when the difference in utility between the two bundles is small. To check the specification of the model, we ran a Probit regression of the probability of correct classification as a function of the absolute utility difference $|\mathrm{BU}^r - \mathrm{BU}^l|$. As predicted by RUM, most subjects have positive coefficients (better classification when utility differences are large). Also, subjects with negative coefficients are those with highest number of violations (hence, those for which RUM is not well specified).

[25] Specifically, hierarchical agglomeration first maximizes the classification likelihood and finds the classification for up to nine clusters for each model. This classification then initializes the Expectation-Maximization algorithm which does maximum likelihood estimation for all possible models and number of clusters combinations. Finally, the BIC is calculated for all combinations with the Expectation-Maximization generated parameters.

the average percentage of RUM misclassifications in **S** and **C** by subjects in each cluster (the main variables used for the clustering). The next two rows present the composition of YA and OA in each cluster. The last five rows summarize the average performance within cluster in the consistency task (GARP violations) and the tests (WM and IQ).

| | **Cluster 1** | **Cluster 2** | **Cluster 3** |
|---|---|---|---|
| *% RUM misclassifications in* **S** | 3.2 (0.6) | 16.2 (0.5) | 36.5 (5.0) |
| *% RUM misclassifications in* **C** | 12.7 (1.4) | 17.0 (1.2) | 40.4 (5.3) |
| *Number of YA* | 13 | 30 | 7 |
| *Number of OA* | 20 | 15 | 10 |
| *Number of violations in* **S** | 1.3 (0.6) | 3.6 (1.6) | 48.1 (9.6) |
| *Number of violations in* **C** | 29.6 (9.7) | 18.1 (6.3) | 189.2 (47.0) |
| *Number of violations in* **A** | 1.6 (0.4) | 0.9 (0.3) | 4.7 (1.0) |
| *Working Memory test* | 173.3 (5.1) | 187.0 (4.2) | 169.8 (8.3) |
| *IQ test* | 9.8 (0.4) | 10.2 (0.4) | 9.2 (0.8) |

standard errors in parentheses

**Table 3:** Summary statistics by cluster.

Clusters are ordered from smallest to largest in the percentage of misclassified observations. Cluster 1 is characterized by almost no misclassification in **S** and few in **C**. Cluster 2 exhibits also limited misclassifications in both **S** and **C** (but more than cluster 1) whereas cluster 3 has substantial misclassifications in both treatments. OA and YA populations are allocated differently to our clusters. The majority of YA belong to cluster 2 and few to cluster 3. By contrast, OA are represented in similar proportions in all groups.

When we look at the performance in the choice tasks and tests, we immediately notice that cluster 3 stands out as a group of inconsistent subjects exhibiting a large number of GARP violations and low performance in the working memory and IQ tests. These subjects also fail our trivial trials much more frequently than the rest. Not surprisingly, the vast majority of minimizers (6 in treatment **S** and 5 in treatment **C**) belong to this cluster. The remaining (1 in treatment **S** and 2 in treatment **C**) belong to cluster 2.

Clusters 1 and 2 are composed of relatively consistent subjects and differ mostly in the way their behavior compares between treatments. In treatment **S**, subjects in cluster 1 are very well classified and have almost no violations while subjects in cluster 2 remain more consistent. In treatment **C**, subjects in cluster 1 decrease significantly their performance while subjects in cluster 2 remain more consistent.[26] Overall, cluster 2 is a group

[26] A series of t-tests shows that the percentage of misclassified trials between clusters are significantly

of "consistently consistent" subjects. By contrast, subjects in cluster 1 are remarkably consistent in **S** but significantly less in **C**.

Figure 9 provides two different representations of the three clusters. In the left graph, clusters 1, 2 and 3 are displayed according to the % of RUM misclassifications in treatments **S** and **C** (rows 1 and 2 in Table 3).[27] In the right graph, these same subjects and clusters are represented based on (a log transformation of) the average number of violations in treatments **S** and **C** (rows 5 and 6 in Table 3).

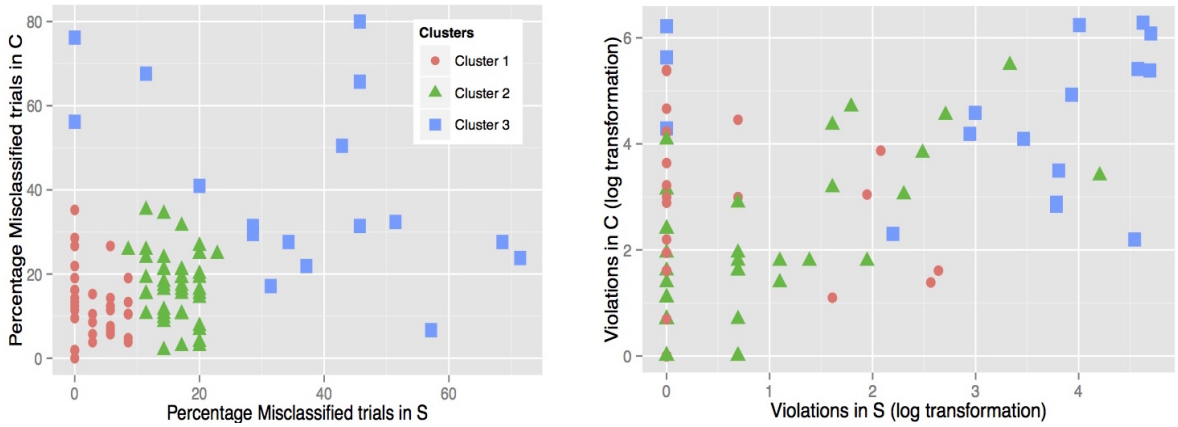


**Figure 9:** Cluster representation. Misclassified trials (left) and number of violations (right) in treatments **S** and **C**.

Clusters are clearly differentiated in the left graph. This is is not surprising since the variables presented are (up to a transformation) the ones used for grouping the individuals. The figure highlights the differences across clusters emphasized above, namely the small percentage of RUM misclassifications in both treatments for cluster 1, the slightly larger in **S** for cluster 2 and the substantial fraction of misclassifications for cluster 3 in both treatments.

Perhaps more interestingly, the right graph also shows a relatively clear distinction in violations across clusters. Cluster 1 has (with a few exceptions) almost no violations in **S** and some in **C**, cluster 2 has a more even distribution of violations between **S** and **C** than cluster 1 and cluster 3 is, again, an outlier in both types of violations. This reasonable mapping is quite remarkable given that subjects are not clustered on the basis of that variable. It suggests a tight relationship between classification by RUM and likelihood of

different (p-value = .000 in **S** and p-value = 0.020 in **C**.

[27] Recall that the exact variables used to group the individuals are % of RUM misclassifications in **S** and difference between the % of RUM misclassifications in **C** and **S**. Our display helps visual clarity (both measures are between 0 and 100) while keeping the essence of the clustering.

violations but also that the transition between the simple and complex situation is not equally difficult for all individuals. In section 5.3, we investigate this issue in more detail.

Finally, to better understand the differences between clusters 1 and 2, we analyze the working memory and IQ scores of the two clusters. Subjects in cluster 1 have significantly worse working memory scores than subjects in cluster 2 (p-value = .040), suggesting a relationship between working memory and the ability to remain consistent across treatment. They also have lower IQ scores but the difference is not statistically significant.

### 5.3 Heuristics

The extreme degree of consistency (and lack of misclassifications) by cluster 1 subjects in treatment **S** is somewhat puzzling. Indeed, 26 subjects out of 33 have zero violations in **S**. Examining in more detail the value estimates of the model (the $\beta_{ij}$-coefficients), we find that for some subjects one value estimate in **S** and two value estimates in treatment **C** are close to 0. These are subjects whose behavior is consistent with maximizing the quantity of one item. For some other subjects, the value estimates of all goods are almost identical to each other. These are subjects whose behavior is consistent with maximizing the total quantity in the bundle. These two choice strategies are reminiscent of heuristics, or quick rules that eliminate the need to perform sophisticated trade-offs between items. We therefore hypothesize that the use of heuristics may potentially explain cluster 1's extremely high level of rationality in treatment **S**.

With this idea in mind, we construct two heuristics for subjects in clusters 1 and 2: heuristic $O$ where the subject maximizes the quantity of *one* of the items and heuristic $T$ where the subject maximizes the *total* quantity in the bundle. We assign type $O$ or $T$ to a subject if (i) it generates the same number of misclassifications as RUM and (ii) this number is smaller than 3 in treatment **S** and smaller than 10 in treatment **C**. These arbitrary thresholds are simply meant to reflect the nature of a heuristic: a quick and simple rule that can be implemented with "few" errors. Otherwise, we assign type $RUM$ to the subject (even though we know that RUM may not always be an accurate representation of behavior). In other words, we divide the sample into subject who employ simple heuristics ($O$ and $T$) and subjects whose choice cannot be summarized by a quick rule, in which case we *assume* that they act as if they trade-off values and quantities ($RUM$). Table 4 summarizes the percentage of each type of subject in clusters 1 and 2. For $RUM$ types, we also add in parentheses the average number of violations (violations are typically small for heuristic users, so the numbers are omitted).

All but one subject in cluster 1 are heuristic users in treatment **S**, mostly using heuristic $O$. More than half of these subjects change their strategy in treatment **C** and are then best classified as $RUM$. Subjects in cluster 2 use heuristics considerably less than subjects

| | Cluster 1 | | | Cluster 2 | | |
|---|---|---|---|---|---|---|
| | *O* | *T* | *RUM* | *O* | *T* | *RUM* |
| Treatment **S** | 27 | 5 | 1 (0) | 1 | 10 | 34 (3.9) |
| Treatment **C** | 10 | 4 | 19 (46.7) | 7 | 3 | 35 (21.4) |

**Table 4:** Heuristic and non-heuristic users

in cluster 1 and there is no treatment effect. Notice also that in treatment **S**, the *RUM* subjects of cluster 2 incur very few violations, resulting in a similar average number of violations in both clusters (row 5 of Table 3). By contrast, in treatment **C**, violations of *RUM* subjects are substantially higher for cluster 1 than for cluster 2, which explains the differences in violations across clusters (row 6 of Table 3). Finally, it is interesting to see such sharp difference of heuristic usage across clusters even though subjects are *not* grouped based on that dimension.

Intuitively, simple rules are more natural in treatment **S**, where the same goods are offered in both bundles: if one good is strongly preferred, the subject can (lexicographically) settle for it; if both goods are of similar value, the subject can focus on total quantities. In treatment **C**, subjects are forced to compare "apple to oranges" so heuristics are less intuitive.[28] Subject are more likely to explicitly trade-off the different alternatives, which explains why more of them are better classified as *RUM*. Finally, since trade-offs are difficult, *RUM* types have more violations than either *O* or *T* types. These behavioral differences are consistent with the working memory differences illustrated earlier.

Overall, the individual analysis reveals interesting insights regarding the different strategies of our subjects. For one group of subjects (cluster 3), the RUM provides a poor fit. These individuals perform badly in both treatments of the consistency task. For two groups of subjects (clusters 1 and 2), the RUM works reasonably well, although differently across treatments. Cluster 1 is mostly composed of OA who use a heuristic in treatment **S** (maximize the amount of the preferred good), resulting in fully rational behavior. Their consistency decreases substantially in treatment **C** due to the extra difficulty to use a heuristic.[29] Cluster 2 is mostly composed of YA who use heuristics less frequently but perform better value-quantity tradeoffs. These subjects make more consistency mistakes than heuristics users in **S** but less in **C**. More generally, the result is consistent with the version of Hypothesis 3 where some subjects who are aware of their

[28]Interestingly, the majority of subjects make significantly more violations when one specific item is common, which suggests that trade-offs are more or less difficult depending on the composition of bundles.

[29]The higher propensity of heuristic usage by OA is also in accordance with the existing literature (Thornton and Dumke, 2005; Mata et al., 2007; Peters et al., 2007; Besedes et al. (2012a, 2012b)).

compromised working memory and fluid intelligence (mostly OA) use simple rules. Such strategy can be applied in the simple treatment but not in the complex one.

# 6 Conclusion

In this paper we have studied choice consistency of younger and older adults in simple and complex domains. We have highlighted several differences in behavior across ages. Older adults are less consistent than younger adults but only when the choice task is complex. We can trace the differences in consistency in the complex task to deficiencies of working memory, that is, in the ability to store and retrieve information regarding the value of the different bundles. Finally, consistency across ages is similar in the simple task partly because older adults use simple rules of choice, such as maximizing the quantity of one of the items in the bundle.

The importance of working memory in ensuring choice consistency is a key result of the paper with fundamental medical and policy implications. Our experimental design, characterized by two bundles presented in a screen, a left-right choice and the possibility of multiple repetitions (see Figure 2) is suitable to be implemented in the scanner. In future research, we plan to use fMRI techniques to study the neural correlates of choice consistency. We already know that simple choices between items involve the ventromedial prefrontal cortex (Hare et al., 2008; Hare et al., 2009) that represents the value difference between options. Our objective is to study how the working memory system (which involves the dorsolateral prefrontal cortex) and the ventromedial prefrontal cortex interact to produce consistent choices, and why this interaction differs across ages.

# References


1. Afriat, S. N. (1967). The Construction of Utility Functions from Expenditure Data. *International Economic Review*, 8 (1): 67-77.

2. Albert, S. M., and Duffy, J. (2012). Differences in risk aversion between young and older adults. *Neuroscience and Neuroeconomics*, 3-9.

3. Ameriks, J., Caplin, A., Leahy, J. and Tyler, T. (2007). Measuring Self-Control Problems. *The American Economic Review, 97*(3), 966-972.

4. Andreoni, J., and Harbaugh, W. (2009). Unexpected utility: Experimental tests of five key questions about preferences over risk. *Mimeo, University of Oregon.*

5. Andreoni, J., and Miller, J. (2002). Giving according to GARP: An experimental test of the consistency of preferences for altruism. *Econometrica, 70*(2), 737-753.

6. Andreoni, J., and Sprenger, C. (2012a). Estimating time preferences from convex budgets. *The American Economic Review, 102*(7), 3333-3356.

7. Andreoni, J., and Sprenger, C. (2012b). Risk preferences are not time preferences. *The American Economic Review, 102*(7), 3357-3376.

8. Baker, S. C., Frith, C. D., Frackowiak, S. J., and Dolan, R. J. (1996). Active representation of shape and spatial location in man. *Cerebral Cortex, 6*(4), 612-619.

9. Battalio, R. C., Kagel, J. H., Winkler, R. C., Fisher, E. B., Basmann, R. L., and Krasner, L. (1973). A test of consumer demand theory using obersvations of individual consumer purchases. *Economic Inquiry, 11*(4), 411-428.

10. Bellemare, C., Kroger, S., and Van Soest, A. (2008). Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. *Econometrica, 76*(4), 815-839.

11. Bernheim, B.D., and A. Rangel (2009). Beyond Revealed Preference: Choice-Theoretic Foundations for Behavioral Welfare Economics. *The Quarterly Journal of Economics, 124*(1), 51-104.

12. Besedes, T., Deck, C., Sarangi, S., and Shor, M. (2012a). Age effects and heuristics in decision making. *Review of Economics and Statistics, 94*(2), 580-595.

13. Besedes, T., Deck, C., Sarangi, S., and Shor, M. (2012b). Decision-making strategies and performance among seniors *Journal of Economic Behavior & Organization, 81* 524-533.

14. Bradbury, H., and Nelson, T. M. (1974). Transitivity and the patterns of children's preferences. *Developmental Psychology, 10*(1), 55.

15. Brand, M., and Markowitsch, H. J. (2010). Aging and decision-making: a neurocognitive perspective. *Gerontology, 56*(3), 319-324.

16. Braver, T. S., Cohen, J. D., Nystrom, L. E., Jonides, J., Smith, E. E., and Noll, D. C. (1997). A parametric study of prefrontal cortex involvement in human working memory. *Neuroimage, 5*(1), 49-62.

17. Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica*, 693-698.

18. Burghart, D. R., Glimcher, P. W., and Lazzaro, S. C. (2013). An expected utility maximizer walks into a bar... *Journal of Risk and Uncertainty, 46*(3), 215-246.

19. Cappelen, A. W., Kariv, S., Sorensen, E., and Tungodden, B. (2014). Is There a Development Gap in Rationality? *Mimeo, Norwegian School of Economics*.

20. Carlson, S., Martinkauppi, S., Rm, P., Salli, E., Korvenoja, A., and Aronen, H. J. (1998). Distribution of cortical activation during visuospatial n-back tasks as revealed by functional magnetic resonance imaging. *Cerebral Cortex, 8*(8), 743-752.

21. Carstensen, L. L., and Mikels, J. A. (2005). At the intersection of emotion and cognition aging and the positivity effect. *Current Directions in Psychological Science, 14*(3), 117-121.

22. Castillo, M., Dickinson, D. L., and Petrie, R. (2014). Sleepiness, Choice Consistency, and Risk Preferences. *Mimeo, Institute for the Study of Labor (IZA)*.

23. Castle, E., Eisenberger, N. I., Seeman, T. E., Moons, W. G., Boggero, I. A., Grinblatt, M. S., and Taylor, S. E. (2012). Neural and behavioral bases of age differences in perceptions of trust. *Proceedings of the National Academy of Sciences, 109*(51), 20848-20852.

24. Chao, L. W., Szrek, H., Pereira, N. S., and Pauly, M. V. (2009). Time preference and its relationship with age, health, and survival probability. *Judgment and Decision Making, 4*(1), 1.

25. Charness, G., and Villeval, M. C. (2009). Cooperation and Competition in Intergenerational Experiments in the Field and the Laboratory. *The American Economic Review, 99*(3), 956-978.

26. Choi, S., Fisman, R., Gale, D., and Kariv, S. (2007). Consistency and heterogeneity of individual behavior under uncertainty. *American Economic Review, 97*(5), 1921-1938.

27. Choi, S., Kariv, S., Muller, W., and Silverman, D. (2014). Who Is (More) Rational? *American Economic Review, 104*(6), 1518-1550.

28. Christoff, K., Prabhakaran, V., Dorfman, J., Zhao, Z., Kroger, J. K., Holyoak, K. J., and Gabrieli, J. D. (2001). Rostrolateral prefrontal cortex involvement in relational integration during reasoning. *Neuroimage, 14* (5), 1136-1149.

29. Cohen, J.D., Perlstein, W. M., Braver, T. S., Nystrom, L. E., Noll, D. C., Jonides, J., and Smith, E. E. (1997). Temporal dynamics of brain activation during a working memory task. *Nature*, 386, 604-607.

30. Cox, J.C. (1997). On Testing the Utility Hypothesis. *The Economic Journal, 107*(443), 1054-1078.

31. Dean, M., and Martin, D. (2013). Measuring Rationality with the Minimum Cost of Revealed Preference Violations. *Mimeo, Brown University.*

32. Demb, J. B., Desmond, J. E., Wagner, A. D., Vaidya, C. J., Glover, G. H., and Gabrieli, J. D. (1995). Semantic encoding and retrieval in the left inferior prefrontal cortex: a functional MRI study of task difficulty and process specificity. *Journal of Neuroscience, 15*(9), 5870-5878.

33. Dove, A., Pollmann, S., Schubert, T., Wiggins, C. J., and Yves von Cramon, D. (2000). Prefrontal cortex activation in task switching: an event-related fMRI study. *Cognitive brain research, 9*(1), 103-109.

34. Dror, I. E., Katona, M., and Mungur, K. (1998). Age differences in decision making: To take a risk or not? *Gerontology, 44*(2), 67-71.

35. Echenique, F., Lee, S., and Shum, M. (2011). The money pump as a measure of revealed preference violations. *Journal of Political Economy, 119*(6), 1201-1223.

36. Engel, C. (2011). Dictator games: a meta study. *Experimental Economics,14*(4), 583-610.

37. Engle, R. W., Tuholski, S. W., Laughlin, J. E., and Conway, A. R. (1999). Working memory, short-term memory, and general fluid intelligence: a latent-variable approach. *Journal of Experimental Psychology: General, 128*(3), 309.

38. Fehr, E., Fischbacher, U., Von Rosenbladt, B., Schupp, J., and Wagner, G. G. (2003). A nation-wide laboratory: Examining trust and trustworthiness by integrating behavioral experiments into representative surveys, *IZA Discussion paper series.*

39. Fevrier, P., and Visser, M. (2004). A study of consumer behavior using laboratory data. *Experimental economics, 7*(1), 93-114.

40. Finucane, M. L., Mertz, C. K., Slovic, P., and Schmidt, E. S. (2005). Task complexity and older adults' decision-making competence. *Psychology and aging, 20*(1), 71.

41. Finucane, M. L., Slovic, P., Hibbard, J. H., Peters, E., Mertz, C. K., and MacGregor, D. G. (2002). Aging and decision-making competence: an analysis of comprehension and consistency skills in older versus younger adults considering health-plan options. *Journal of Behavioral Decision Making, 15*(2), 141-164.

42. Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 1858-1876.

43. Fraley, C., and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association, 97*(458), 611-631.

44. Fraley, C., and Raftery, A. E. (2006). MCLUST version 3: an R package for normal mixture modeling and model-based clustering. *Mimeo, University of Washington.*

45. Frith, C. D., Friston, K. J., Liddle, P. F., and Frackowiak, R. S. J. (1991). A PET study of word finding. *Neuropsychologia, 29*(12), 1137-1148.

46. Geary, D. C. (2005). *The Origin of Mind: Evolution of Brain, Cognition, and General Intelligence.* American Psychological Association.

47. Grady, C. L., Springer, M., Hongwanishkul, D., McIntosh, A., and Winocur, G. (2006). Age-related changes in brain activity across the adult lifespan. *Journal of Cognitive Neuroscience, 18*(2), 227-241.

48. Gray, J. R., Chabris, C. F., and Braver, T. S. (2003). Neural mechanisms of general fluid intelligence. *Nature Neuroscience, 6*(3), 316-322.

49. Harbaugh, W. T., Krause, K., and Berry, T. R. (2001). GARP for Kids: On the Development of Rational Choice Behavior. *American Economic Review, 91*(5), 1539-1545.

50. Hare, T. , O'Doherty, J. , Camerer, C. , Schultz, W., and A. Rangel (2008) "Dissociating the Role of the Orbitofrontal Cortex and the Striatum in the Computation of Goal Values and Prediction Errors", *The Journal of Neuroscience* 28(22), 5623-5630.

51. Hare, T. A., Camerer, C. F., and Rangel, A. (2009). Self-control in decision-making involves modulation of the vmPFC valuation system. *Science, 324*(5927), 646-648.

52. Harrison, G. W., Lau, M. I., and Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review*, 1606-1617.

53. Henninger, D. E., Madden, D. J., and Huettel, S. A. (2010). Processing speed and memory mediate age-related differences in decision making. *Psychology and aging, 25*(2), 262.

54. Holm, H., and Nystedt, P. (2005). Intra-generational trust?a semi-experimental study of trust among different generations. *Journal of Economic Behavior and Organization, 58*(3), 403-419.

55. Horn, J. L., and Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. *Acta psychologica, 26,* 107-129.

56. Houthakker, H.S. (1950). Revealed Preference and the Utility Function. *Economica, 17,* 159-174.

57. Houtman, M., and Maks, J. (1985). Determining all maximal data subsets consistent with revealed preference. *Kwantitatieve methoden, 19*, 89-104.

58. Jaeggi, S. M., Buschkuehl, M., Jonides, J., and Perrig, W. J. (2008). Improving fluid intelligence with training on working memory. *Proceedings of the National Academy of Sciences, 105*(19), 6829-6833.

59. Jahanshahi, M., Dirnberger, G., Fuller, R., and Frith, C. D. (2000). The role of the dorsolateral prefrontal cortex in random number generation: a study with positron emission tomography. *Neuroimage, 12*(6), 713-725.

60. Kane, M. J., and Engle, R. W. (2002). The role of prefrontal cortex in working-memory capacity, executive attention, and general fluid intelligence: An individual-differences perspective. *Psychonomic bulletin and review, 9*(4), 637-671.

61. Kaufman, A. S., and Horn, J. L. (1996). Age changes on tests of fluid and crystallized ability for women and men on the Kaufman Adolescent and Adult Intelligence Test (KAIT) at ages 17-94 years. *Archives of clinical neuropsychology, 11*(2), 97-121.

62. Kim, S., and Hasher, L. (2005). The attraction effect in decision making: Superior performance by older adults. *The Quarterly Journal of Experimental Psychology Section A, 58*(1), 120-133.

63. Kovalchik, S., Camerer, C. F., Grether, D. M., Plott, C. R., and Allman, J. M. (2005). Aging and decision making: A comparison between neurologically healthy elderly and young individuals. *Journal of Economic Behavior and Organization, 58*(1), 79-94.

64. Kroger, J. K., Sabb, F. W., Fales, C. L., Bookheimer, S. Y., Cohen, M. S., and Holyoak, K. J. (2002). Recruitment of anterior dorsolateral prefrontal cortex in human reasoning: a parametric study of relational complexity. *Cerebral Cortex, 12*(5), 477-485.

65. Lewandowsky, S., Oberauer, K., Yang, L. X., and Ecker, U. K. (2010). A working memory test battery for MATLAB. *Behavior Research Methods, 42*(2), 571-585.

66. Lichtenstein, S., and Slovic, P. (1973). Response-induced reversals of preference in gambling: An extended replication in Las Vegas. *Journal of Experimental Psychology, 101*(1), 16.

67. Mata, R., Josef, A. K., Samanez-Larkin, G. R., and Hertwig, R. (2011). Age differences in risky choice: a meta-analysis. *Annals of the New York Academy of Sciences, 1235*(1), 18-29.

68. Mata, R., Schooler, L. J., and Rieskamp, J. (2007). The aging decision maker: cognitive aging and the adaptive selection of decision strategies. *Psychology and aging, 22*(4), 796.

69. Mather, M., and Carstensen, L. L. (2005). Aging and motivated cognition: The positivity effect in attention and memory. *Trends in Cognitive Sciences, 9*(10), 496-502.

70. Mather, M., Mazar, N., Gorlick, M. A., Lighthall, N. R., Burgeno, J., Schoeke, A., and Ariely, D. (2012). Risk preferences and aging: The "certainty effect" in older adults' decision making. *Psychology and aging, 27*(4), 801.

71. Mattei, A. (2000). Full-scale real tests of consumer behavior using experimental data. *Journal of Economic Behavior and Organization, 43*(4), 487-497.

72. Mohr, P. N., Li, S. C., and Heekeren, H. R. (2010). Neuroeconomics and aging: neuromodulation of economic decision making in old age. *Neuroscience and Biobehavioral Reviews, 34*(5), 678-688.

73. Nielsen, L., and Mather, M. (2011). Emerging perspectives in social neuroscience and neuroeconomics of aging. *Social cognitive and affective neuroscience, 6*(2), 149-164.

74. Olesen, P. J., Westerberg, H., and Klingberg, T. (2004). Increased prefrontal and parietal activity after training of working memory. *Nature neuroscience, 7*(1), 75-79.

75. Peters, E., Hess, T. M., Vstfjll, D., and Auman, C. (2007). Adult age differences in dual information processes: Implications for the role of affective and deliberative processes in older adults' decision making. *Perspectives on Psychological Science, 2*(1), 1-23.

76. Prabhakaran, V., Smith, J. A., Desmond, J. E., Glover, G. H., and Gabrieli, J. D. (1997). Neural substrates of fluid reasoning: an fMRI study of neocortical activation during performance of the Raven's Progressive Matrices Test. *Cognitive psychology, 33*(1), 43-63.

77. Rangel, A., and Clithero, J. (2013). The computation of stimulus values in simple choice. *Neuroeconomics: decision making and the brain.*

78. Raven, J., Raven, J.C., and Court, J.H. (1998). *Manual for Raven's progressive matrices and vocabulary scales.*

79. Raz, N., Lindenberger, U., Rodrigue, K. M., Kennedy, K. M., Head, D., Williamson, A., Dable, C., Gerstorf, D., and Acker, J. D. (2005). Regional brain changes in aging healthy adults: general trends, individual differences and modifiers. *Cerebral Cortex, 15*(11), 1676-1689.

80. Read, D., and Read, N. L. (2004). Time discounting over the lifespan. *Organizational behavior and human decision processes, 94*(1), 22-32.

81. Resnick, S. M., Pham, D. L., Kraut, M. A., Zonderman, A. B., and Davatzikos, C. (2003). Longitudinal magnetic resonance imaging studies of older adults: a shrinking brain. *The Journal of Neuroscience, 23*(8), 3295-3301.

82. Samuelson, P.A. (1938). A Note on the Pure Theory of Consumer Behavior. *Economica, 5,* 61-71.

83. Sippel, R. (1997). An Experiment on the Pure Theory of Consumer's Behaviour. *The Economic Journal, 107*(444), 1431-1444.

84. Sutter, M., and Kocher, M. G. (2007). Trust and trustworthiness across different age groups. *Games and Economic Behavior, 59*(2), 364-382.

85. Tentori, K., Osherson, D., Hasher, L., and May, C. (2001). Wisdom and aging: Irrational preferences in college students but not older adults. *Cognition, 81*(3), B87-B96.

86. Thornton, W. J., and Dumke, H. A. (2005). Age differences in everyday problem-solving and decision-making effectiveness: a meta-analytic review. *Psychology and aging, 20*(1), 85-99.

87. Varian, H.R. (1982). The Nonparametric Approach to Demand Analysis. *Econometrica, 50*(4): 945-74.

88. Varian, H.R. (1990). Goodness-of-fit in optimizing models. *Journal of Econometrics, 46*(1), 125-140.

89. Visser, M., Harbaugh, B., and Mocan, N. (2006). An experimental test of criminal behavior among juveniles and young adults. *NBER Working Paper 12507.*

90. Zamarian, L., Sinz, H., Bonatti, E., Gamboz, N., and Delazer, M. (2008). Normal aging affects decisions under ambiguity, but not decisions under risk. *Neuropsychology, 22*(5), 645.

# Appendix (for online publication)

### Appendix A1. List of trials

Denote $\{(q_x, q_y), (q'_x, q'_z)\}$ the trial where a subject chooses between bundle $a_{xy} = (q_x, q_y)$ and bundle $a'_{xz} = (q'_x, q'_z)$, where $y = z$ in treatment **S** and $y \neq z$ in treatment **C**. The list of 35 trials is:

$\{(1,5),(2,1)\}$, $\{(1,5),(3,1)\}$, $\{(1,5),(4,1)\}$, $\{(1,5),(5,1)\}$, $\{(1,5),(2,2)\}$,
$\{(1,5),(3,2)\}$, $\{(1,5),(4,2)\}$, $\{(2,5),(3,1)\}$, $\{(2,5),(4,1)\}$, $\{(2,5),(3,2)\}$,
$\{(1,4),(2,1)\}$, $\{(1,4),(3,1)\}$, $\{(1,4),(4,1)\}$, $\{(1,4),(5,1)\}$, $\{(1,4),(3,2)\}$,
$\{(1,4),(4,2)\}$, $\{(1,4),(5,2)\}$, $\{(2,4),(3,1)\}$, $\{(2,4),(4,1)\}$, $\{(2,4),(5,1)\}$,
$\{(2,4),(3,2)\}$, $\{(1,3),(3,1)\}$, $\{(1,3),(4,1)\}$, $\{(1,3),(5,1)\}$, $\{(1,3),(3,2)\}$,
$\{(1,3),(4,2)\}$, $\{(1,3),(5,2)\}$, $\{(2,3),(3,1)\}$, $\{(2,3),(4,1)\}$, $\{(2,3),(5,1)\}$,
$\{(2,3),(4,2)\}$, $\{(2,3),(5,2)\}$, $\{(1,2),(4,1)\}$, $\{(1,2),(5,1)\}$, $\{(2,2),(5,1)\}$.

### Appendix A2. Food items (with portions) used in the experiment

Almond (2); Barbecue popped potato chip (1); Cashew (2); Cheddar cracker (2); Mini cheese sandwich cracker (1); Citrus gum drop (2); Roasted gorgonzola cracker (2); Fruit gum candy, "Gummy bears" (2); Popcorn (2); Chocolate candy, "M&Ms" (2); Dark chocolate peanut butter cup (1); Mini chocolate-covered pretzel (1); Mini sandwich cookie, "Mini Oreo" (1); Onion-flavored corn snack, "Funyuns" (1); Peanut (3); Pistachio (2); Potato chip (1); Mini pretzel (1); Pretzel nugget (2); Sweet potato chip (1); Yogurt-covered raisin (1);

**Instructions**

**PART 1 - Prep and Introduction (10-15 minutes):**

EXP 1 and EXP 2: *Prepare computers, label seats, and have ready a list of confirmed subjects. Have ready consent forms with Items Sheet attached to front. Place a pen at each table. Lay out one serving of each type of food on the counter in the waiting area. The food items should be labeled both by their name and by the image that will represent them during the experiment.*

EXP 1: *Call in subjects one at a time and check their IDs.*

EXP 1: "Hello and welcome. Before we start we need to ask you when your last meal was. When did you last eat or drink something besides water?"

EXP 1: *Wait for response. If last meal was less than three hours ago, thank them for coming and explain that they cannot participate due to their noncompliance to pre-experiment instructions. If last meal was at least three hours ago, proceed.*

EXP 1: "Today, you will be making choices between bundles of different foods. We want to make sure you like the food items between which you are deciding. Please take some time to look at the different items laid out here and think of which ***five*** you like the most. Keep in mind that you ***may*** be consuming some of these foods together in different amounts. The images you see above each food will be the ones you see during the experiment. This is **NOT** part of the experiment -- please pick the food items that you are most interested in eating."

EXP 1: *Give subject a few minutes to survey the foods.*

EXP 1: "Have you chosen your five items?"

EXP 1: *In the case that they have not chosen their items, wait another couple of minutes. Otherwise, continue.*

EXP 1: "Please let me know which items you have chosen. Would you enjoy eating any combination of these items? Again, this is **NOT** part of the experiment but you may be consuming some of these foods together so we want to be sure that you like them."

EXP 1: *After ensuring their choices are indeed desirable in combination with one another, write item names on subject's Items Sheet.*

EXP 1: "Attached to this sheet is a consent form. As you wait for the experiment to begin, please read the form and sign the last page to consent."

EXP 1: *Direct subject to their seat.*

EXP 1: *Repeat above steps until all subjects have been seated.*

EXP 2: *Modify subject's MATLAB code to ensure only chosen items will be displayed during the experiment.*

EXP 2: "Please make sure your phone is off or on silent mode and do not touch anything as you wait for further instructions."

*After all subjects have been seated and their MATLAB code modified...*

EXP 1: "Dear participants: hello and thank you for coming to this experiment. Today, you will be making choices between bundles of different food items that you like. After you have made your choices, you will complete two short tests and a questionnaire. You will receive food at the end, based on your responses during the experiment. More specifically, one of the choices you make during the experiment will be randomly selected, and, at the end you will receive the amount of food represented in that choice. So, make every choice today as if it were the **ONLY** choice you were making. For example, if your choice of "three chips and two cookies" is randomly selected, at the end of the experiment this is exactly what you will be receiving -- and, eating. You will be given fifteen minutes after the experiment to eat what you receive. You are asked to stay in the waiting area for the whole fifteen minutes. During that time you will have to consume your food items and nothing else. Water will be provided upon request. After that, you will be paid $20 in cash for your participation. You may leave at any time during the experiment, but if you leave before the end, you will not receive the full compensation.

Before each part of the experiment, I will be giving you brief instructions. You can ask questions during these times. "

**PART 2 - GARP Task (15-20 minutes):**

EXP 1: "Now, you will be choosing between different combinations of food items displayed on your computer."

EXP 1: *Show sample screenshot.*

EXP 1: "Here is a sample of what your screen may look like. This is a screenshot for someone that had chosen - *say what the items are* - in the beginning. The ***only*** foods you will see on your screen are the ones you chose in the beginning. Similar to here, you will always have a choice between two combinations: one shown on the right side of the screen, and one shown on the left. If you like the combination shown on the *right* side more, tap the right side of the computer. If you like the combination on the *left* side of the screen more, tap the left side. You cannot tap both sides at once. Remember to make every choice as if it were the **ONLY** one that counted because you will be receiving exactly one of your choices at the end. For example, if I were to tap the left side, there is a chance that I will receive and eat - *say what the foods and quantities of each food are* - at the end.

The experiment is broken down into four parts. You will be making about 35 such choices in each part. When you are done with each part, a screen that reads 'Break' will appear. Please do not touch your screen at that time, but wait for instructions from me to proceed. We will always wait for everyone to finish a part before moving on.

Raise your hand if you have any questions now."

EXP 1: *Look around for raised hands and answer any questions that may arise.*

EXP 1: "Let us proceed with Part 1 of the experiment. Remember, when you are done with this part, a screen that reads 'Break' will appear. Do not press anything but wait for further instruction from me at that point. Remember to make every choice as if it were the **ONLY** one that counted

because you will be receiving exactly one of your choices at the end. Tap the screen to begin the experiment."

EXP 1: *Wait until everyone has completed Part 1. Wait 30 seconds after the last person has finished.*

EXP 1: "Now we will move on to Part 2. As before, tap the side of the screen displaying the combination you like more. When you are done with this part, a screen that reads 'Break' will appear. Do not press anything but wait for further instruction from me at that point. Remember to make every choice as if it were the **ONLY** one that counted because you will be receiving exactly one of your choices at the end. Tap the screen to begin."

EXP 1: *Wait until everyone has completed Part 2. Wait 30 seconds after the last person has finished.*

EXP 1: "Now we will move on to Part 3. As before, tap the side of the screen displaying the combination you like more. When you are done with this part, a screen that reads 'Break' will appear. Do not press anything but wait for further instruction from me at that point. Remember to make every choice as if it were the **ONLY** one that counted because you will be receiving exactly one of your choices at the end. Tap the screen to begin."

EXP 1: *Wait until everyone has completed Part 3. Wait 30 seconds after the last person has finished.*

EXP 1: "Now we will move on to Part 4. As before, tap the side of the screen displaying the combination you like more. When you are done with this part, a screen that reads 'Break' will appear. Do not press anything but wait for further instruction from me at that point. Remember to make every choice as if it were the **ONLY** one that counted because you will be receiving exactly one of your choices at the end. Tap the screen to begin."

EXP 1: *Wait until everyone has completed Part 4.*

**PART 3 - Working Memory Test (10-15 minutes):**

EXP 1: "You are done with the decision-making portion of the experiment. We will now begin the first test. This test is designed to measure your short-term memory abilities."

EXP 1: *Show a sample image of the 10-by-10 matrix they will be seeing during the experiment.*

EXP 1: "During the test you will see a 10-by-10 checkerboard as shown here. Solid black dots will appear, and quickly thereafter, disappear, in some of the spaces. You will see anywhere between two to six black dots appear and disappear in succession. After a short time, the entire checkerboard will disappear, and in its place, an empty checkerboard will appear. You are to tap the spaces of the empty checkerboard where you remember the dots to have been. In this test, it is not important that you accurately recall the positions of the dots; it is more important that you remember the relative positions of the dots. For example, if three dots appeared, one in the top center, one on the bottom right, and one on the bottom left, it would be more beneficial to recall the triangular pattern and recreate it to the best of your abilities, than to accurately remember the position of one of the dots of that triangle. Also, you do not need to remember the order in which the dots appeared - you can tap the spaces of the empty checkerboard in whatever order you like.

If you would like to undo a selection, you can tap the dot to erase it. You will first do two practice trials and then the test will begin.

Are there any questions?

Let us begin the practice trials. Please tap the screen to begin."

EXP 1: *Wait for all subjects to complete practice trials.*

EXP 1: "Are there any questions about this test?"

EXP 1: *Look around for raised hands and answer any questions that may arise.*

EXP 1: "You may begin now."

EXP 2: *Once all subjects have completed the test, collect tablets from subjects and begin preparing their rewards.*

**PART 4 - IQ Test (15-20 minutes):**

EXP 1: "This next part is a test of perception and clear thinking. We will first do two practice problems to familiarize you with the format of the test and method of thought required.

The top part of the first sample problem is a pattern with a bit cut out of it. Look at the pattern, think what the piece needed to complete the pattern correctly both along and down must be like. Then find the right piece out of the eight bits shown below.

Only one of these pieces is perfectly correct. No. 2 completes the pattern correctly going downwards, but is wrong going the other way. No. 1 is correct going along, but is wrong going downward.

Think about which piece is correct both ways.

No. 4 is the right bit, isn't it? So the answer is No. 4, and you select No. 4."

EXP 1: *Check that everyone has selected "4" for the first sample problem.*

EXP 1: "Now turn to the next page and do the second sample problem by yourselves."

EXP 1: *Allow 20 seconds.*

EXP 1: "The answer is No. 8. See that you have selected No. 8. Have you all done that?"

EXP 1: *Check that everyone has selected No. 8.*

EXP 1: "Is everyone clear about what it is you are to do on this test?"

EXP 1: *Answer any questions that subjects may have.*

EXP 1: "You can have as much time as you like for the rest of the test. You will find that the problems soon get difficult. Whether the problems are easy or difficult, you will notice that to solve them you have to use the same method all the time. Keep in mind, it is accurate work that counts. Attempt each problem in turn. Do your best to find the correct piece to complete it before going on the next problem. If you get stuck, you can move on and come back to the problem later. But remember, in every case, the next problem is harder and it will take you longer to check your answers carefully. When you get to the end of the test, please wait for further instructions.

Are there any questions?"

EXP 1: *Pause briefly. Check that everyone is ready to start.*

EXP 1: "You may begin now."

EXP 1: *Wait for all subjects to complete test.*

**PART 5 - Demographic Questionnaire (10-15 minutes):**

EXP 1: "You will now complete a brief questionnaire, which begins on the following page. After you have completed the questionnaire please remain in your seat. Are there any questions?"

EXP 1: *Look around for raised hands and answer any questions that may arise.*

*After subjects have completed questionnaire...*

EXP 1: " The computer has randomly selected one of the bundles you chose today. The other experimenter will now call you one-by-one by your subject ID number. They will hand you your randomly-selected food items. As stated earlier, you will be receiving portions that correspond exactly with one of your choices during the experiment.

Once you have received your items, please remain in the waiting area. You may begin to consume your food once received, however you are required to stay in the waiting area for fifteen minutes after the last subject arrives there. Raise your hand if you have any questions now."

EXP 1: *Look around for raised hands and answer any questions that may arise.*

**PART 6 - Consumption (15 minutes):**

EXP 2: *Call the first subject to the waiting area using subjects' number. Give the subject their food items and call the next subject. Repeat until all subjects have received their bundles.*

EXP 2: "You now have fifteen minutes to eat the items you received. You are asked to stay in this room for the whole fifteen minutes. After that period we will pay you the $20 participation fee and you will be free to leave."

EXP 2: *After fifteen minutes, call each subject one-by-one using subject ID numbers and pay subjects their participation fee. Have subjects sign receipt upon receiving their compensation. Thank them and let them know they are free to leave.*