Do people disagree with themselves? Exploring the internal consistency of complex, unfamiliar and risky decisions.

Douglas L. Bessette *[1], Robyn S. Wilson[2], & Joseph L. Arvai [3, 4]

[1] Department of Community Sustainability
College of Agriculture & Natural Resources
Michigan State University

[2] School of Environment & Natural Resources
The Ohio State University

[3] Erb Institute for Global Sustainable
School of Environment & Sustainability, *and*
Ross School of Business
University of Michigan

[4] Decision Research, Eugene, OR

*Corresponding Author
Email: bessett6@msu.edu

ABSTRACT

It is commonly accepted that people disagree with one another. In this paper, we present results that suggest people may disagree with themselves. Using eight decision-making contexts ranging in familiarity, complexity and risk we show that a nationally representative sample (n = 1,874) of respondents made choices that were inconsistent across two complimentary methods of eliciting preferences. We show that on average individuals demonstrate higher levels of internal consistency, or alignment between their choices and their stated values and concerns, when decisions are "easy," or simple, familiar and have little risk. However, this consistency declines when people are confronted with difficult choices involving unfamiliar, complex contexts involving high risk. Moreover, providing additional and salient contextual information about alternatives, such as brand names, model information or the specific components of a risk mitigation strategy, results in significantly lower levels of consistency when compared to situations where this information is withheld. This finding suggests that people rely on simplifying heuristics when making easy decisions; however, this kind of information is less influential when choices are difficult. Importantly, we show that higher levels of education also have a significant and positive association with the consistency of people's choices.

1. INTRODUCTION

On this we can all agree: people disagree. Not everyone buys the same vehicle, supports the same political candidate, or agrees on what to do about climate change. In this research, we explore a related question: do people disagree with *themselves*?

Researchers have invested considerable time and effort in developing an understanding of the nature and construction of preferences [1, 2]. Stemming from this work, two perspectives on preferences have been identified: that of 'articulated values' and that of 'basic values' [3].

The articulated values perspective holds that individuals' choices are accurate representations of their values and priorities. Any demonstrated inconsistencies in preferences are, therefore, the result of people—e.g., researchers, practitioners—asking poorly articulated or misleading questions during elicitation processes, and not a respondent's inability to articulate what they want [3]. The basic values perspective assumes, by contrast, that people lack well-established preferences for all but the most familiar judgmental contexts. Thus, when people are confronted with unfamiliar contexts, they must construct their preferences as part of an inferential process, one highly dependent upon the elicitation process and the specific context [3, 4].

The articulated values perspective holds considerable appeal for researchers because of the practical ease with which it can be applied. Political scientists and sociologists, for example, adopt this perspective when conducting public opinion polls [5, 6]; so too do economists when seeking responses to stated preference surveys [7, 8]. Research in the decision sciences however has shown that people are not rational maximizers [9], that they in fact construct their judgments in response to contextual and experiential cues [2, 10, 11], and often face obstacles in the form of judgmental biases [12], such as often falsely perceiving risks and benefits to be positively

correlated [13]. As a result, information obtained from certain kinds of surveys—e.g., about preferences for options, or about the tradeoffs people are willing to make when selecting among options—may be misleading.

Debates about the validity of each perspective have largely been philosophical in nature, with some taking a rather hard line [14], and others taking a more moderate position. Both Fischhoff [3] and Slovic [4], for example, acknowledge the possibility that the accuracy of people's preferences—in terms of meeting desired objectives—is a function of their familiarity or experience with the context of the elicitation. So, preferences about choosing a restaurant for an enjoyable dinner out, which is a context that most adults have ample experience with, will be based on more articulated values than will preferences for specific climate change mitigation strategies. Evaluating the latter requires more cognitively demanding judgmental tasks with which far fewer individuals have had direct or meaningful experience.

To be sure, a wealth of research compares alternative *frames* of fundamentally the same elicitation task; i.e., frames that are functionally similar or logically identical [15]. The use of these alternative frames often results in so-called "preference reversals," demonstrating what some argue to be the intransitivity of preferences [16]. Others disagree [17], arguing that implicit information provided by the speaker's choice of frame result in different, but completely justifiable and rational, choices [18]. Little experimental research, however, involves tests of alternative, but functionally similar elicitation approaches using large representative samples of respondents [3].

Here, we attempt to address this gap by using a nationally representative sample and two complimentary elicitation approaches. In doing so, we investigate the extent to which people's

preferences are consistent with their stated priorities as a function of both the elicitation method, and the level of difficulty, or the complexity, familiarity and risk of the decision context. If internal consistency remains high both across elicitation approaches and across "easy" and difficult decision contexts, it would suggest that the articulated values perspective prevails. On the other hand, if competing elicitation approaches yield different levels of internal consistency, or consistency varies as a function of complexity, familiarity or risk, then we might argue that the basic values perspective prevails.

Regardless of the perspective adopted, participants' choices would be expected to exhibit low levels of internal consistency if the elicitation *tasks* are convoluted or impenetrable. As such, the focus of our research was two straightforward elicitation approaches: direct ranking and attribute weighting. Ranking is what we term a holistic evaluation approach in that respondents simply order a series of stimuli (e.g., options in a choice set) from most to least preferred. The assumption behind such an approach is that participants are able to evaluate the information presented about alternatives in an integrated, holistic manner prior to articulating a preference order. Indeed, the ability to rank options in a manner consistent with one's values and priorities is a central element in the theory of rational choice [9, 19]. Moreover, it is a judgmental approach that is familiar to everyone. Indeed, the intuitive nature of ranking tasks contributes to their widespread use in more elaborate preference elicitation methodologies such as metric conjoint analysis [7, 20].

The same is true of decomposed choice by the weighting of attributes. Here, respondents are asked to evaluate the individual attributes that characterize the options under consideration, prioritize them in order of importance, and assign relative weights to each. Once elicited, these weights can be used by analysts to determine preference orders for options based on multi-

attribute utility theory [21, 22]. Applied scientists routinely rely on decomposed choice tasks in studies requiring optimization across potential options [23], as do decision analysts when attempting to formalize strategic planning and policy-making processes [24].

In the design of this research, we also took note of the argument by some proponents of the basic values perspective that individuals' preferences are heavily influenced by both past experience and contextual cues available during the presentation and consideration of options [2]. For example, if individuals were asked to indicate a preference for Coca Cola or Pepsi Cola by evaluating a series of attributes (e.g., price, sugar content, etc.) that characterize the two sodas, it is likely they would only conduct a cursory evaluation of the attributes. Individuals' preferences are likely to instead reflect an innate response to the soft drinks, based on prior experience with the brands and existing cultural values [25]. The same could be said for "no-fly zones" vs. "boots on-the-ground" as competing humanitarian intervention strategies. Individuals may ignore the technical attributes of these strategies and instead base their judgments on the positive or negative affective halos associated with them.

Research on internal consistency should, therefore, control for these contextual cues. In our study we did so by creating two between-subject versions of the experiment; one that included identifying information about the options (e.g., brand or strategy names like "no fly zone"), and one that did not. Based on previous work [26, 27] we speculated that individuals provided with identifying information would exhibit lower levels of internal consistency than those who did not. Furthermore, as suggested above, we expected consistency to decline as individuals encountered more difficult decision contexts.

By using ranking and weighting in a complementary manner, controlling for the effect of identifying information, and varying the difficulty of decision scenarios, research on the consistency of people's preferences and how consistency varies between contexts becomes more straightforward. Direct ranking provides a holistically established preference order; elicited weights are operationalized using a value function to reveal an implied preference order. Assessing individuals' internal consistency requires only comparing the two orders. Considerable inconsistency between the orders in certain contexts may have serious ramifications for both those engaged in measuring and enacting policy based on the public's preferences and the public themselves.

2. METHODS

2.1. Design

This study used a mixed between-within-subjects design, which respondents completed online. Before proceeding, respondents were provided an introduction to two elicitation tasks that accounted for the dependent variables in the study. Task 1, *holistic ranking*, asked respondents to imagine they were confronted with a particular decision context; e.g., purchasing a new midsize automobile. For each decision context, respondents were asked to review a menu of five options (e.g., alternative midsize sedans to choose from). Information about each option was characterized according to five attributes (e.g., price, safety, quality, etc.)—see Table 1 for a complete list of decision contexts, options and attributes. After reviewing the information presented in the table, respondents were asked to rank the five options from most to least preferred. Ties were allowed.

Insert Table 1 here.

In Task 2, *attribute weighting*, respondents were asked to review the attributes characterizing those options and make a series of judgments about the relative importance of each in guiding their choices. In order to elicit these judgments, we used a swing-weighting approach [28], which asks respondents to assign weights in a manner that accounts for the range of performance of each attribute based on the options in Task 1. In this method, participants are shown only the best performance and worst performance they can achieve on each attribute. They are then asked to provide a weight (out of 100) to represent how important it is for the participant to move each attribute from its worst performance to its best. Participants apply these weights only while looking at the attributes and their range of performance; the participant is not looking at the alternatives themselves. These weights are then put into a linear additive value model to establish an implied preference order for the options in each context [28]. Once again, ties were allowed.

Each respondent was presented with eight different decision scenarios ranging in difficulty, or complexity, familiarity and risk. To ensure we captured complexity accurately, 50 randomly selected individuals who were not included in the main experiment were asked to rate each scenario's complexity, as well as its importance to them personally and to society as a whole (see Supporting Information: S1 Table). Based on these results, the easiest choices included the tutorial, i.e, selecting a restaurant for dinner, comparing mobile phones (1), purchasing a new vehicle (2), and selecting a home or apartment to rent (3). The most difficult choices included evaluating humanitarian intervention options in Syria (6) and comparing climate change mitigation strategies (7). Two choices, comparing different types of fuel (4) and evaluating investment portfolios (5), were rated as moderately complex, and were considered relatively unfamiliar and risky, and thus fell in between easy and difficult.

In the main experiment, all respondents began with the easiest scenario: choosing among restaurants for a dinner out. This scenario served as an initial tutorial so that respondents could familiarize themselves with the holistic ranking and attribute weighting tasks. To ensure that respondents understood these tasks and completed them correctly, we developed a video tutorial that was shown to participants before they began the experiment. The remaining seven scenarios were counterbalanced to address potential order effects. Participants engaged in Tasks 1 and 2 respectively for each scenario, before proceeding on to the next scenario.

In order to test for the effect of providing contextual information about the options (e.g., mobile phone brands and specific model information) on consistency, participants were subdivided at random into two treatments. In Treatment 1 (labeled *exposed*) participants received contextual information (Table 1); in Treatment 2 (labeled *blind*) respondents received only generic information to characterize each option (i.e., options were labeled Mobile Phone A, Mobile Phone B, etc.).

Demographic data was collected from each respondent before (age and gender) and after (respondents' political orientation, education level, and annual income) they completed the tasks associated with each scenario.

2.2. Analysis

We compared each respondent's *implied* preference order (based on attribute weights) to their *ranked* preference order and calculated Kendall's coefficient of concordance (W) for each individual across each of the eight scenarios, and across both treatments (exposed and blind).

Kendall's W is typically used to measure the agreement between a group of raters across a single set of alternatives, for instance to show how consistent a group of reviewers are in rating films.

Here we calculated W to represent the agreement of a single individual across each method of elicitation. It may be helpful to know that a Kendall's coefficient is a direct transformation of a Spearman correlation coefficient ($\boldsymbol{\rho}$). So, when $W = 1$ (and $\boldsymbol{\rho} = 1$) perfect agreement exists between an individual's holistic rank and implied preference order. When $W = 0.5$ ($\boldsymbol{\rho} = 0$) no agreement exists between a respondent's holistic rank order and implied preference order, and when W equals zero (or a $\boldsymbol{\rho} = -1$) perfect *dis*agreement exists.

In additional to individual coefficients, overall mean consistency measures ($\overline{W}$) were calculated for each scenario in each treatment. Repeated measures analysis of variance (ANOVA) was used to test for the main effect of both scenario and treatment on consistency, with multiple demographic variables included as covariates. We also tested the interaction effect of the treatment with individual demographic differences, as well as the interaction between the scenarios and treatment. Finally, pairwise t-tests were used to test for differences between mean scenario consistency scores within treatments.

2.3. Respondents

This research was carried out in Canada. A total of 2,382 respondents were randomly drawn using a national probability sample from a pre-recruited web-enabled panel maintained by Insightrix Research. Of these, 508 respondents (21%) were removed from the dataset because they spent less than the minimum time (20 minutes) deemed necessary (based on pretesting) to complete the survey, or because they did not complete the swing-weighting task correctly. In this final sample (n = 1,874), 945 respondents were assigned to Treatment 1 and 929 respondents were assigned to Treatment 2.

Forty-seven percent of the sample self-reported as male and 53% as female. The median age-range was 40-49; 70% of the sample was between the ages of 18 and 60; 66% earned at least $50,000 annually, and all but 25 subjects had at least a high school diploma (or equivalent).

3. RESULTS

3.1. Consistency between treatments

Respondents achieved an overall mean consistency ($\overline{W}$) of 0.71 *(SE* = 0.004) in the blind treatment and 0.67 (*SE* = 0.004) in the exposed treatment. The highest mean consistency demonstrated by an individual in the blind treatment was 0.94 (the lowest mean consistency demonstrated by an individual in this same treatment was 0.17), while in the exposed treatment it was 0.95 (the lowest mean consistency demonstrated by an individual in this same treatment was 0.23).

There was a main effect of treatment on consistency (F = 48.76, df =1, $p < 0.001$; see Table 2) with significantly higher levels of consistency observed in the blind (rather than exposed) scenarios 1, 2, 3 and 5 ($p < 0.01$). Significantly higher consistency was observed in the exposed (rather than blind) treatment of scenario 4 ($p < 0.01$), and no significant difference was found between the blind and exposed consistencies in scenarios 6 and 7. The highest mean level of consistency was observed in the blind treatment of scenario 2 ($\overline{W}_{2\text{-blind}} = 0.81$), and the lowest mean level of consistency was observed in the exposed treatment of scenario 1, ($\overline{W}_{1\text{-exposed}} = 0.58$); see Table 3. The greatest difference between treatments was also observed in scenario 1 ($\Delta\overline{W}_1 = 0.13$).

Insert Table 2 here.

There was a main effect of the scenarios (F = 5.75, df = 1, $p < 0.001$), with 18 of 21 pairwise comparisons between overall mean consistency measures ($\overline{W}$) differing significantly ($p < 0.01$); the exceptions were Scenarios 1 & 7; 3 & 4; 5 & 6; see Tables 2 & 3. There was also a significant interaction effect of the scenarios and the treatment (F = 8.39, df = 6, $p < 0.001$). Regarding this effect, in the simplest scenarios (see S1 Table), i.e., scenarios 1, 2 & 3, individuals demonstrated significantly higher ($p < 0.001$) consistency in the blind treatment than they did in the exposed treatment (Table 3). However, as decisions became more difficult, the difference between treatments was attenuated (i.e., in scenarios 5, 6 & 7).

Insert Table 3 here.

3.2. Consistency within treatments

To understand differences in consistency across the scenarios, we draw attention to the scenario ratings provided in the pre-test and the mean consistency scores in the blind treatment, which relative to the exposed treatment provides a more accurate measure of individuals' consistency, as argued in the introduction. We observed the lowest levels of consistency for the two most difficult and important decision contexts (to society, see Table S1); scenario 6, humanitarian interventions in Syria, $\overline{W}_{6\text{-blind}} = 0.68$, and scenario 7, climate change abatement, $\overline{W}_{7\text{-blind}} = 0.62$. Based on t-tests, these consistency scores were significantly lower ($p < 0.001$) than the consistency scores for the simplest scenarios, i.e., 1, 2 and 3 (see Supporting Information: S2 Table). T-tests showed scenario 2, $\overline{W}_{2\text{-blind}} = 0.81$, to be the highest in consistency ($p < 0.001$); while scenario 3, $\overline{W}_{3\text{-blind}} = 0.75$, demonstrated the next highest consistency ($p < 0.001$).

3.3. Consistency as a function of demographic characteristics

Neither respondents' political orientation, income nor gender had a significant effect on consistency across either treatment ($p > 0.05$). Respondents' age ($F = 4.32$, df =1, $p = 0.038$) had a significant effect, but no discernible trend was demonstrated within treatments. Education ($F = 18.83$, df =1, $p < 0.001$) had a highly significant effect, with individuals who had earned an advanced (i.e., undergraduate, graduate or professional) degree demonstrating significantly higher ($p < 0.001$) levels of internal consistency ($\overline{W}_{\text{Adv.Deg}} = 0.70$) than did respondents who only graduated high school ($\overline{W}_{\text{HS}} = 0.67$). No significant interaction effects were found between these demographic variables and the treatment.

4. DISCUSSION

Choices which are found to be consistent with people's stated priorities would support an articulated values perspective on decision-making [3, 19]. Widespread inconsistency or greater inconsistency for more difficult choices would support a basic values perspective [2, 4]. Our results support the latter perspective, suggesting preferences elicited regarding complex, unfamiliar and risky decisions are less internally consistent than preferences elicited for simpler, more familiar and less risky choices.

Certainly, individuals are capable of establishing preferences across all kinds of contexts [29]; indeed, people have been shown to do it all the time [30, 31]. However, when a decision is highly complex, the degree to which holistically-derived preferences align with a decision-maker's priorities may be called into question. These lower levels of consistency may be due in part to limits regarding the amount of information that people can assess [32, 33], and thus perhaps opens the door to biases that negatively affect the manner in which objectives—or attributes—are prioritized.

In our research, the complexity of the more difficult scenarios was further compounded because these decisions, i.e., about climate change mitigation and humanitarian interventions in Syria, invoked the possibility of "taboo tradeoffs", which represent a form of constitutive incommensurability whereby decision-makers struggle with making tradeoffs across competing alternatives (or attributes) that are symbolically or morally significant [34, 35]. In these two scenarios, respondents were asked to make tradeoffs among saving money, saving lives, and mitigating catastrophic risk. These kinds of choices may lead to lower levels of consistency because the most straightforward way to eliminate incommensurability is to simply avoid the specific tradeoffs that cause it in the first place. Doing so is easy in cases where holistic evaluation is employed; a decision-maker can simply engage in satisficing by downplaying or ignoring certain attributes. Doing so in a decomposed choice task, when each attribute is specifically singled out for prioritization however is far more difficult. In such cases, a lack of consistency between the preference orders generated by these two methods is virtually inevitable.

A decision-maker's ability to establish consistent preferences is also certainly linked to their level of familiarity with the decision context. In this research, we selected the contexts, options, and attributes based on a combination of expert elicitation, e.g., phone interviews with scholars, and a review of popular consumer magazines, i.e., *Consumer Reports*. We did not elicit from a random subsample of respondents a set of options or attributes, which some would argue helps to better ground the decision–making process in respondents' own reality e.g.[e.g., 36]. This lack of familiarity may increase the difficulty of establishing consistent preferences. Others however have considered this position and found that, when eliciting their own attributes, decision-makers often fail to identify key attributes and focus on others not germane to the task at hand [37].

Our experiment included explicit instructions for respondents to consider *only* the information presented to them, which is a practice typical of stated preference methods like contingent valuation [38]. It is important to note that the scenarios and tasks presented to respondents in this experiment were representative of the kinds of situations that the public engages with routinely in the real world (e.g., in consumer choice, during elections, etc.). Rarely in these situations are people asked to provide attributes based on their own values; instead, such information is provided to them by others, sometimes leading to considerable backlash. For example, food labels in the U.S. have been criticized for not displaying information that is in line with consumers' values [39].

A growing body of scientific knowledge also suggests that instinctive, affect-based responses to decision problems—like the ones presented in our study—may preempt a deliberative assessment of options and attributes [40-42]. In some cases—e.g., when certain contexts invoke strong emotions [43, 44], or when certain attributes are prominent [45] —an instinctive positive or negative feeling may intervene when more careful and thoughtful analysis across multiple attributes is in order. Based on these findings, we believe the affect heuristic played an important role in driving preferences in some of the scenarios used here, especially for certain attributes that might have instinctively been singled out as affect-rich by respondents; e.g., military and civilian casualties, the likelihood of global catastrophe, and human rights considerations. These attributes may have received disproportionally more attention during holistic evaluation than in decomposed choice, leading to the lack of consistency.

The affect heuristic may also account for some of the differences we observed between our exposed and blind treatments. Overall, respondents' choices exhibited lower levels of consistency when contextual information such as the brand names or models of phones and cars,

or the specific practices comprising the risk mitigation strategies, were displayed. A range of studies has looked at the effect of brands, and the *perceived* quality of options on individuals' hedonic or sensory preferences [25, 46, 47]. Certain products and brands or alternatives—e.g., an Apple iPhone vs. a Samsung Galaxy, Donald Trump vs. Hillary Clinton, etc.—elicit positive or negative affective responses, which carry significant weight during decision making in the presence or absence of additional details about their individual attributes [25]. In our research, this effect may have been prominent in the scenarios that involved familiar consumer products where brand recognition was likely to be high (mobile phones and automobiles), but became less so in the tougher scenarios, which contained more complex alternatives and less salient attributes.

In addition to examining respondents' consistency under different conditions, we explored the role that demographic differences play. Our results suggest that one's experience, i.e, their age, but even more so their education plays a significant role. Individuals with college or professional degrees made choices that were significantly more consistent than did respondents with lower levels of education. It has been argued that both our values and our moral reasoning develop as a result of the interplay between our cognitive abilities and other socio-cultural factors [48, 49]. While much of this development occurs during adolescence, the ability to deploy values and reason in challenging or unfamiliar contexts likely relies upon—and is fostered by—learning experiences that disrupt existing cognitive structures [50]. Such disruptions, like those afforded by an advanced education, allow people to explore the unfamiliar, and to synthesize and incorporate this information into increasingly differentiated and integrated structures that can be used during the construction of preference.

4.1 Conclusion

Our research offers an answer to the question raised in the introduction: Yes, people do disagree with themselves in that there is significant misalignment between their implied (based on their weighting of attributes) and ranked preference orders.

However, our research also raises a critical question: which approach—preferences based on ranking or preferences based on weighting— results in a more accurate reflection of respondents' priorities? Is it the priorities revealed by respondents' holistically derived preference order, or is it the stated priorities established through respondents' direct weighting of attributes? We rely primarily on the direct weighting of attributes as representing what truly matters to respondents, since this is the approach that focuses respondents' attention *specifically and directly* on the task of priority-setting (i.e., attribute weighting).

As such, it is our contention that achieving high consistency is a function of one's ability to establish holistically-derived preferences that fall in line with these stated priorities. Our results demonstrate that difficult choices resulted in the lowest levels of consistency—though more research is necessary to delineate which specific element, i.e., complexity, familiarity or risk, causes the least consistency. We argued that these findings can be attributed to both cognitive and affective barriers to deliberative reasoning during decision-making, and that higher levels of education help people to overcome these obstacles.

We also acknowledge that some readers of this paper may take an opposite perspective on consistency, arguing that it is more strongly associated with *revealed* preferences and what these choices say about a decision-maker's values. If one adopts this perspective, our main findings still apply in that the disagreement between what people say they want and what they actually choose remains—this presents a serious dilemma for those eliciting the public's preferences,

policy-makers relying on those preferences, and of course the public themselves. However, adopting this alternative perspective would also imply that higher levels of education exacerbate the level of preferential disagreement, a seemingly counterintuitive result.

Taken from our perspective, our results are both hopeful and concerning. On the one hand, they support the notion that education can enhance a decision-maker's capabilities for making choices that are in line with their priorities. On the other, our results highlight what may be a gulf between what individuals say they value, and what they ultimately choose—or prefer, in certain contexts—as well as how different methods may contribute to this gulf. While individuals' consistency was relatively high in the easy scenarios, it remained far from perfect (a W of 0.81 is equivalent to a Pearson's correlation coefficient of only 0.62). And regarding difficult decisions—decisions that were also considered to be most important to society, people were even less consistent. While more work is necessary to establish the meaning of such values in the absolute, it is concerning that society's most consequential decisions may be, relatively speaking, the least reflective of what people actually care about.

The good news is that several techniques exist, which can help people—working individually and in groups—to make higher quality, and what we believe to be more internally consistent decisions [27, 51, 52]. These techniques aim to overcome well-known biases in decision making by helping individuals more thoughtfully consider their values, the consequences of different options to one's values, and the implicit tradeoffs required by a choice. One technique for helping people focus more clearly on their values may be to simply ask individuals to engage in attribute weighting (Task 2) *before* ranking alternatives (here, Task 1); previous work suggests doing so may increase the internal consistency of people's choices [26].

Especially as the complexity and gravity of choices facing individuals and society increases, we would argue that adopting such techniques and continuing to examine the consistency of people's choices across elicitation methods is critical.

REFERENCES

1. Schwartz N. Self reports: How the questions shape the answers. American Psychologist. 1999;54:93-105.

2. Lichtenstein S, Slovic P. The construction of preference: Cambridge University Press; 2006.

3. Fischhoff B. Value elicitation: is there anything in there? American Psychologist. 1991;46(8):835.

4. Slovic P. The Construction of Preference. American Psychologist. 1995;50(3):364-71.

5. Wojcieszak M. Preferences for political decision-making processes and issue publics. Public Opinion Quarterly. 2014;78(4):917-39.

6. Blumer H. Public opinion and public opinion polling. American Sociological Review. 1948;13(5):542-9. doi: 10.2307/2087146.

7. Adamowicz W, Boxall P, Wilhams M, Louviere J. Stated preference approaches for measuring passive use values: Choice experiments and contingent valuation. American Journal of Agricultural Economics. 1998;80:64-7.

8. Louviere JJ, Hensher DA, Swait JD. Stated Choice Methods: Analysis and Application. Cambridge, UK: Cambridge University Press; 2000.

9. Simon H. A behavioral model of rational choice. Quarterly Journal of Economics. 1955;69(1):99-118.

10. Payne JW, Bettman JR, Johnson EJ. Behavioral decision research: A constructive processing perspective. Annual Review of Psychology. 1992;43:87-131.

11. Slovic P, Finucane ML, Peters E, MacGregor DG. Risk as analysis and risk as feelings: Some thoughts about affect, reason, risk, and rationality. Risk analysis. 2004;24(2):311-22.

12. Gilovich T, Griffin D, Kahneman D. Heuristics and biases: The psychology of intuitive judgment: Cambridge University Press; 2002.

13. Alhakami AS, Slovic P. A psychological study of the inverse relationship between perceived risk and perceived benefit. Risk analysis. 1994;14(6):1085-96.

14. Samuelson PA. A note on the pure theory of consumer's behaviour. Economica. 1938:61-71.

15. Tversky A, Kahneman D. The framing of decisions and the psychology of choice. Science. 1981;211(4481):453-8.

16. Tversky A, Slovic P, Kahneman D. The causes of preference reversal. American Economic Review. 1990;80:204-17.

17. Regenwetter M, Dana J, Davis-Stober CP. Transitivity of preferences. Psychological Review. 2011;118(1):42.

18. McKenzie CR. Framing effects in inference tasks—and why they are normatively defensible. Memory & cognition. 2004;32(6):874-85.

19. Von Neumann J, Morgenstern O. Theory of games and economic behavior: Princeton university press; 2007.

20. Louviere JJ. Analysing decision making: Metric conjoint analysis. Sage University Papers Series in Quantitative Applications in the Social Sciences. Newbury Park, CA: Sage Publications; 1988.

21. Keeney RL, Raiffa H. Decisions with multiple objectives: preferences and value trade-offs: Cambridge university press; 1993.

22. Edwards W, Newman JR. Multiattribute Evaluation. Beverly Hills, CA: Sage Publications; 1982.

23. Marler RT, Arora JS. Survey of multi-objective optimization methods for engineering. Structural and Multidisciplinary Optimization. 2004;26(6):369-95. doi: 10.1007/s00158-003-0368-6.

24. Keeney RL. Decision analysis: An overview. Operations Research. 1982;30(5):101-35.

25. McClure SM, Li J, Tomlin D, Cypert KS, Montague LM, Montague PR. Neural Correlates of Behavioral Preference for Culturally Familiar Drinks. Neuron. 2004;44(2):379-87. doi: http://dx.doi.org/10.1016/j.neuron.2004.09.019.

26. Bessette DL, Arvai J, Campbell-Arvai V. Decision support framework for developing regional energy strategies. Environmental Science & Technology. 2014;48:1401-8. doi: 10.1021/es4036286.

27. Bessette DL, Campbell-Arvai V, Arvai J. Expanding the Reach of Participatory Risk Management: Testing an Online Decision-Aiding Framework for Informing Internally Consistent Choices. Risk analysis. 2016;36(5):992-1005.

28. Clemen RT, Reilly T. Making Hard Decisons with Decision Tools. Independence, KY: Cengage Learning; 2014.

29. Carson R, Flores N, Meade N. Contingent valuation: Controversies and evidence. Environmental and Resource Economics. 2001;19(2):173-210. doi: 10.1023/A:1011128332243.

30. Kahneman D, Slovic P, Tversky A. Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press; 1982.

31. Bishop GF, Oldendick RW, Tuchfarber AJ, Bennett SE. Pseudo-opinions on public affairs. Public Opinion Quarterly. 1980;44(2):198-209.

32. Miller GA. The magical number seven, plus or minus two: some limits on our capacity for processing information. Psychological review. 1956;63(2):81.

33. Wilson TD, Lisle DJ, Schooler JW, Hodges SD, Klaaren KJ, LaFleur SJ. Introspecting about reasons can reduce post-choice satisfaction. Personality and Social Psychology Bulletin. 1993;19(3):331-9. doi: 10.1177/0146167293193010.

34. Tetlock PE, Kristel OV, Elson SB, Green MC, Lerner JS. The psychology of the unthinkable: Taboo tradeoffs, forbidden base rates, and heretical counterfactuals. Journal of Personality and Social Psychology. 2000;78(5):853-70.

35. Chang R. Against constitutive incommensurability or buying and selling friends. Philosophical Issues. 2001;11(1):33-60.

36. Keeney RL. Value-focused thinking. A path to creative decision making. Cambridge, MA: Harvard University Press; 1992.

37. Bond SD, Carlson KA, Keeney RL. Generating objectives: Can decision makers articulate what they want? Management Science. 2008;54(1):56-70.

38. Mitchell RC, Carson RT. Using Surveys to Value Public Goods: The Contingent Valuation Method. New York, NY: McGraw Hill; 1989.

39. Hawley KL, Roberto CA, Bragg MA, Liu PJ, Schwartz MB, Brownell KD. The science on front-of-package food labels. Public health nutrition. 2013;16(03):430-9.

40. Loewenstein G, Hsee CK, Weber E, Welch N. Risk as feelings. Psychological Bulletin. 2001;127(2):267-86.

41. Slovic P, Finucane ML, Peters E, MacGregor DG. The affect heuristic. European Journal of Operational Research. 2007;177(3):1333-52.

42. Kahneman D. Thinking, fast and slow. New York: Farrar, Straus and Giroux; 2011.

43. Wilson RS, Arvai JL. When Less is More: How Affect Influences Preferences When Comparing Low and High-risk Options. Journal of Risk Research. 2006;9(2):165-78.

44. Wilson RS, Arvai J. Why less is more: Exploring affect-based value neglect. Journal of Risk Research. 2010;13:399-409.

45. Slovic P. The prominence effect: Confronting the collapse of humanitarian values in foreign policy decisions. In: Slovic S, Slovic P, editors. Numbers and Nerves: Information, Emotion, and Meaning in a World of Data. Corvallis, OR.: Oregon State University Press; 2015. p. 53-61.

46. Siegrist M, Cousin M-E. Expectations influence sensory experience in a wine tasting. Appetite. 2009;52(3):762-5.

47. Lee L, Frederick S, Ariely D. Try it, you'll like it the influence of expectation, consumption, and revelation on preferences for beer. Psychological Science. 2006;17(12):1054-8.

48. Piaget J. The moral judgement of the child: Simon and Schuster; 1997.

49. Kohlberg L. The philosophy of moral development moral stages and the idea of justice. 1981.

50. Rest JR. Why does college promote development in moral judgement? Journal of moral education. 1988;17(3):183-94.

51. Weaver CP, Lempert RJ, Brown C, Hall JA, Revell D, Sarewitz D. Improving the contribution of climate model information to decision making: The value and demands of robust decision frameworks. Wiley Interdisciplinary Reviews: Climate Change. 2013;4(1):39-60.

52. Gregory R, Satterfield T, Hasell A. Using decision pathway surveys to inform climate engineering policy choices. Proceedings of the National Academy of Sciences. 2016;113(3):560-5.

Table 1. Scenarios, options, and attributes. Scenarios are ordered according to increasing complexity. In one between-subjects treatment of this experiment (labeled *exposed*), identifying labels were shown to respondents (e.g., the mobile phone brands in Scenario 1); in the other between-subjects treatment (labeled *blind*), identifying labels were withheld from respondent, and only option numbers were shown (e.g., Mobile Phone 1, Mobile Phone 2, Mobile Phone 3, and so forth). Arrows, (↗) and (↘) indicate the preferred direction of each attribute.

Tutorial. Dinner Out		**Options**				
Attribute	**Units**	*Alice's*	*In Thyme*	*Chop*	*Diplomat*	*The Earle*
Average price of entrée	*Dollars (↘)*	$15	$17	$21	$27	$35
Variety of entrées	*Number of entrees (↗)*	20	13	14	15	8
Critics' ratings of the food	*1-5 star rating (↗)*	2	4	3.5	4	5
Customers' ratings of serv/atmo	*10-point Likert scale (↗)*	7	3	7	5	8
Cleanliness of the kitchen	*Health code violations (↘)*	4	1	3	0	0
Scenario 1. Mobile Phone Purchase						
Attribute	**Units**	*Apple iPhone*	*Samsung Galaxy*	*HTC One*	*Motorola Droid*	*Blackberry*
Price you pay	*Dollars (↘)*	$350	$275	$150	$200	$350
Camera resolution	*Megapixels (↗)*	8	16	5	10	8
Battery life	*Hours of talk time (↗)*	10	17	20	48	10
Memory	*Gigabytes (↗)*	64	16	16	32	64
Quality & reliability	*Consumer Reports: 0-5(↗)*	5	3	4	4	5
Scenario 2. New Vehicle Purchase						
Attribute	**Units**	*Toyota Camry*	*Ford Fusion*	*VW Passat*	*Hyundai Sonata*	*Volvo S60*
Price	*Dollars (↘)*	$27,850	$27,015	$34,100	$23,995	$37,750
Fuel consumption	*Liters of fuel/100km (↘)*	7.8	9	7.7	7.3	10.2
GHG emissions	*Tonnes of CO_2/Yea r(↘)*	4.7	5.1	4.4	4.8	5.4
Safety	*Consumer Reports: 0-5(↗)*	4.7	4.3	4.8	4.8	5.0
Quality & reliability	*Consumer Reports: 0-5(↗)*	4.3	3.3	4.3	3.7	3.0
Scenario 3. Rental Property						
Attribute	**Units**	*Dwntwn 2br Condo*	*Rural 3br Home*	*Subrbn 3br Home*	*Suburban 3br Home*	*Urban 3br Home*
Walkability	*Walk Score (↗)*	100	40	65	25	80
Size	*Square footage (↗)*	1,150	2,600	2,000	1,200	1,600
Commute to work	*Time in minutes by car (↘)*	2	60	40	25	10
Crime rate	*Crimes reported/Yr. (↘)*	450	85	90	275	100
Price	*Dollars per month (↘)*	$2,800	$1,800	$2,000	$900	$2,100

Table 1 (continued). Scenarios, options, and attributes. Scenarios are ordered according to increasing complexity. In one between-subjects treatment of this experiment (labeled *exposed*), identifying labels were shown to respondents (e.g., the mobile phone brands in Scenario 1); in the other between-subjects treatment (labeled *blind*), identifying labels were withheld from respondent, and only option numbers were shown (e.g., Mobile Phone 1, Mobile Phone 2, Mobile Phone 3, and so forth). Arrows, (↗) and (↘) indicate the preferred direction of each attribute.

Scenario 4. Gasoline Purchase						
Attribute	**Units**	*Canada Oil Sands*	*Saudi Arabia*	*Nigerian Delta*	*Venezuela Bitumen*	*USA Offshore*
Price per litre	*Dollars (↘)*	$1.35	$1.35	$1.35	$1.35	$1.35
GHG emissions	*Kilograms CO_2/litre (↘)*	3.7	2.7	3.3	3.3	2.9
Environmtal impact of extraction	*7-point Likert Scale (↘)*	7	1	4	4	2
Human rights score (COI)	*7-point Likert Scale (↗)*	7	1	4	3	7
Environmental performance (COI)	*10-point Likert Scale (↗)*	7.3	6.7	3.9	5.8	6.8
Scenario 5. Investment Portfolios						
Attribute	**Units**	*90% Domestic*	*50% Domestic*	*33% Domestic*	*12% Domestic*	*100% Domestic*
Expected growth in 10 yrs	*% Growth (↗)*	5	10	28	48	97
Level of confidence in growth	*10-point Likert Scale (↗)*	9	5	6	4	1
Volatility of portfolio	*10-point Likert Scale (↘)*	1	3	2	7	9
% invest in sustainable comp.	*% of Portfolio (↗)*	100	89	60	16	5
% invest promoting human rights	*% of Portfolio (↗)*	100	85	91	21	10
Scenario 6. Humanitarian Intervention in Syria						
Attribute	**Units**	*Condem-nation*	*No-Fly Zone*	*Air Strikes*	*Safe Zone*	*Military Engage.*
Civilians likely killed by Assad	*Human Lives (↘)*	1,000,000	250,000	100,000	10,000	1,000
Canadian military casualties	*Human Lives (↘)*	0	5	20	20	1,600
Cost	*Dollars (↘)*	$0	$50B	$20B	$150B	$500B
Duration	*Months (↘)*	0	18	6	36	48
Likelihood of peace in 5 Yrs.	*Probability (%) (↗)*	5	15	40	20	80
Scenario 7. Climate Change Abatement						
Attribute	**Units**	*Status Quo*	*Efficient improve.*	*Decarbon-ization*	*Geoengin-eering*	*CCS*
Potential to reduce GHGs	*5-point Likert Scale (↗)*	0	2	4	0	3
Potential to stabilize glbl temps	*5-point Likert Scale (↗)*	0	1	2	4	2
Catastrophic potential	*5-point Likert Scale (↘)*	0	0	2	3	2
Public opposition	*5-point Likert Scale (↘)*	2	0	4	5	4
Cost	*5-point Likert Scale (↘)*	0	3	5	1	3

Table 2. Effects on consistency. Main and interaction effects of treatment, scenario and key demographic variables. Statistics are result of repeated measures ANOVA using individuals' consistency in each scenario as dependent variable (x7).

	df	F	Partial Eta Squared
Variable(s)			
Scenario	6	5.753**	0.018
Treatment	1	48.761**	0.026
Gender	1	3.168	0.002
Age	1	4.315*	0.002
Political orientation	1	0.141	0.000
Education level	1	18.825**	0.010
Income	1	0.462	0.000
Treatment * Scenario	6	8.386**	0.026
Treatment * Gender	1	2.036	0.001
Treatment * Age	1	1.043	0.001
Treatment * Political orientation	1	0.179	0.000
Treatment * Education level	1	0.919	0.000
Treatment * Income	1	0.568	0.000

*$p < 0.05$; **$p < 0.001$

Table 3. Consistency scores. Mean consistency scores ($\overline{W}$) and standard deviation (*SD*) are shown for each scenario (per treatment and overall). Differences in ($\overline{W}$) based on ANOVA are shown with a * for $p < 0.01$ and ** for $p < 0.001$. Statistics reflect all 1874 participants.

Scenario	$\overline{W}$ *(SD)*	
Tu. Dinner	**Treatment**	**Overall**
T1 Exposed **T2** Blind	- -	**0.62**[a] (0.28)
1. Mobile Phone Purchase		
T1 Exposed **T2** Blind	**0.58** (0.26) **0.71**** (0.22)	**0.64** (0.25)
2. New Vehicle Purchase		
T1 Exposed **T2** Blind	**0.73** (0.24) **0.81**** (0.21)	**0.77** (0.23)
3. Rental Property		
T1 Exposed **T2** Blind	**0.70** (0.22) **0.75**** (0.21)	**0.73** (0.22)
4. Gasoline Purchase		
T1 Exposed **T2** Blind	**0.74**** (0.20) **0.69** (0.20)	**0.71** (0.20)
5. Investment Portfolios		
T1 Exposed **T2** Blind	**0.67** (0.31) **0.71*** (0.30)	**0.69** (0.30)
6. Humanitarian Intervention in Syria		
T1 Exposed **T2** Blind	**0.66** (0.22) **0.68** (0.23)	**0.67** (0.23)
7. Climate Change Abatement		
T1 Exposed **T2** Blind	**0.63** (0.23) **0.62** (0.25)	**0.63** (0.24)

[a] All participants saw the same Tutorial scenario.

ACKNOWLEDGEMENTS

To be included.

SUPPORTING INFORMATION

S1 Table. Scenario characteristics. Table displays mean statistics for each scenario based on pre-test (n = 50). Standard deviations are shown in parentheses.

S2 Table. Correlation coefficients and pairwise comparisons of scenario consistency scores. Table displays spearman correlation coefficients (δ) and results of paired sample t-tests for each scenario ($df_{T1} = 944$; $df_{T2} = 928$). Statistics are significant at $p < 0.001$ unless noted with a * ($p < 0.05$) or $^{\wedge}$ ($p > 0.05$).