Consistency in Simple vs. Complex Choices by Younger and Older Adults *

Isabelle Brocas
University of Southern California and CEPR

Juan D. Carrillo
University of Southern California and CEPR

T. Dalton Combs
Dopamine Labs

Niree Kodaverdian
Pomona College

October 2018

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 to YA in the simple condition, both in terms of the number and severity of GARP violations. Working memory and IQ scores correlate significantly with consistency levels, but only in the complex treatment. Our findings suggest that the age-related deterioration of neural faculties responsible for working memory and fluid intelligence is an obstacle for consistent decision-making.

Keywords: laboratory experiments, revealed preferences, aging, complexity.

JEL Classification: C91, D11, D12.

*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 thank Cary Deck, Mara Mather, John Monterosso, and participants at the LABEL Experimental Economics Conference (University of Southern California), and the Social Neuroscience retreat (Catalina island, USC) for useful comments. We also thank Rosa Aguirre for helping to organize sessions at OASIS. 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. Address for correspondence: Juan D. Carrillo, Department of Economics, University of Southern California, 3620 S. Vermont Ave., Los Angeles, CA 90089, USA, <juandc@usc.edu>.

1 Introduction

As most day-to-day decisions involve comparing options and making trade-offs between them, understanding how people attribute value to options is crucial in understanding how people make decisions. Economics builds theories under the assumption that individuals have unambiguous values for options 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). By contrast, there is only some empirical evidence regarding consistency in older adults (Choi et al., 2014) and our knowledge of that population is still incomplete.[1] 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 different choices as compared to younger adults (YA) in a variety of domains.[2] Such differences can potentially be due to two very different mechanisms: either *preferences change* with age or preferences remain stable but the *ability to act consistently* on them changes with age. There is indirect evidence for both possibilities. In line with the first prong, aging brings substantial changes in our motivations, which in turn affects the decisions we make (Carstensen and Mikels, 2005; Mather and Carstensen, 2005). At the same time, the aging process affects many brain structures and brain mechanisms, hindering the ability to evaluate alternatives and select among options (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). We aim to resolve this problem by controling for differences in preferences and testing those preferences for consistency.

In this paper, we propose to use the Generalized Axiom of Revealed Preference (GARP) to test the internal consistency of the preferences of YA and OA by offering repeated choices

[1] The study by Choi et al., 2014 reports evidence from a large sample of adults of all ages and does not provide a balanced comparison between old and young. The study also focuses on consistency in the risk domain.

[2] See e.g. Fehr et al., 2003; Ameriks et al., 2007; Bellemare et al., 2008; Engel, 2011; Albert and Duffy, 2012; Castle et al., 2012. 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). Yet others find curvilinear age effects (Harrison et al., 2002; Read and Read, 2004). Some studies offer to resolve these mixed findings, by arguing that results are highly sensitive to differences in the learning requirements (Mata et al., 2011), the completeness of information (Zamarian et al., 2008), the number of options to choose between (Brand and Markowitsch, 2010) and the contents of choice sets (Mather et al., 2012).

between bundles of goods. Additionally, we vary the complexity of the task by changing the number of unique goods that are present in a choice. Our goal is to understand consistency at different ages and as a function of the *complexity* of the situation.

More specifically, we use a controlled laboratory experiment with a 2×2 design, where YA and OA make choices in two different domains: simple and complex. In the simple domain, subjects decide between two bundles each composed of different quantities of the same two goods (e.g., 5 pistachios plus 1 cheese vs. 2 pistachios plus 2 cheese). In the complex domain, subjects choose between two bundles, each also composed of different quantities of two goods, but now with exactly one common good (e.g., 5 pistachios plus 1 cheese vs. 2 pistachios plus 2 crackers). Complexity refers to the number of goods and combinations of goods that need to be evaluated to make a decision. This definition of complexity is motivated by earlier studies in which complex choices were related to the number of items involved in the task (Wright, 1981; Cappell et al., 2010).

Besides the contribution of comparing YA and OA in a simple and a complex domain, our design has three new elements relative to the existing literature (reviewed below). We ask subjects to choose between two bundles presented pictorially. This simplifies the choice problem relative to presenting a large number of bundles (as in Harbaugh et al., 2001) or relative to presenting a budget set on a coordinate plane (as in Choi et al., 2007; Fisman et al., 2007; Choi et al., 2014). We also include trivial trials to our task, where subjects choose between a smaller and a larger quantity of one desirable good. Subjects who fail these trials are likely to violate one or more assumptions of the model; they are inattentive, they do not monotonically value the good over the tested range, and/or they misunderstand the task. This allows us to conduct the consistency analysis both with the full sample and with the subsample of subjects for whom we are most confident the model is appropriate. In addition, our subjects perform a working memory and IQ task. This allows us to study the determinants of consistency. Sample selection issues limit the extent to which causal relationship can be assessed.

Notice that, despite our best efforts to match samples, differences found across age groups may be due to cohort-specific factors and may be unrelated to age.[3] Our analysis will take these limitations into account to draw conclusions. With this caveat in mind, we next summarize the two main findings of our study. First, both OA and YA are reasonably (and roughly equally) consistent in the simple treatment whereas the OA in our sample are significantly more inconsistent than the YA in the complex treatment. This difference across populations applies generally: to the number of total violations, to the number

[3]For instance, these differences could be driven by differences in sociability, experience, opportunity cost of time or income (in particular, the mean household income of our YA sample is greater than that of our OA sample).

of violations by type (direct and indirect) and to the severity of violations (using two different criteria). Surprisingly, a significant fraction of subjects (12% of YA and 33% of OA) fails the trivial trials. This calls into question the reliability and interpretability of the consistency results for those individuals. We then conduct the same analysis with the subsample of subjects who pass the trivial trials. Not surprisingly, the total number of violations is substantially smaller in this subsample. Importantly, however, the treatment effect is identical: marginal differences between OA and YA in the simple domain and significant differences in the complex domain.

Second, we find that differences in violations in the complex treatment correlate 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 more heavily recruited in the complex task than in the simple one. The result echoes the studies reviewed below, which show this precise relationship between complexity and working memory demands. Interestingly, the result also extends to IQ (although less strongly) but it should be noted that working memory and IQ are highly correlated.

Finally, we also conduct an individual and cluster analysis (see the Appendix). One group of subjects is very inconsistent in both the simple and complex treatments. A second group, mostly composed of OA, are individuals who commit almost no violations in the simple treatment. Interestingly, these subjects have a preference that can be implemented with a simple rule: maximize the quantity of the favorite good in the bundle. Their behavior becomes significantly more inconsistent in the complex domain, possibly because that simple rule is less intuitive to implement in that context. The last group, mostly composed of YA, are subjects who do not exhibit preferences that can be implemented with simple rules. They are slightly less consistent than the previous group in the simple treatment but significantly more consistent in the complex one.

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, such as goods (bundles with positive quantities of two or more desirable items), risk (bundles of quantities and probabilities) and social (bundles of money for oneself and money for another party). Studies find that YA make choices generally consistent with revealed preference theory.[4]

[4]See e.g. Sippel (1997), Mattei (2000) and Fevrier and Visser (2004) for studies in the good domain, Choi et al. (2007), Andreoni and Harbaugh (2009) and Choi et al. (2014) for studies in the risk domain and Andreoni and Miller (2002) and Fisman et al. (2007) for studies in the social domain. 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., 2017) subjects. In a cross cultural study, Tanzanian YA are found to commit more GARP violations as compared to YA from the United States (Cappelen

Second, experiments have concurred in the finding that consistency increases between 8 and 12 years old children (Bradbury and Nelson, 1974) and thereafter stabilizes (Harbaugh et al., 2001). By contrast, the full trajectory across the lifetime has not been established. Indeed, some laboratory (Tentori et al., 2001; Kim and Hasher, 2005) and field (Dean and Martin, 2016) experiments find that OA are more consistent than YA while other laboratory experiments (Finucane et al., 2002; Finucane et al., 2005) and panel data studies (Echenique et al., 2011) find the opposite. These disparate findings may be partly due to two methodological choices. First and contrary to standard practices in experimental economics, decisions in those YA vs. OA 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.

Additional support for an inverse relationship between age and consistency can be found in the recent work by Choi et al. (2014). In this comprehensive online study, the authors show that GARP consistency in the risk domain decreases with age and increases with household wealth.[5] The paper combines benefits of field (large sample size) and laboratory (incentivized) experiments. Moreover, subjects are drawn from a sample designed to be representative of the Dutch population. The study, however, does not address the two questions we are interested in, namely (i) how choice inconsistencies depend on the combination of age and task complexity and (ii) whether they can be traced to compromised working memory and fluid intelligence.

Third, studies have demonstrated that task complexity imposes demands on working memory.[6] Working memory is the short-term mental maintenance (Cohen et al., 1997; Curtis and D'Esposito, 2003) and manipulation of information (Pochon et al., 2001), and this process is less efficient in OA. Varying levels of task complexity may account for differences in choice consistency between OA and YA. Indeed, neuroimaging studies have shown that the working memory regions of the brain are recruited during more difficult tasks, such as those requiring task-switching (MacDonald et al., 2000), integral-solving

et al., 2014). Finally, 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 consistent.

[5] A related study by Gaudecker et al. (2011) considers a large sample to study the determinants of risk preferences. The authors use a multiple price list design and report that risk preferences themselves depend on a variety of socioeconomic variables. This complements the finding related to the consistency of preferences per se obtained by Choi et al. (2014).

[6] We want to emphasize that working memory is not what is usually referred informally to as "memory." It is not about remembering choices in past trials. Instead, working memory is a term used in cognitive neuroscience to describe the ability to hold pieces of information in mind for a few seconds to complete a reasoning (in our case, to make a choice in a given trial).

(Krueger et al., 2009), and attention-shifting (Kondo et al., 2004). Crucially, the circuitry is 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; Greene et al., 2004).[7] 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 the age-related atrophy of regions involved in working memory (Raz et al., 2005) could be a main cause of that decline: these regions are activated less in OA as compared to YA in working memory tasks (Rypma and D'Esposito, 2000), especially when the number of items to be maintained (Cappell et al., 2010) or manipulated (Wright, 1981) in memory is high.

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

2 Theoretical background

Consider a subject making choices between pairs of bundles, each with two goods that are assumed to be desirable, in the sense that *more of each good is strictly preferred to less*.[8] A choice between a pair of bundles is called a "trial." Denote $a_{xy} := (q_x^a, q_y^a)$ 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 strictly more quantity of one good and strictly less of the other ($q_x^a > q_x^{a'} \Leftrightarrow q_y^a < q_y^{a'}$). In the experimental section, this is called treatment **S** (for simple). When a trial is considered in isolation, the question of consistency does not arise, and *any* choice between pairs of bundles with the aforementioned properties is consistent with the maximization of monotonic and transitive preferences. However, when we jointly consider a pair of trials, some combinations of choices may constitute a violation of revealed preferences (which we call $\mathcal{D}_{\mathcal{S}}$, for direct

[7] This relationship extends to tasks requiring the explicit representation and manipulation of knowledge, when the ability to reason relationally is essential (Kroger et al., 2002), when the number of dimensions to be considered simultaneously is increased (Christoff et al., 2001), or when the number of objects to maintain in working memory is increased (Gould et al., 2003).

[8] This corresponds to the standard monotonicity assumption in the revealed preference literature.

violation in the simple treatment).[9] 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 to the maximization of monotonic and transitive preferences.

Definition 1 sets conditions for a direct violation in a pair of trials of treatment **S**.

Definition 1 *Direct violation in a pair of trials of the 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 $q_x^{a'} \geq q_x^b$ for all x (with at least one strict inequality) and $q_x^{b'} \geq q_x^a$ for all x (with at least one strict inequality).

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

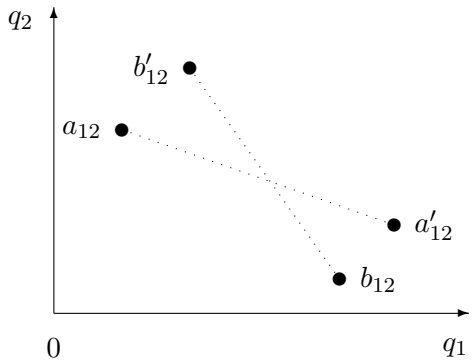

Figure 1: Trials a_{12} vs. a'_{12} and b_{12} vs. b'_{12}

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 options per choice is substantially reduced. Therefore, a choice in one trial only reveals that the selected option, or "bundle," is preferred to the only other bundle proposed rather than to any bundle on the "budget line."[10] Notice that Definition 1 is made of two parts. Part (i) provides conditions such that choices in a pair of trials *may* result in a violation. Intuitively, the requirement is that for each trial one bundle dominates (i.e., has weakly more quantity of both goods and strictly more of at least one than) a bundle in another trial whereas the remaining

[9] The seminal work on revealed preference theory is due to Samuelson (1938). It was subsequently extended by Houthakker (1950), Afriat (1967) and Varian (1982) and more recently by Nishimura et al. (2017) among others.

[10] Obviously, under no satiation, each option is preferred to any bundle that has weakly less quantity of both goods, which is why in the example depicted in Figure 1, we have $b'_{12} \succ a_{12}$ and $a'_{12} \succ b_{12}$.

bundle is dominated by (i.e., has weakly less quantity of both goods and strictly less of at least one 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 is dominated by a bundle in the other trial. In our example, the dominated bundles are a_{12} and b_{12}. Hence, only one out of the four possible choice combinations will result in a $\mathcal{D}_{\mathcal{S}}$-violation.

Given n trials, there are $n(n-1)/2$ pairs of trials. By considering all pairs of trials and checking whether the condition in Definition 1(i) is satisfied, we can identify all possible violations between pairs of trial. Then, actual violations are determined simply by checking whether a subject's selected bundles (in those pairs of trials in which a violation is possible) satisfy the condition in Definition 1(ii).

A precise test of GARP requires checking whether the data satisfies cyclical consistency. The exercise described here, based on pairs of trials, consists in checking that property on the *minimum cycle*. In some cases, the minimum cycle is enough to account for all violations (Banerjee and Murphy, 2006), but in others, longer cycles should be included for an exhaustive analysis. By the discrete nature of our choice problem, minimum cycles are not enough in our case. In other words, it is possible that a direct violation occurs between a triplet of trials (or more) even though the condition in Definition 1(i) is not satisfied by any pair of trials in that triplet (and therefore no direct violation occurs between pairs of trials in the triplet). In Appendix A1, we construct an example of such case. Given our choice of bundles, conditions such that direct violations can occur between triplets of trials – but not between pairs of trials in that triplet – are very rare but still possible. We will not consider them in the analysis, which means that our experimental study may miss some (small number of) GARP violations.

It is also worth noting that it is not possible 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.[11]

[11] 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 $q^c_x < q^a_x$ and $q^{c'}_x > q^{a'}_x$ for all x. By choosing a_{12} over a'_{12} and b_{12} over b'_{12} the subject incurs 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}).

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 as before, bundles are composed of two goods. Consider trials between pairs of 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). As the choice problem involves more goods, the decision is arguably more complicated. Since a trial still has two bundles and each bundle still has positive quantities of exactly two goods, the two treatments remain comparable. By definition, each bundle now has strictly more quantity of at least one good (only a_{xy} has a positive quantity of good y and only a'_{xz} has a positive quantity of good z). Again, when a trial is considered in isolation, any choice between pairs of bundles is consistent with the maximization of monotonic and transitive preferences.

In this new treatment, there are two qualitatively different types of violations: (i) violations that arise among bundles that have one good in common and (ii) violations that arise across all types of bundles. As in the previous section, we will check cyclical consistency on the minimum cycles. If we consider the set of trials that share one good, the minimum cycle involves pairs of trials. This is the analog of direct violations in the previous section. If we consider all types of bundles, the minimum cycle to detect violations involves triplets of trials, each trial with a different good in common.[12]

Definition 2 identifies conditions for a direct violation in a pair of trials of treatment **C** to occur. Given those violations are the analog of $\mathcal{D}_{\mathcal{S}}$-violations, the definitions are very similar to the conditions described in Definition 1.

Definition 2 *Direct violation in a pair of trials of the complex treatment ($\mathcal{D}_{\mathcal{C}}$).*

(i) Trials a_{xy} vs. a'_{xz} and b_{xz} vs. b'_{xy} may involve a $\mathcal{D}_{\mathcal{C}}$-violation if and only if $q_x^{a'} \geq q_x^b$ and $q_z^{a'} \geq q_z^b$ (with at least one strict inequality) and $q_x^{b'} \geq q_x^a$ and $q_y^{b'} \geq q_y^a$ (with at least one strict inequality).

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

Just like in the example of Figure 1, when the conditions of Definition 2(i) and (ii) are satisfied, we get $a_{xy} \succ a'_{xz} \succ b_{xz}$ and $b_{xz} \succ b'_{xy} \succ a_{xy}$ which is a contradiction to the maximization of monotonic and transitive preferences.

Definition 3 describes violations that occur across all types of bundles and involve three trials, each with a different common good. We call them indirect violations.

Definition 3 *Indirect violation in a triplet of trials of the complex treatment ($\mathcal{I}_{\mathcal{C}}$).*

[12] There cannot be violations involving two trials that have the same good in common and one trial that has a different good in common.

(i) Trials a_{xy} vs. a'_{xz}, b_{xz} vs. b'_{yz} and c_{yz} vs. c'_{xy} may involve an $\mathcal{I}_{\mathcal{C}}$-violation if and only if $q_x^{a'} \geq q_x^b$ and $q_z^{a'} \geq q_z^b$ (with at least one strict inequality), $q_y^{b'} \geq q_y^c$ and $q_z^{b'} \geq q_z^c$ (with at least one strict inequality), and $q_x^{c'} \geq q_x^a$ and $q_y^{c'} \geq q_y^a$ (with at least one strict inequality).

(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 remaining bundle is dominated by the bundle composed of the same goods in the other trial. In Definition 3(i) and given that more quantity is always desirable, we have $a' \succ b$, $b' \succ c$, and $c' \succ a$. When this condition is satisfied, an indirect violation occurs if the subject chooses bundles a, b, and c. Indeed, these choices imply $a \succ a' \succ b \succ b' \succ c$ on one hand, and $c \succ c' \succ a$ on the other, which forms a contradiction.

For the same reasons as in the second remark of the simple treatment, in the complex treatment it may be the case that a direct violation involving three or more trials occurs but no violation occurs between any subset of two trials. Similarly, it may be the case that an indirect violation involving four or more trials occurs but no violation occurs between any subset of three trials. For simplicity, we will again ignore those violations.[13]

3 Experimental design and procedures

To study choice consistency of younger adults (YA) and older adults (OA) with different levels of complexity, we conduct an experiment based on the setup described in the theory section using the MatLab extension Psychtoolbox (Brainard, 1997; Pelli, 1997).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 59-89) were recruited through the OASIS activities catalogue.[14] Six subjects were omitted from analysis: four subjects experienced software malfunctioning; one spontaneously reported miscomprehension of the task halfway through the experiment; the only male subject in the pool was excluded to make the sample more demographically homogeneous.

[13]The reader shall note that these difficulties arise because of the discrete nature of our choice set and the specific constraints imposed on the composition of the bundles.

[14]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. Recruitment is mostly word-of-mouth, with existing members referring new members. More information can be found at `http://www.oasisnet.org`.

We therefore retained 45 female OA for the analysis.[15]

OA in our sample are highly educated.[16] Given their education level, we deemed it appropriate to recruit college students for our YA sample.[17] YA subjects were recruited from the Los Angeles Behavioral Economics Laboratory (LABEL) pool, which consists of over 2,500 USC students, and sessions were conducted at LABEL, in the department of Economics at the University of Southern California. In order to match gender, we recruited 50 YA female USC students, age 18-34.[18] All subjects were compensated with a fixed amount of $20 plus an incentive payment (described below).

As discussed in the introduction, the potential selection problem limits our ability to make causal inferences. Experimental evidence regarding differences in behavior between our YA and OA does not prove a causal effect of age. In section 5 we review different channels through which differences in behavior between age groups could arise and address how plausible these alternative explanations are in light of the specific differences we obtain in our study. In particular and as developed below, we find a clear differential effect across task complexity which is consistent with age-related changes. This finding seems inconsistent with the alternative channels we consider, as we would not expect them to discriminate across task complexity.[19]

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 each composed of two goods, and were not allowed to express indifference. There were 35 trials of the simple treatment **S**, where the same two goods $(1, 2)$ appeared in both bundles (a_{12} vs. a'_{12}, b_{12} vs. b'_{12}, etc.). There were also three sets of 35 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

[15]The overwhelming majority of OASIS members are female (88%), which explains the extreme gender selection in our sample but also raises some concerns about self-selection. Besedes et al. (2012b) also report a larger fraction of female participation (75%), although the difference is not as extreme as ours.

[16]The distribution of their highest educational attainment is: PhD (4%), MA (22%), Professional degree (2%), BA (29%), AA (11%), some college credit (26%), and trade/technical/vocational school (4%). This is representative of the OASIS members and substantially above national averages. It is not surprising that an organization dedicated to the sharing of knowledge and promotion of research-based programs attracts individuals with above average levels of education and intellectual curiosity.

[17]All the YA in our sample are USC students. Based on national average statistics (reported by U.S News in 2009), we expect that 26% of undergraduates will pursue a graduate degree. Therefore, education of our OA is comparable to the final education that can be expected for our YA.

[18]For more information about the laboratory, see `http://dornsife.usc.edu/label`. We had 45 undergraduate and 5 master students in our YA sample from all 4 disciplines: Arts and Humanities (10%), Natural Sciences (30%), Social Sciences (50%), and Technical Sciences (10%).

[19]For example, if the self-selected OA had a lower opportunity cost of time than the general OA population, and if opportunity cost of time affects decision consistency, then we would expect the effect to be present for both the simple and complex version of the GARP task.

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}). Importantly, 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 a second bundle that was dominated by a bundle in as many other trials as possible. There are two reasons for this choice. First, to give as many chances as possible to observe $\mathcal{D}_{\mathcal{S}}$-, $\mathcal{D}_{\mathcal{C}}$-, and $\mathcal{I}_{\mathcal{C}}$-violations if subjects were inconsistent. Second, to minimize the chances of violations that are not identified in our analysis (e.g., direct violations between triplets of trials that are not captured with pairs of trials, as explained in the second remark of section 2.1 and Appendix A1).

Figure 2 depicts the 35 trials in treatment **S**. The x- and y-axis represent the quantities of the two goods. Each point represents a bundle of some quantity of good x and some quantity of good y. Each segment corresponds to one trial in which the two bundles it connects were offered against one another. For example, the bold red segment represents a trial in which a bundle of 1x and 5y were offered against a bundle of 2x and 2y. We used the same quantities for trials in treatment **C**, in order to facilitate the comparison of violations across treatments. The only difference is that bundles have only one good in common.[20]

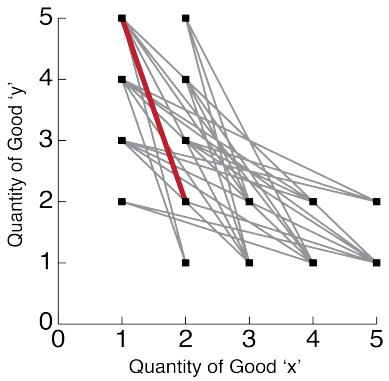

Figure 2: The 35 trials in treatment **S**.

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). Including trivial trials are typical in psychology experiments (under the misleading terminology of "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

[20] If we were to depict it, we would use a three-dimensional graph. Each bundle would then have positive quantities of exactly two goods.

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 that subject's choices in treatments **S** and **C**.

All subjects faced 150 trials: 140 core trials and 10 attention trials, as summarized in Table 1. The 150 trials were intertwined and presented in a randomized and counterbalanced order.

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 popular salty and sweet snacks and we asked each subject to pick five of them for consumption: two were then randomly used in treatment **S** and the other three in treatment **C**. Therefore, each subject completed the task with a personalized set of snacks. 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.[21] 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 3 presents a sample screenshot of a trial in treatment **C**. In this example, the subject had to choose between a bundle of 5 portions of chips plus 1 portion of peanuts and a bundle of 4 portions of chips plus 2 portions of pretzels. At the end of the experiment, one trial was randomly selected for each subject, and the subject's choice in that trial was given to them to consume. Subjects were kept in the experimental room for 15 minutes

[21] 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 might have to consume a combination of two items at the end of the experiment. Appendix A2 presents the list of food items and quantities per portion.

following the end of the experiment. This was to ensure that all the foods would be fully consumed, that they would be consumed by the intended subject, and that they would not be consumed in combination with foods other than those in each subject's bundle. An advantage of using food items is that subjects cannot trade goods at the end of the experiment. Every subject complied with the procedure.

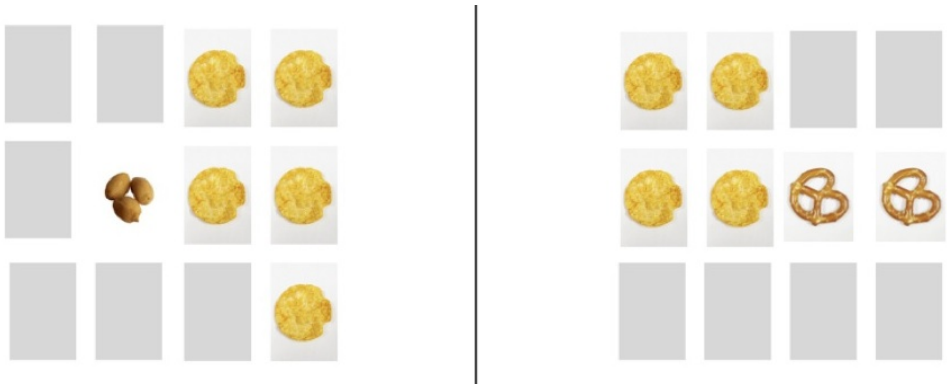

Figure 3: Screenshot of one trial in treatment **C**

Working memory and Raven's IQ tests. After the GARP task, subjects performed a spatial working memory test and an IQ test. To measure 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 10×10 grid. A trial 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. After the final dot in a sequence disappears the subject attempts tap the cells where the dots appeared in any order. Score decreases with the distance between the correct and the selected cells. The entire test consists of 32 such trials, including 2 practice trials. No feedback is given between trials or upon completion of the test. For IQ, we used the short version of Raven's IQ test, namely Set I of 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 one used by the Emotion and Cognition Lab at USC. It includes questions about their highest diploma, occupation, income, ethnicity, various stress rankings and health levels, as well as information relative to current medications.

Summary. From a design viewpoint, there are two new elements relative to the existing experimental tests of revealed preferences. First, we study choice across ages and choices across task complexity but, most importantly, we study the interaction between the two. Second, we correlate choices with measures of memory and fluid intelligence, to better understand the source of differences in consistency over the life cycle. From a methodological viewpoint, there are also two 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 (such as inattention, satiation, disliking or miscomprehension) from those who violate consistency even though they are likely to 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.[22] It also allows to incorporate a complexity component without modifying the dimensions of the task (e.g. number of items on screen and number of choices). A sample copy of the instructions can be found in the Appendix.

4 Analysis

4.1 Frequency of violations

Our first and central objective is to assess choice consistency across populations (YA vs. OA) and treatments (simple vs. complex). 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 $\frac{35\times34}{2} = 595$ pairs of trials, of which 170 can potentially result in direct violations. Therefore, there is a

[22] Our 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)). Choi et al. (2007) also perform many trials thanks to their ingenious software presentation, although the decision problem in their setting is substantially more complex.

total of 170 possible $\mathcal{D}_S$-violations and 510 possible $\mathcal{D}_C$-violations. By contrast, of the $35^3 = 42,875$ triplets of trials in treatment **C**, only 188 can result in an $\mathcal{I}_C$-violation. This means that at most 28.6% of choices can result in direct violations between pair of trials but only 0.4% can result in indirect violations between triplets of trials.[23] A more informative measure to assess the extent of violations across treatments is to compare them with the number of violations incurred by a simulated subject choosing randomly between bundles. Figure 4 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.[24]

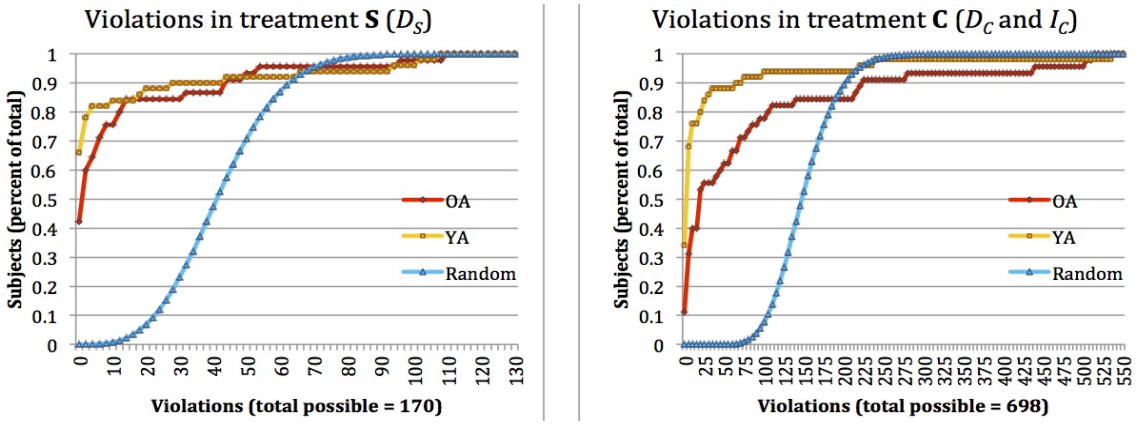

Figure 4: Number of violations in treatments **S** (left) and **C** (right)

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 less than 105 in treatment **C**. Using these numbers as a benchmark, we get instead that 88% of our YA and 84% of our OA incur

[23] Recall the second remark in section 2.1, stating that choices which induce some violations may preclude some others. As such, 170 and 188 are upper-limits on the number of effectively feasible direct violations of pairs ($\mathcal{D}_S$, $\mathcal{D}_C$) and indirect violations of triplets ($\mathcal{I}_C$), respectively.

[24] We use a uniform distribution for this exercise. An alternative was to apply a bootstrap method to generate the empirical distribution of actual choices and use such distribution to simulate random players. However, given our population is composed of two distinct groups and is not a representative sample of the general population, the empirical distribution is not a particularly meaningful tool to draw inferences from.

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), and others), the majority of our subjects incur relatively few violations.

Perhaps more interestingly, we can compare violations across age groups. YA in our sample incur fewer violations than OA in our sample and differences are more pronounced in treatment **C** than in treatment **S**. More precisely, non-parametric Kolmogorov-Smirnov (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 $<$ .001, respectively).[25] As can be seen 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 4 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 the behavior in the tail of the distribution: 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 some assumption of the model. Overall, we find that treatment **C** is more difficult than treatment **S** and generates relatively more mistakes in both populations. Also, the results in this section are consistent with a strong age effect on the number of violations in the complex task and a weak or no age effect in the simple task.

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 4. Figure 5 separates violations in treatment **C** into direct ($\mathcal{D}_{\mathcal{C}}$) and indirect ($\mathcal{I}_{\mathcal{C}}$) for each population. As before, it also represents the distribution of violations under a random choice rule.

According to KS and WRS tests, differences in the distributions between YA and OA in treatment **C** are substantially more pronounced for direct violations (p-value $<$ .001

[25] As it is well-known, KS is sensitive to any difference in distributions (shape, spread, median, etc.) whereas WRS is mostly sensitive to changes in the median. In an attempt to remain agnostic about which test is more appropriate for our sample, we will report results for both tests in all of our comparisons of distributions.

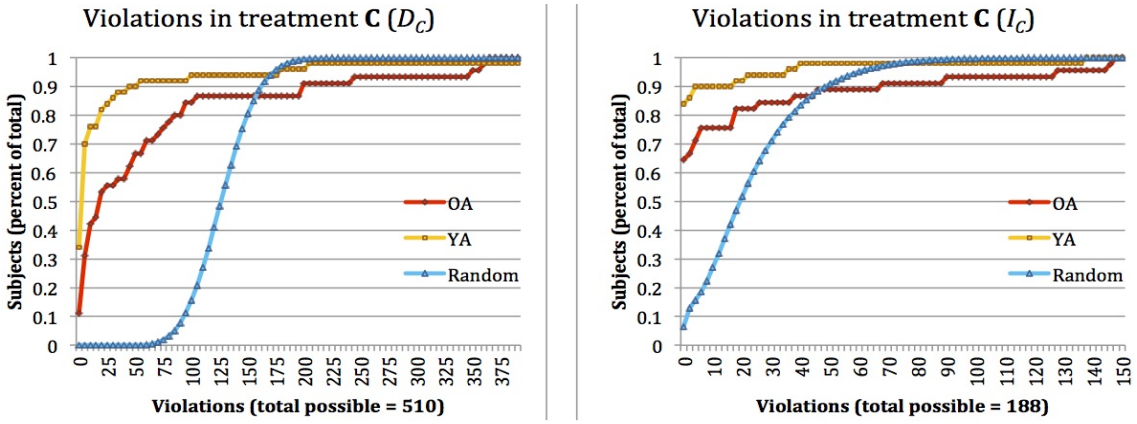

Figure 5: Direct (left) and Indirect (right) violations in treatment **C**

for both) 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 focused on the *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 (2016) among others, one should also take into account the *severity* of violations. Populations may differ in frequency of violations but not in severity and vice-versa.

There are several ways to study severity. One possibility is to consider a severity index that puts a measurement to the intuition that, if the differences in quantities between bundles is small, the violation is less severe than if the differences are large.[26] Recall from Definitions 1, 2, and 3 that the condition for a GARP violation to occur between a pair (direct) or a triplet (indirect) of trials is that for both trials, the chosen bundle must have less quantity of both goods than the bundle not chosen in the other trial. This is independent of how much smaller these quantities are. For example (and, again, assuming

[26] 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 way to quantify their extent.

monotonic preferences) if my choices reveal a preference for (1,1) over (2,2), the violation is less acute than if they reveal a preference for (1,1) over (5,5). This is consistent with the theory behind stochastic choices, whereby rational individuals are more likely to incur smaller mistakes than bigger ones.

To formalize this idea of severity, we take each pair (triplet) of trials involved in a direct (indirect) violation, and measure the euclidean distance between the amounts in the chosen bundles and the amounts in the bundles that have weakly more quantity of both goods and were not chosen. We then take the minimum of these distances, which we call **d**. This value captures the minimum quantity by which we should change one of the choices of the individual in order to remove the violation. It can also be interpreted as the magnitude of the "mistake" incurred by not choosing the bundles with more quantities of both goods.

To illustrate the concept, consider the case of a $\mathcal{D}_{\mathcal{S}}$-violation described in Figure 1. 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 error is small and reversing two very similar choices would remove the violation. For the case of $\mathcal{D}_{\mathcal{C}}$ and $\mathcal{I}_{\mathcal{C}}$-violations in treatment **C**, the severity is given by $\mathbf{d} \equiv \min\left\{d(a_{xy}, b'_{xy}), d(b_{xz}, a'_{xz})\right\}$ and by $\mathbf{d} \equiv \min\left\{d(a_{xy}, c'_{xy}), d(b_{xz}, a'_{xz}), d(c_{yz}, b'_{yz})\right\}$ respectively. Notice that the euclidean distance is always taken between two bundles containing positive quantities of the same two goods.

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 fictitious effect, 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 6 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, 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). In treatment **S** the distribution of severity of violations is not significantly different across 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).[27]

[27] We performed the exact same analysis with the average amount (instead of the minimum amount) choices of an individual should be changed in order to remove the violation. So, for example, in Figure

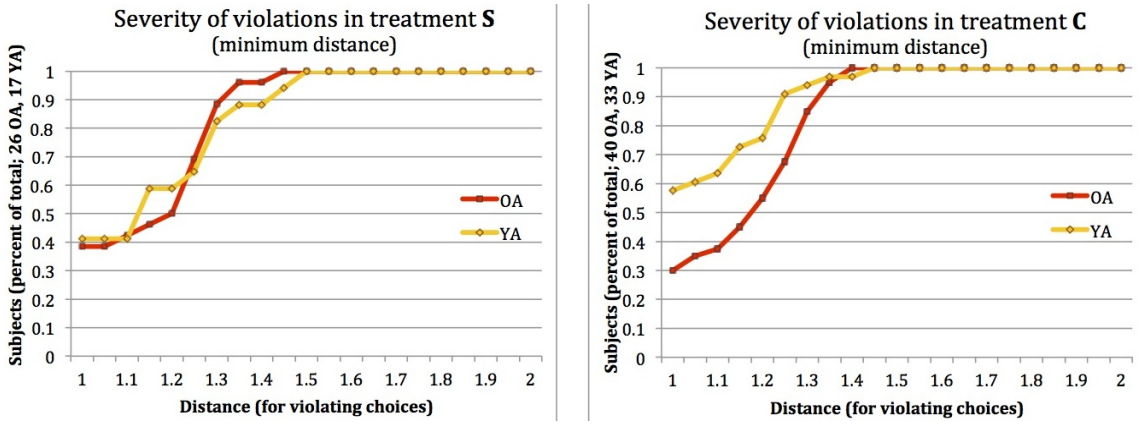

Figure 6: Severity of violations in treatments **S** (left) and **C** (right)

An alternative measure of severity of violations consists of finding, for each individual, the minimum number of trials that need to be removed in order to suppress all violations for that individual. 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.[28] As before, we exclude the individuals with no violations to avoid artificially exacerbating differences between OA and YA. This means that the minimum number of trials to be removed is 1. Figure 7 presents the c.d.f. of the number of choices to be removed for perfect consistency by population and treatment.

This severity measure yields results similar to the previous one. The distribution of the number of choices that need to be removed to achieve consistency is not statistically different between 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 example, 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 OA we need to remove 14 trials. Taken together, the results in this section lend further

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 and the treatment effect sharper than before: still no significant difference between OA and YA in treatment **S** (p-value = .727 and .820 for KS and WRS tests) and significantly more severe violations for OA than YA in treatment **C** (p-value = .023 and .004 for KS and WRS tests). The graphs are omitted for brevity.

[28] 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 does not include any cycles. Intuitively, the index provides a corrected number of violations: two individuals who commit one mistake may end up exhibiting drastically different violation counts depending on how their mistake contradicts other choices. The Houtman-Maks index reflects that both made the same number of mistakes.

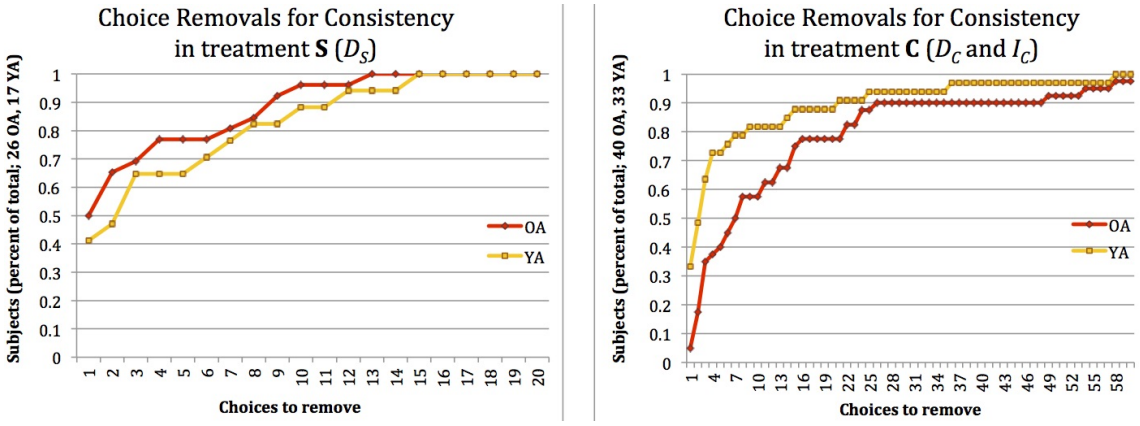

Figure 7: Choices to remove for consistency in treatments **S** (left) and **C** (right)

support to our previous finding: OA in our experiment are marginally more inconsistent than YA in the simple treatment but substantially more inconsistent than YA in the complex treatment, both in terms of the number and the severity of violations. These results are further reinforced by the analysis we conduct in Appendix A.3.1, where we estimate a random utility model for each subject and compute the proportion of misclassified trials. This exercise allows to estimate the rational preference that is closest to the observed data and to measure how many mistakes each subject makes around it. As expected, these mistakes are strongly correlated with GARP violations and severity measures.[29]

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 each good and always prefer more to less – are satisfied. Figure 8 presents the number of violations incurred by YA and OA in the 10 trivial trials.

The results are highly surprising: we expected some mistakes but not quite as many as we got. 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 trivial 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

[29] A variety of other severity indices have been proposed in the literature to measure the distance between actual and rational choices (Echenique et al. (2011), Apesteguia and Ballester (2015)). All are consistent with an underlying stochastic choice model in which the behavior of individuals who make more noisy choices result in more severe violations.

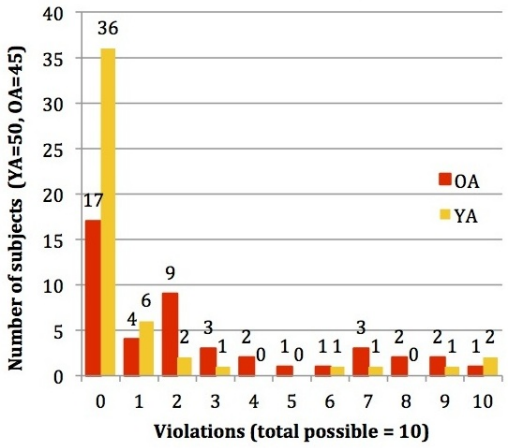

Figure 8: Number of violations in treatment **A** (trivial trials)

of our subjects 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 food, even though our protocol imposed the strongest possible emphasis into having hungry subjects, desirable goods, and small portions.[30] For subjects who fail 4 or 5 trivial trials, it is more difficult to disentangle between inattentiveness, miscomprehension, and 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 suggest the importance of including trivial trials in studies of consistency to test whether the assumptions of the model are satisfied by the experimental subjects.

A natural next step is to conduct the same study as in section 4.1 keeping only those subjects that *we think* satisfy the assumptions of our model. This substantially reduces the sample size, and asymmetrically so for YA and OA. Furthermore, it creates its own selection problem since the subsample is selected based on a variable (choices where less is preferred to more) which is linked to the dependent variable of the study. However, we still think it is a useful exercise, and we find it more satisfactory than ignoring the problem altogether.

Below, we present the results when we restrict our attention to subjects who fail at most two trivial trials. We choose that number in order to exclude 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

[30] 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.

two errors is admittedly ad-hoc. Figure 9 is the analogue Figure 4 for those individuals.

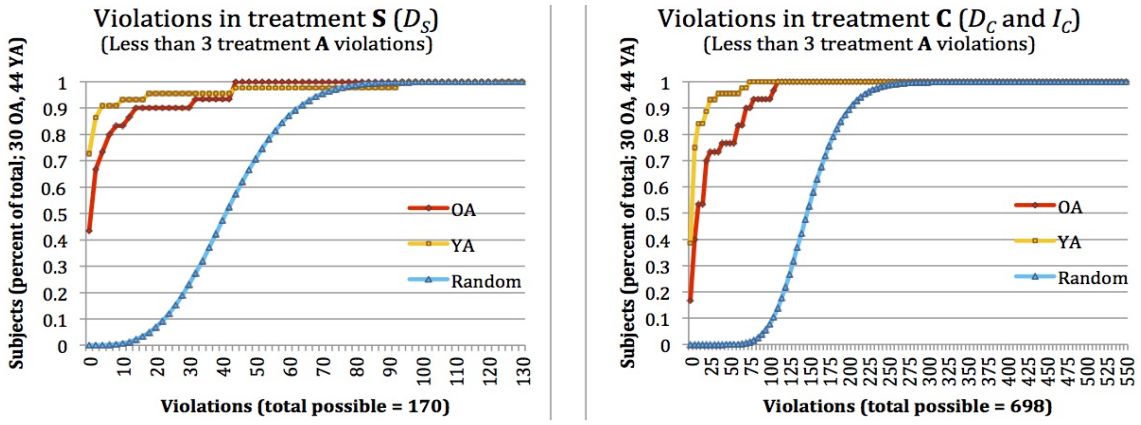

Figure 9: Choice violations by subjects with at most two treatment **A** violations

As expected, violations are significantly reduced when we consider only the subjects with at most two 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 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).[31]

5 Understanding violations

An obvious reason why an individual might commit violations is that her preferences do not satisfy the main GARP assumptions. Given the behavior of subjects in treatment **A**, there might be a non-negligible fraction of those individuals. Since the interpretation of the results is radically different for those subjects, we investigate below the determinants of consistency using the entire population and also using the subsample of subjects for

[31] 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 to that of the entire population: KS and WRS tests show no significant differences in distributions in treatment **S** (p-value = .534 and .305, respectively) and show significant differences in treatment **C** (p-value = .060 and .022, respectively). Again, the graphs are omitted for brevity.

whom we are most confident that the model is appropriate (those who fail at most two trivial trials).

Recent neuroscience studies have shown that some functions or abilities play a critical role in the quality of decisions and tend to decline with age (Gould et al. (2003), Cappel et al. (2010)). We therefore expect to find an association between decreased performance in tasks that assess those functions and abilities and decreased consistency in complex choices. Naturally, we also expect to replicate known associations between performance in those tasks and aging.

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. Working memory is the ability for storing information for immediate processing (Baddeley and Hitch, 1974; Baddeley, 1992). Subjects with low working memory and high working memory perform similarly in simple discrimination or detection tasks, but in complex tasks, working memory predicts task performance (Cerella et al., 1980; Gick et al., 1988). If non-consistent choices are a result of the subject's inability to simultaneously maintain a representation of many values, then GARP violations are expected to be more pronounced in treatment **C** –where more item values must be held in mind– than in treatment **S**. To investigate this hypothesis, we study scores in the spatial working memory test performed in the experiment.[32]

Performance in the working memory test is higher for YA (mean = 203, st. error = 3) than for OA (mean = 152, st. error = 1.73) and the difference is highly significant (p-value $< .001$).[33] This is consistent with many previous findings (see e.g., Salthouse and Babcock, 1991; Park et al., 2002).[34] A regression between working memory scores and a group dummy shows that the two are highly correlated both when we consider the full sample (p-value $< .001$, Adj. $R^2 = 0.71$) and when we restrict attention to subjects with at most two violations in treatment **A** (p-value $< .001$, Adj. $R^2 = 0.69$).

Another candidate to explain differences in consistency across age groups is IQ. General intelligence has two main components: fluid intelligence, which is 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, her objective is to accurately represent her true preferences and act accordingly. This task requires a certain level of reasoning about true values, which may rely on fluid intelligence.

[32] We shall emphasize that the working memory hypothesis is about the relationship between the general ability to reliably compute and compare value in a trial-by-trial basis and the number of items involved, not about the ability to remember how one chose in the past to not contradict that choice in the future.

[33] Due to software malfunction, 2 OA did not complete the working memory test and are excluded from all following analyses which include the measure.

[34] A correlation analysis between age and the working memory score shows the same result.

To test this hypothesis, we can use the answers to the Raven's IQ test, which is designed to measure fluid intelligence.

Performance in Raven's IQ test is again higher for YA than for OA both for the full sample (11.44 vs. 8.16) and for subjects with at most two violations in treatment **A** (11.39 vs. 8.77) and the differences are highly significant (p-value $< .001$ for both).[35] 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 Raven's test measures fluid intelligence, OA are expected to perform worse.

Having established that working memory (WM) and fluid intelligence (IQ) are lower for older subjects, we now study the correlation between these two measures and the number of violations in treatment **S** ($Viol\text{-}S$) as well as the total number of direct and indirect violations in treatment **C** ($Viol\text{-}C$).[36] The results are presented in Table 2 for the entire sample (left) and the subsample of subjects who fail at most two trivial trials (right).

All subjects

	Viol-S	*Viol-C*	*WM*
Viol-C	0.56***		
WM	-0.10	-0.23*	
IQ	-0.01	-0.22*	0.68***

*, **, ***: significant at 5%, 1%, 0.1% level

Subjects with 2 or less violations in **A**

	Viol-S	*Viol-C*	*WM*
Viol-C	0.12		
WM	-0.06	-0.26*	
IQ	-0.06	-0.29*	0.65***

*, **, ***: significant at 5%, 1%, 0.1% level

Table 2: Pearson correlations of memory, intelligence and GARP violations.

Except for the correlation between violations in both treatments, the results are very similar when we consider the entire sample or only the subjects who fail at most two trivial trials. We find no relationship between the number of violations in treatment **S** and performance in the working memory or IQ tests. By contrast, violations in treatment **C** are negatively correlated with both working memory and IQ scores.

The findings related to working memory are consistent with the hypothesis that subjects use a decision-making process that requires them to encode the value of items. They are not consistent with interpretations that subjects are attending to only the count of items or attending to only a single element of the options. The findings related to IQ suggest that fluid intelligence is heavily involved in choice processing only for the most

[35] Again, a correlation analysis between age and the Raven IQ score shows the same result.

[36] We also conducted the analysis for direct and indirect violations separately and found qualitatively similar results. It is also worth noting that $\mathcal{D}_C$ and $\mathcal{I}_C$ are strongly correlated (Pearson correlation = 0.84).

complex tasks. It should be noted however that working memory and fluid intelligence are very strongly correlated. This is in line with previous studies (see e.g., Engle et al., 1999) and reflects 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).[37]

To further investigate the relationship between violations in the complex treatment and performance in working memory and IQ tests, we conduct a set of ordinary least squares (OLS) regressions where the dependent variable is the number of violations in **C**. Explanatory variables include the variables presented above (violations in **S**, working memory scores, IQ scores) as well as a Younger Adult dummy (*YA-d*) and household income (*Income*). The results are presented in Table 3.

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Const.	142*	118**	50.5**	38.5	44.9	204**	148**	79.1***	85.6*	197
	(57)	(37)	(14)	(29)	(123)	(68)	(44)	(16)	(35)	(155)
Viol-S	2.45***	2.48***	2.46***	2.60***	2.74***					
	(0.37)	(0.37)	(0.37)	(0.43)	(0.42)					
WM	-0.66*				0.43	-0.85*				-0.44
	(0.31)				(0.82)	(0.38)				(1.0)
IQ		-9.33**			-9.72		-9.60*			-6.21
		(3.5)			(5.9)		(4.3)			(7.6)
YA-d			-46.2*		-43.3			-50.2*		-2.15
			(18)		(49)			(22)		(62)
Income				-3.05	5.91				-7.39	0.05
				(7.5)	(8.2)				(9.2)	(11)
Adj. R^2	0.35	0.35	0.35	0.34	0.39	0.04	0.04	0.04	-0.01	-0.01
obs.	93	95	95	72	70	93	95	95	72	70

(standard errors in parentheses); *, **, *** = significant at 5%, 1% and 0.1% level

Table 3: Ordinary least squares (OLS) regression of number of violations in treatment **C** (all subjects)

After controlling for violations in **S**, working memory, IQ, and age, *group* has significant explanatory power to understand consistency in **C** (regressions 1-3), but income does not (regression 4).[38] The similarities in significance of the regressions are not surprising since

[37] Given the age heterogeneity in our OA population, we performed a within-sample analysis. We found that the correlation between violations in **C** and scores in working memory and IQ tests keep the same sign but lose significance, in part due to the lower number of observations (data omitted for brevity).

[38] We should notice however a further selection problem since not all individuals reported the income of

we know from the previous analysis that *WM* and *IQ* are highly correlated and age is a strong predictor of performance in those tests.[39] When violations in treatment **C** are regressed on all of the variables (regression 5), the coefficients on working memory scores, IQ scores, and the YA dummy lose significance as these are highly correlated. Finally, the results are very similar when the variable *Viol-S* is excluded (regressions 6-10). However, the adjusted R^2 values are drastically lower, indicating a worse fit.

For robustness, we then perform the same regressions with the subset of subjects who failed two or less trials in treatment **A**. The results are presented in Table 4. While violations in treatment **S** are no longer a significant predictor for violations in treatment **C**, working memory scores, IQ scores, and age group are still highly significant in their explanatory power (regressions 1-3) and income is still not (regression 4). Similar qualitative conclusions are obtained when we exclude violations in treatment **S** (regressions 6-10).

	1	*2*	*3*	*4*	*5*	*6*	*7*	*8*	*9*	*10*
Const	56.6**	48.3***	23.7***	13.1	25.7	58.5**	49.9***	24.7***	15.5	33.2
	(19)	(14)	(4.5)	(9.5)	(41)	(19)	(14)	(4.3)	(9.0)	(40)
Viol-S	0.18	0.17	0.17	0.20	0.26					
	(0.2)	(0.2)	(0.2)	(0.3)	(0.2)					
WM	-0.23*				0.12	-0.24*				0.09
	(0.1)				(0.3)	(0.1)				(0.3)
IQ		-3.37*			-3.85		-3.45*			-3.83
		(1.4)			(1.9)		(1.4)			(1.9)
YA-d			-17.1**		-21.4			-17.4**		-19.4
			(5.6)		(14)			(5.6)		(14)
Income				0.10	4.49				-0.32	3.87
				(2.4)	(2.4)				(2.3)	(2.3)
Adj. R^2	0.05	0.07	0.10	-0.03	0.17	0.06	0.07	0.11	-0.02	0.17
obs.	74	74	74	53	53	74	74	74	53	53

(standard errors in parentheses); *, **, *** = significant at 5%, 1% and 0.1% level

Table 4: OLS Regression of number of violations in treatment **C** (subjects with 2 or less violations in treatment **A**)

Overall, the results are consistent with a cognitive decline theory of behavioral differ-

their household. Comparisons are also difficult since for most YA income refers to that of their parents whereas for OA it is theirs and their spouses.

[39] A principal component analysis on *WM* and *IQ* suggests that working memory data contains the largest fraction of the relevant information: the first component is mostly driven by working memory score and explains 70% of the data.

ences in complex tasks between YA and OA. As noted earlier, recent neuroscience findings support the hypothesis that brain regions involved in working memory and fluid intelligence are critical to processing information in complex situations. Other studies have evidenced that aging impairs those regions yielding a decrease in working memory and IQ measurements as well as performance in complex tasks. Our findings suggest that the same logic applies to economic decision-making: GARP consistency is mediated by the brain structures involved in working memory and 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 and retrieve information as well as to perform logical reasoning is reflected in the level of consistency of her choices.

Next, we examine the responses obtained in our questionnaire. We find that OA self-report a lower stress level compared to YA (p-value $< .001$) and that reported stress correlates with working memory scores (Pearson correlation = .29, p-value = .006).[40] Interestingly, self-reported health rankings are similar across groups and uncorrelated to any relevant element of our analysis. Last, we 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.

Finally, there are inevitably some unobservable factors that may have different effects on subjects of different ages. One such factor is fatigue. The case could be made that fatigue affects OA more severely than it affects YA, leading to disparate levels of consistency. There is a stream of psychological research that is relevant to the question of whether older adults are more susceptible to fatigue. This research is on the phenomenon of "ego-depletion" – the impairment of decision-making immediately following a task requiring thoughtfulness or self control. If a subject is highly susceptible to ego-depletion then the quality of their decisions worsens as an experiment progresses. Research shows that older adults are less susceptible to ego-depletion than are younger adults (Dahm et al., 2011), so the main effect of ego-depletion in our experiment should go in the opposite direction.[41]

[40] Note however that reported stress is not correlated with the number of violations in the complex GARP task (Pearson correlation = -.09, p-value = .41).

[41] When violations per trial are regressed on their order of appearance, a Younger Adult dummy, and the interaction of these variables, order is a significant predictor of violations but the interaction of order and age dummy is not. This suggests that fatigue may affect consistency but not differently across age

Other possible factors include differences in the opportunity cost of time, sensitivity to hunger, cognitive skills or experience in individual decision making. Unfortunately, it is not feasible to rule out all these factors, given our design. While we agree that they raise caution as to the interpretation of our findings, we do not feel that any of them has a clear and unambiguous differential effect on our populations.[42] Perhaps more importantly, if either age group were to be more susceptible to any of these factors, it would be reasonable to believe that their performance would be affected in treatments **S** and **C** alike.

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 our two populations. Our older adults are less consistent than our younger adults, both in terms of the number and severity of violations, when the choice task is complex. Also, differences in consistency in the complex task is associated to deficiencies of working memory, that is, in the ability to store and retrieve information regarding the value of the different items in bundles. A few comments are in order.

First, the results are consistent with evidence reported in a growing literature, in particular in neuroscience, that links behavior, cognition and aging. There is indeed converging evidence that structures involved in working memory are responsible for performance in complex situations. At the same time, decline in working memory is associated with decreased performance in such tasks. Our study suggests that behavior in a simple economic decision-making task is consistent with the same mechanism. In future research, we plan to use fMRI techniques to study the neural correlates of choice consistency and test the theory outlined here. Note that our experimental design, characterized by two bundles presented in a screen, a left-right choice and the possibility of multiple repetitions (see Figure 3) is suitable to be implemented in the scanner. We also 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, in particular in the complex domain, and how this interaction differs across ages.

groups.

[42] For example, OA are likely to have a lower opportunity cost of time, but this may very well increase consistency by, other things equal, inducing them to take more time, effort and care in their decision making. Cognitive skills help consistency but this is affected by education and our OA are at least as educated as our YA. Finally, YA in our sample are possibly more experienced with experimental tasks but it is unlikely that it eclipses the additional decades of decision-making experience held by the OA.

Second, the individual analysis (see the appendix) suggests that consistency across ages is similar in the simple task partly because the older adults in our sample have preferences consistent with a simple rule (typically, to maximize the quantity of the favorite item in the bundle) that can be easily implemented without errors, or on-line reference to subjective value. An important question for future study is whether these preferences are intrinsic to subjects or if it is a second-best strategy employed by individuals who are aware of their compromised working memory and fluid intelligence.

Third, we have shown that working memory is associated with decreased choice consistency in the context of complex choices and aging. Decreased working memory ability and aging go hand in hand. This means in particular that working memory and aging cannot be dissociated in our study. However, poor working memory is also linked to young age and behavioral disorders such as attention disorders. As noted in the introduction, there is also evidence that consistent decision-making improves during childhood (Harbaugh et al., 2001). We conjecture that this tracks the development of brain areas involved in working memory (Gathercole et al. (2004)). In a similar vein, we expect that attention deficits in children and adolescents to be associated with decreased consistency levels in choice paradigms because such deficits correlate with poor performance in working memory tasks (Martinussen (2005)). Further research in new populations should prove useful to establish a *causal* link between working memory and rational behavior in the context of complex choices.

Fourth, our study is part of a larger agenda that investigates the contribution of cognitive functions, and the underlying brain mechanisms that sustain them, to economic decision-making. With respect to the growing literature that looks for associations between cognition and rationality (Dohmen et al., (2018), Andersson et. al. (2016)),[43] our results suggest that this relationship is context-dependent rather than universal. Some tasks do not require the involvement of specific functions while others do. At the same time, people differ in their ability to engage such functions. This is consistent with a dual process interpretation in which specific brain regions are recruited as a function of task demands and individual specific characteristics. We conjecture that this mechanism is operating under other paradigms.[44]

Last, from a methodological perspective, our study introduces a very simple experimental task to test for consistency. This task is visual and intuitive enough to be implemented with populations differing in attention span or comprehension ability. An alley for fu-

[43]These articles are focusing on risk preferences. Andersson et al. (2016) provides evidence suggesting that the elicitation of risk preferences interferes with the propensity to make errors, which is itself related to cognitive abilities.

[44]This mechanism could rationalize the observations in Andersson et al. (2016).

ture research is to provide an in-depth theoretical framework to study consistency in such binary choice experiments and extend our attempt to measure severity in that context. A more systematic theoretical analysis of the relationship between measures of GARP consistency and measures of error in the random utility model would be also valuable in that context.

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, 1,* 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. Andersson O., H. Holm J-R Tyran and E. Wengström (2016). Risk aversion relates to cognitive ability: preferences or noise? *Journal of the European economic Association, 14* (5), 1129-1154.

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

6. 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.

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

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

9. Apesteguia, J. and M.A. Ballester (2015). A measure of rationality and welfare. *Journal of Political Economy, 123*, 1278-1310.

10. Baddeley, A. (1992). Working Memory. *Science, 255*(5044), 556-559.

11. Baddeley, A., and Hitch, G. (1974). Working Memory. In G. H. Bower (Ed.), *The Psychology of Learning and Motivation* (47-89). Academic Press.

12. 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.

13. Banerjee S. and J.H. Murphy (2006). A simplified test for preference rationality of two-commodity choice. *Experimental Economics, 9*(1), 67-75.

14. 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.

15. 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.

16. 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.

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

18. Besedes, T., Deck, C.A., Sarangi, S., and Shor, M. (2012b). Decision-making strategies and performance among seniors *Journal of Economic Behavior & Organization, 81*(2), 524-533.

19. Bradbury, H., and Nelson, T. M. (1974). Transitivity and the patterns of children's preferences. *Developmental Psychology, 10*(1), 55-64.

20. Brainard, D. (1997). The psychophysics toolbox. *Spatial Vision*, 433436.

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

22. 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.

23. Bronars, S. G. (1987). The power of nonparametric tests of preference maximization. *Econometrica, 55*(3), 693-698.

24. 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.

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

26. Cappell, K.A., Gmeindl, L., and Reuter-Lorenz, P.A. (2010). Age Differences in Prefrontal Recruitment During Verbal Working Memory Maintenance Depend on Memory Load. *Cortex, 46*(4), 462-473.

27. 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.

28. 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.

29. Castillo, M., Dickinson, D. L., and Petrie, R. (2017). Sleepiness, Choice Consistency, and Risk Preferences. *Theory and Decision, 82*(1), 41-73.

30. 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.

31. Cerella, J., Poon, L. W., and Williams, D. M. (1980). Age and the complexity hypothesis. *Aging in the 1980s: Psychological issues*, 332-340.

32. 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.

33. 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.

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

35. 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.

36. 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.

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

38. Curtis, C.E., and D'Esposito, M. (2003). Persistent activity in the prefrontal cortex during working memory. *TRENDS in Cognitive Sciences, 7*(9), 415-423.

39. Dahm, T., Neshat-Doost, H. T., Golden, A. M., Horn, E., Hagger, M., and Dalgleish, T. (2011). Age shall not weary us: Deleterious effects of self-regulation depletion are specific to younger adults. *PLoS ONE, 6*(10), 811.

40. Dean, M., and Martin, D. (2016). Measuring Rationality with the Minimum Cost of Revealed Preference Violations. *Review of Economics and Statistics, 98* (3), 524-34.

41. 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.

42. Dohmen T., Falk A., Huffman D. and Sunde U. (2018). On the Relationship Between Cognitive Ability and Risk Preference. *Journal of Economic Perspectives, 32*(2), 115-134.

43. 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.

44. 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.

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

46. 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), 309-331.

47. 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*.

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

49. 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.

50. 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.

51. Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review, 97*(5), 1858-1876.

52. 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.

53. 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*.

54. Gathercole S.E., S. J. Pickering, B. Ambridge, and H. Wearing (2004). The Structure of Working Memory From 4 to 15 Years of Age. *Developmental Psychology, 40* (2), 177-190.

55. von Gaudecker H, van Soest A. and Wengström E. (2011). Heterogeneity in Risky Choice Behavior in a Broad Population. *American Economic Review, 101*, 664-694.

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

57. Gick, M. L., Craik, F. I., and Morris, R. G. (1988). Task complexity and age differences in working memory. *Memory and Cognition, 16*(4), 353 - 361.

58. Gould, R.L., Brown, R.G., Owen, A.M., Ffytche, D.H., and Howard, R.J. (2003). fMRI BOLD response to increasing task difficulty during paired associates learning. *NeuroImage, 20*(2), 1006-1019.

59. 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.

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

61. Greene, J.D., Nystrom, L.E., Engell, A.D., Darley, J.M., and Cohen, J.D. (2004). The Neural Bases of Cognitive Conflict and Control in Moral Judgment. *Neuron, 44*(2), 389-400.

62. 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.

63. 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.

64. 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.

65. Harrison, G. W., Lau, M. I., and Williams, M. B. (2002). Estimating individual discount rates in Denmark: A field experiment. *American Economic Review, 92*(5), 1606-1617.

66. 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.

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

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

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

70. 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.

71. 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.

72. 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.

73. 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.

74. Kondo, H., Osaka, N., and Osaka, M. (2004). Cooperation of the anterior cingulate cortex and dorsolateral prefrontal cortex for attention shifting. *NeuroImage, 23*(2), 670-679.

75. 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.

76. 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.

77. Krueger, F., Spampinato, M.V., Pardini, M., Pajevic, S., Wood, J.N., Weiss, G.H., Landgraf, S., and Grafman, J. (2009). Integral calculus problem solving: An fMRI investigation. *Neuroreport, 19*(11), 1095-1099.

78. 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.

79. 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.

80. MacDonald, A.W., Cohen, J.D., Stenger, V.A., and Carter, C.S. (2000). Dissociating the Role of the Dorsolateral Prefrontal and Anterior Cingulate Cortex in Cognitive Control. *Science, 288*(5472), 1835-1838.

81. Martinussen R., J. Hayden,s. Hogg-Johnson and R. Tannock (2005). Meta-Analysis of Working Memory Impairments in Children With Attention-Deficit/Hyperactivity Disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 44* (4), 377-384.

82. 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.

83. 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.

84. 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.

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

86. 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.

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

88. Nishimura H., E.A. Ok and J. K.-H. Quah. (2017). A Comprehensive Approach to Revealed Preference Theory. *The American Economic Review, 107*, 1239-1263.

89. 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.

90. Park, D.C., Lautenschlager, G., Hedden, T., Davidson, N.S., and Smith, A.D. (2002). Models of visuospatial and verbal memory across the adult life span. *Psychology and Aging, 17*(2), 299-320.

91. Pelli, D. G. (1997). The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spatial Vision.*

92. Pochon, J.B., Levy, R., Poline, J.B., Crozier, S., Lehéricy, S., Pillon, B., Deweer, B., Le Bihan, D., and Dubois, B. (2001). The role of dorsolateral prefrontal cortex in the preparation of forthcoming actions: an fMRI study. *Cerebral Cortex, 11*(3), 260-266.

93. 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.

94. Rangel, A., and Clithero, J. (2013). The computation of stimulus values in simple choice. In P.W. Glimcher and E. Fehr (Eds.), *Neuroeconomics: decision making and the brain* (125-147). Academic Press.

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

96. 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.

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

98. Rypma, B., and D'Esposito, M. (2000). Isolating the neural mechanisms of age-related changes in human working memory. *Nature Neuroscience, 3*(5), 509-515.

99. Salthouse, T.A., and Babcock, R.L. (1991). Decomposing Adult Age Differences in Working Memory. *Developmental Psychology, 27*(5), 763-776.

100. Samuelson, P.A. (1938). A Note on the Pure Theory of Consumer Behavior. *Economica, 5*(17), 61-71.

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

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

103. 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.

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

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

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

107. Wright, R.E. (1981). Aging, Divided Attention, and Processing Capacity. *Journal of Gerontology, 36*(5), 605-614.

108. 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

Appendix A1. Example of direct violation in a triplet of trials of treatment S

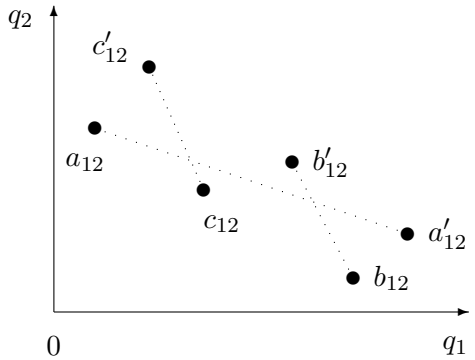

Figure 10: Trials (a_{12} vs. a'_{12}), (b_{12} vs. b'_{12}), (c_{12} vs. c'_{12})

In the example of Figure 10, no pair of trials satisfies condition (i) of Definition 1, so there cannot be a direct violation between any pair of trials. Suppose now that a_{12} is chosen over a'_{12}, b_{12} is chosen over b'_{12} and c_{12} is chosen over c'_{12}. Since $q_x^{a'} > q_x^b$ for all x and $q_x^{b'} > q_x^c$ for all x, we have $a_{12} \succ a'_{12} \succ b_{12} \succ b'_{12} \succ c_{12}$. Since $q_x^{c'} > q_x^a$ for all x, we have $c_{12} \succ c'_{12} \succ a_{12}$. This forms a contradiction to the maximization of monotonic and transitive preferences.

Notice that the key reason we have a direct violation between triplets of trials but not between any pair of trials is that $q_2^{b'} < q_2^a$ and $q_2^{a'} < q_2^c$. This issue would not arise if, instead of a discrete number of alternatives we were to offer subjects the entire budget set (see Banerjee and Murphy, 2006).

Appendix A2. List of all 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); Gummy bears (2); Popcorn (2); M&M (2); Dark chocolate peanut butter cup (1); Mini chocolate-covered pretzel (1); 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);

Appendix A3. 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 declines 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.

A3.1. A random utility model (RUM)

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 BU^l, and a bundle on the right (r) of the screen, denoted by BU^r. A decision is obtained by comparing the utility derived from each option. We assume that utility depends linearly on the observable quantities of the goods 1 and 2, as well as on a stochastic unobserved error component ϵ_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 BU^l in trial o is therefore:

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

and $P_{io}^r = 1 - P_{io}^l$. We assume that error terms are i.i.d. and follow an extreme value distribution: the cumulative 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 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 β_{i1} and β_{i2}, which we achieve by maximum likelihood.[45]

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:

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

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

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

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

and the probability that subject i chooses option BU^l is the logistic function:[46]

$$P^l_{io}(q^l_{wo}, q^l_{so} - q^r_{so}, q^r_{po}) = \frac{1}{1 + e^{-\left(\beta_{iw} q^l_{wo} + \beta_{is}(q^l_{so} - q^r_{so}) - \beta_{ip} q^r_{po}\right)}}.$$

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 misclassified trials. Importantly, we find that misclassification rates in each treatment are strongly correlated with the number of violations (Pearson coefficient = .77 in treatment **S** and .77 in treatment **C**). 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. Finally, notice that RUM presupposes more errors when the difference in utility between the two bundles is small, which means that it is also related to severity of violations.[47] It is therefore natural that we also find a significant correlation between severity of violations and misclassification rates (Pearson coefficient = .52 in treatment **S** and .32 in treatment **C**).

A3.2. Clustering

We then 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 understanding the treatment effect between simple and complex choices. A wide

[46] The log-likelihood is now:

$$\log L_i = \sum_{o=1}^{O} \log \left[P^l_{io}(q^l_{wo}, q^l_{so} - q^r_{so}, q^r_{po}) \mathbf{1}_l + [1 - P^l_{io}(q^l_{wo}, q^l_{so} - q^r_{so}, q^r_{po})][1 - \mathbf{1}_l] \right]$$

[47] 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 $|\text{BU}^r - \text{BU}^l|$. 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, that is, those for which RUM is not well specified. Finally, as another robustness check, we correlated RUM misclassifications with GARP violations $\mathcal{D}_\mathcal{C}$ and $\mathcal{I}_\mathcal{C}$ separately, and obtained the same results.

array of heuristic clustering methods are commonly used, however 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). For our data, the ellipsoidal, equal shape model that endogenously yields *three* clusters maximizes the BIC.

Table 5 provides summary statistics of the three clusters. The first two rows display the average percentage of RUM misclassifications in **S** and **C** by subjects in each cluster, the 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 each cluster in the consistency task (GARP violations) and the tests (WM and IQ). Clusters are ordered from smallest to largest in the percentage of misclassified observations.

	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 5: Summary statistics by cluster.

Cluster 1 is characterized by almost no misclassification in **S** and few in **C**. Cluster 2 also exhibits limited misclassifications in both **S** and **C** (although more than cluster 1). Cluster 3 has substantial misclassifications in both treatments. The first surprising finding is the allocation of OA and YA across clusters. Given our previous results, one would expect more YA in cluster 1 and more OA in cluster 2. We find the reverse. Cluster 3 is a mix of subjects.

When we consider performance in the choice tasks and tests, we notice that cluster

3 stands out as a group of inconsistent subjects exhibiting a large number of GARP violations and low performance in WM and IQ tests. These subjects also fail our trivial trials much more frequently than the rest of the subjects. Not surprisingly, the vast majority of minimizers (6 in treatment **S** and 5 in treatment **C**) belong to this cluster.

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 are slightly more inconsistent. In treatment **C**, subjects in cluster 1 decrease significantly in their performance while subjects in cluster 2 remain more consistent. Overall, cluster 2 is a group of "consistently consistent" subjects. By contrast, subjects in cluster 1 are remarkably consistent in **S** but significantly less in **C**.

Figure 11 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 5).[48] 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 5).

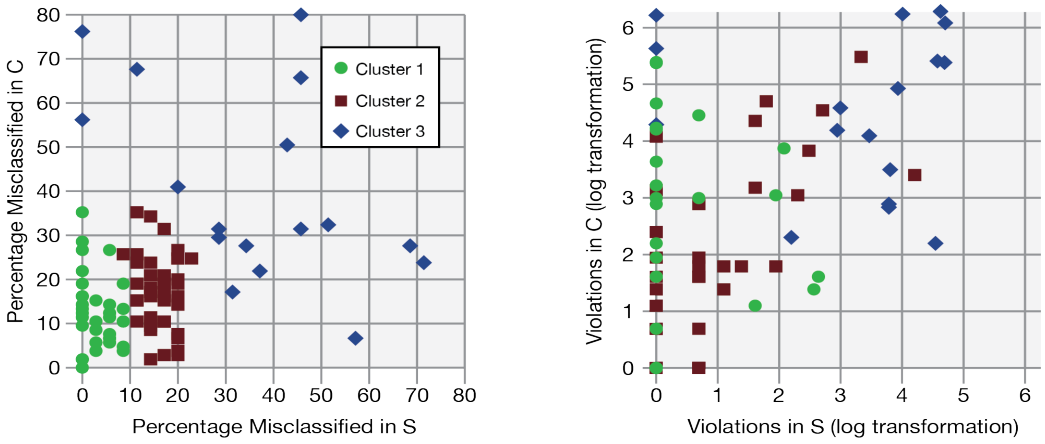

Figure 11: Cluster representation. Misclassified trials (left) and number of violations (right) in treatments **S** and **C**.

Clusters are clearly differentiated in the left graph. This is is not surprising since the variables are, up to a transformation, the ones used for grouping the individuals. The figure highlights the differences across clusters emphasized above: small percentage of

[48] 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.

RUM misclassifications in both treatments for cluster 1, slightly larger in **S** for cluster 2, and a substantial fraction of misclassifications for cluster 3 in both treatments.

More interestingly, the right graph also shows clear differences 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 GARP violations. It also suggests that the transition from the simple to the complex situation is more difficult for individuals in cluster 1 than for those in cluster 2. We investigate this issue in more detail in the next section.

Finally, we find that subjects in cluster 1 have significantly worse working memory scores than subjects in cluster 2 (p-value = .040); they also have lower IQ scores but the difference is not statistically significant. This suggests a relationship between working memory and the ability to *remain consistent* as the complexity of the task increases.

A3.3. Simple choice rules

The extreme degree of consistency and lack of misclassifications by cluster 1 subjects in treatment **S** (26 subjects out of 33 have zero violations), together with the fact that many of them are in the OA population and perform significantly worse in **C** is somewhat puzzling. Examining the value estimates of the RUM model (the β_{ij}-coefficients) in more detail, we find that for some subjects one value estimate in **S** and two value estimates in **C** are close to 0. These are subjects whose behavior is consistent with maximizing the quantity of their most preferred item. For some other subjects, the value estimates of all goods are almost identical to each other. These are subjects for whom goods are perfect substitutes, so that their behavior is consistent with maximizing the total quantity in the bundle. These two choice strategies are clearly consistent with the maximization of monotonic and transitive preferences, resulting in high degrees of consistency. At the same time, subjects with these types of preferences do not need to perform sophisticated mental trade-offs between items and, instead, can use simple choice rules. We therefore hypothesize that having these specific preferences may potentially explain why cluster 1 exhibits such an extremely high level of consistency in treatment **S**.

With this idea in mind, we construct two simple choice rules for subjects in clusters 1 and 2: H (for highest), where the subject maximizes the quantity of one of the items in the bundle (presumably, the one with highest value) and T (for total) where the subject maximizes the total quantity in the bundle (presumably because goods are perfect substitutes). We assign type H (T) to a subject if (i) the rule H (T) generates the same or fewer

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 quick and simple choice rule that can be implemented with "few" errors. Otherwise, we assign type O to the subject (for other). In other words, we assume that the 33 subjects in cluster 1 and 45 subjects in cluster 2 maximize a well defined utility function, linear in the goods present in the bundle, but that they make some errors. As we know from sections A3.1 and A3.2, this is a reasonable description of behavior by subjects in those clusters. We then divide the sample into three types, depending on whether the optimal choice given their preferences can be implemented with a simple rule (types H and T) or not (type O). Table 6 summarizes the number of subjects of each type by cluster. We also add in parentheses the average number of violations for type O subjects (since violations are typically very small for types H and T, the numbers are omitted).

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

Table 6: Types of preferences by subjects in clusters 1 and 2

All but one subjects in cluster 1 have preferences consistent with a simple rule in treatment **S**, mostly H. More than half of these subjects change their strategy in treatment **C** and are there best classified as type O. By contrast, only one-quarter of subjects in cluster 2 have a preference consistent with a simple rule and there is no treatment effect. It is remarkable to see such sharp differences across clusters of choices consistent with simple rules, even though subjects are *not* grouped based on that dimension.

Our conjecture is that 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 (type H); if both goods are of similar value, the subject can focus on total quantities (type T). In treatment **C**, subjects are forced to compare "apples to oranges" so simple rules are less intuitive to implement.[49] Subject are more likely to explicitly trade-off the different alternatives, which explains why more of them are better classified as type O. Finally, since trade-offs are difficult, type O subjects have significantly more violations than either type H or T subjects.

A3.4. Summary

[49] 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.

Overall, the individual analysis reveals interesting insights regarding the preferences and strategies of our subjects. First, a structural model (RUM) – where utility depends linearly on the quantities of goods in each bundle and the subject chooses the bundle that yields the highest utility – provides a good fit for a majority of subjects, but by no means for all of them. Second, a cluster analysis based on RUM misclassifications suggests three distinct groups. The RUM provides a reasonably good fit for two groups of subjects (clusters 1 and 2) and a poor fit for the last one (cluster 3). Third and as expected, RUM misclassifications are correlated with GARP violations. Subjects in cluster 3 perform badly in both treatments of the consistency task, whereas subjects in clusters 1 and 2 perform reasonably well. Surprisingly, however, cluster 1 (the group with the fewest RUM misclassifications) has more violations in treatment **C** than does cluster 2. The composition is also different: two-thirds of subjects in cluster 1 are OA whereas two-thirds of subjects in cluster 2 are YA. Fourth, an analysis of simple rules of behavior consistent with utility maximization sheds light on the differences in age composition and consistency across tasks between clusters 1 and 2. Cluster 1 is mostly composed of OA who use a simple rule in treatment **S** (maximize the amount of the preferred good), resulting in extremely consistent behavior. Their consistency decreases substantially in treatment **C**, possibly due to the difficulty of implementing a simple rule when the bundles contain different goods. By contrast, cluster 2 is mostly composed of YA who use simple rules significantly less often but perform better value-quantity tradeoffs. These subjects make slightly more consistency mistakes in **S** but less in **C**. Finally, a conjecture consistent with the results presented here is that some subjects who are aware of their compromised working memory and fluid intelligence (mostly OA) resort to simple choice rules. Such strategies can be applied in the simple treatment but not in the complex one. This explanation is reasonable and appealing, however it requires the supporting evidence of new experiments.

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.*