Journal of the Economic Science Association (2024) 10:310–329
https://doi.org/10.1007/s40881-024-00161-x

ORIGINAL PAPER

A behaviorally validated warm glow questionnaire

Jeffrey Carpenter[1] · **Alex Lyford**[2] · **Mingfang Zhang**[3]

Received: 22 July 2022 / Revised: 4 January 2024 / Accepted: 5 January 2024 /
Published online: 7 March 2024

Abstract
Measuring the social preferences of economic agents using experiments has become common place. This process, while incentive compatible, is costly and time consuming, making it infeasible in many settings. We combine standard altruism and warm glow choice experiments with a battery of candidate survey questions to construct behaviorally validated questionnaires. We use machine learning to create parsimonious 3-question modules that reliably replicate existing results on general altruism and provide an alternative method for collecting warm glow preferences.

Keywords Experiment · Altruism · Warm glow · Survey validation

JEL Classification C91 · D64 · D91 · H41

1 Introduction

The purely altruistic model of charitable giving predicts that donors don't care where funding comes from as long as the public good is provided. A dollar from taxes, corporate giving or any other source is just as good as a dollar from the donor's own pocket. As a result, other sources of funding should *crowd out* the contribution provided by the altruist. This seems, however, to not be the case. Early studies by Abrahams and Schmitz (1984) and Clotfelter (1985) estimated government grants did little to crowd out private donations to public goods. Since then, a similar lack of crowding has been shown by Payne (1998) among many others and is reviewed in Ribar and Wilhelm (2002). Consequentially, altruism is now thought to be *impure* in the sense that donors must also receive private benefits from the contributions they make. This private benefit, developed in Andreoni (1990), has been termed *warm*

✉ Jeffrey Carpenter
jpc@middlebury.edu

[1] IZA and Department of Economics, Middlebury College, Middlebury, VT 05753, USA

[2] Department of Mathematics, Middlebury College, Middlebury, USA

[3] Department of Economics, Middlebury College, Middlebury, USA

glow. Knowing the extent to which donors are motivated by warm glow has important implications for tax policy (e.g., Andreoni, 2015) and the design of charitable fundraisers (e.g., Lilley & Slonim, 2014; Carpenter, 2021).

The standard way to measure altruistic preferences is using an incentivized experiment, typically the *dictator game*. Here, participants are asked to contribute any amount of a given (or earned) endowment to a selected charity (Eckel & Grossman, 1996) or a charity of their own choosing (Carpenter et al., 2008). The typical endowment in this experiment is at least $10, there are often around 100 participants and a session usually lasts around an hour. In other words, after an hour's work the researcher has spent $1000 and gathered just 100 observations. There must be a more efficient way to collect a substantial sample of altruistic preferences, especially when financial incentives are not feasible.

We run over 300 participants of varied demographics through both a similar incentivized experiment, altered to detect both generally altruistic and warm glow donations, and a large battery of survey questions designed to predict behavior in the experiment. Our goals are to replicate the altruism module created in Falk et al. (2016) with new methods and a new sample and to create a behaviorally validated warm glow questionnaire that researchers can employ much cheaper, much quicker and at much greater scale. Indeed, using a variety of machine learning techniques, we find that just three survey questions can predict incentivized donations in both the general and warm glow contexts with high predictive power. Combined, these 3-question modules constitute short survey instruments that are forged in incentivized behavior and are easy to implement within much larger survey instruments or experimental investigations.

2 Methods

2.1 Incentivized measures of altruism and warm glow

To gather incentivized measures of altruism and warm glow, we modeled our experiment on Crumpler and Grossman (2008) and Carpenter et al. (2008). In this variant of the dictator game, participants were awarded a $2 bonus beyond their participation payment, any fraction of which they could donate to a charity of their choice. The survey is reproduced in the appendix. The standard dictator game (our measure of altruism) was implemented in questions 10–14. Question 10 (*Q10*) provided instructions, *Q11* and *Q12* asked comprehension questions, *Q13* asked participants to choose one of six charities (American Cancer Society, American Red Cross, Doctors Without Borders, World Wildlife Fund, Save the Children or Teachers Without Borders), and *Q14* recorded one's donation using a slider to determine how much of the bonus to donate to their chosen charity and how much to keep. The corresponding questions for the warm glow version of the experiment were *Q5–Q9*.

What made the warm glow version of the dictator game different was that participants were informed that the proctor of the experiment would also donate and that the amount going to the charity would be fixed at $2. The instructions emphasized that "the amount contributed by the proctor will be reduced by the amount

you donate. In other words, your selected charity will always receive exactly $2." Regardless of the amount the participant donated, the sum of her contribution and the proctor's was fixed at $2. For example, as indicated in the instructions, "If you donate $0, the proctor will donate $2; if you donate $1, the proctor will donate $1 and if you donate $2, the proctor will donate $0." The purpose of this variant is to disincentivize any purely altruistic giving—individual donations do not increase the total contribution and, therefore, can only be motivated by warm glow.

2.2 Candidate survey measures

We asked our participants a variety of survey questions that we hypothesized would predict giving in our experiment. Some of the questions we borrowed from the existing literature were designed to detect altruism, broadly speaking, while others we created were more focused on warm glow. To help the exposition, we relabel the questions (and specific responses) with reference to whether they were designed to measure altruism (*A*) or warm glow (*WG*). Based on the success of the questions developed in Falk et al. (2016), we borrowed a hypothetical choice question designed to capture altruism. *A33* asked how much the respondent would donate to charity after winning $1000 in a lottery. We also developed three new variations of this question: first, *A31* asked how much of an unanticipated $1000 bonus for doing a good job at work would the respondent donate to charity. The main difference between these two questions is that the bonus was earned in the first and not in the second. Prior research has shown increases in warm glow giving if the endowment is earned (Luccasen & Grossman, 2017). Second, we created two complementary versions of these questions that targeted warm glow giving (*WG32* and *WG34*) by adding a phrase indicating that a friend or your employer already planned to give $1000 to charity and would reduce the amount they give by whatever you give. Borrowing again from Falk et al. (2016), we also asked a qualitative self-assessment about how willing respondents were to give to good causes without any expectation of a return (*A25*).

To further measure altruistic preferences among our participants, we included two questions (*A27* and *A28*) from a self-reported altruism scale developed by Rushton et al. (1981). The original scale contains twenty questions that elicit the frequency that respondents engage in specific altruistic behaviors, like donating to strangers and charity. In his research measuring altruistic behaviors, Bekkers (2007) discovered that the anonymity of decisions lowered generosity. It is intuitive that warm glow givers may feel better about themselves if donations are made in their names. With this in mind, we added two questions (*WG29, WG30*) to evaluate any impact of recognition on warm glow giving. Nunes and Schokkaert (2003) conducted a contingent valuation study to investigate the effects of altruism on one's willingness to pay for natural resources. We included one of their questions (*A26*) that is associated with the social pressure to donate.

As a last thought on general altruism, we realized that altruism may also affect purchasing decisions. Koschate-Fischer et al. (2012), for example, show how an increasing number of companies now direct some of their earnings to charity

to enhance customer goodwill and improve their brand images. Motivated by this trend, we asked participants (in *A36*) to, consider two identical products and report how much more they would be willing to pay for the one produced by a firm that donates profits to charity.

Considering items that were designed to be warm glow specific, the first question is based on a survey module implemented in the field experiment conducted by Carpenter (2021), which found that the responses to the following question predicted warm glow giving in a similar setting. Respondents were asked, "Think about the last time you gave to charity that needed a fixed amount of money to accomplish a goal (e.g., $1000 to provide a holiday meal to people in need). What was most important to you?" They could give a purely altruistic response, "The total amount given by everyone" (*A21*), a warm glow response, "The amount that you personally gave" (*WG21*) or a catchall third response, "Some other aspect of giving." In the same vein, two other questions were constructed: "What is the most important reason that you donate to charity?" with responses to benefit others (*A22*), self-image or to feel good (*WG22*) and "What do you think is the most common reason that other people donate to charity?" with the same responses (*A23* and *WG23*).

Andreoni's (1990) theory guides the response options of two other questions, one dealing with the possible ways to finance public goods and another one that asks for emotional associations. Public goods can be funded by taxes or private donations. Both ways increase the total supply, but warm glow donors should feel differently about making private donations. In question *A24*, participants are asked how best to fund general public goods like infrastructure and a specific public good—the road in one's neighborhood. In another question, participants describe which factors influence charitable donations: one's ability to help (*A35*) or feeling guilty (*WG35*).

The last two questions of the survey were hypothetical versions of the dictator experiments. For the standard game, we asked (*A38*): "Imagine participating in the following experiment. You are given $100 and told that you can donate as much of the $100 as you like to a charity of your choice and keep the rest for yourself. You can donate any amount between $0 and $100 to your selected charity." For the hypothetical warm glow experiment we asked (*WG37*), which started identically to the previous question but had a different ending: "... You can donate any amount between $0 and $100 to your selected charity. The proctor of the experiment will also donate up to $100 to your selected charity. However, the amount contributed by the proctor will be reduced by the amount you donate. In other words, your selected charity will always receive exactly $100. For example, if you donate $0, the proctor will donate $100. If you donate $50, the proctor will donate $50 and if you donate $100, the proctor will donate $0."

2.3 Recruitment and protocol

The survey experiment was conducted online with a U.S. nationally representative sample from Prolific's participant pool. To ensure the quality of their responses, the survey began with a reCAPTCHA question (*Q2*) for each participant. In addition, the survey access was limited to Prolific participants who had more than twenty-five

prior survey approvals (but no more than 200) and an approval rating higher than 90%. To proceed, consent (*Q3*) was required for every participant at the beginning of the survey. The experiment was approved by the Middlebury College Institutional Review Board.

The survey experiment was composed of four blocks and the questions within a block were presented in random order. The first block of the survey included the two incentivized dictator experiments. Participants were told that one of these two dictator responses would be chosen randomly and implemented at the end of the session. In the second block of the survey, participants were asked six demographic questions: gender, age, zip code, education level, political preference, and religious attitude. In the third block, we asked the sixteen questions designed to differentiate between pure altruism and warm glow. The survey ended with the fourth block in which we presented the hypothetical versions of the dictator experiments. Blocks two and three were strategically placed between the dictator experiments and the hypothetical experiments to attenuate the natural demand for consistency between the two.

After participants completed the questionnaire, they were redirected back to the Prolific website and their participation was recorded automatically. The bonuses for individuals were calculated and distributed within a day. Donations for the chosen charities were totaled and submitted online.

3 Data overview

3.1 Participants

A total of 310 people participated in the study. Considering the demographic characteristics of our nationally representative sample, 51% were female, their ages varied from 18 to 93 years (the mean was 45), 56% graduated from college, 25% were politically conservative to some extent and 40% said that they were not at all religious. On average, participants spent 8 minutes completing the survey, for which they earned a base payment of $1 and an average bonus of another $1.30.

3.2 Dictator game experimental results

Our primary goal is to identify which of the myriad candidate survey questions are useful in understanding a participants' propensity to make donations in the two dictator games. Table 1 summarizes the results of our incentivized experiments. In the standard, altruism, version of the dictator game, while almost 28% of people give nothing, the majority of participants donated a positive amount ($0.72 on average), including 11.4% who give the entire $2. As expected, contributions were lower in the warm glow dictator game because we turn off the purely altruistic motivation. Here, 42.9% of people give nothing, the average amount donated falls to $0.55, but 7.1% of people still give everything.

Table 1 Donation amounts in altruism and warm glow choice tasks

Amount (in dollars)	Altruism N (% of total)	Warm glow N (% of total)
0	86 (27.7%)	133 (42.9%)
(0, 0.5]	63 (20.3%)	50 (16.1%)
(0.5, 1]	108 (34.8%)	95 (30.7%)
(1, 1.5]	15 (4.8%)	10 (3.2%)
(1.5, 2)	3 (1.0%)	0 (0.0%)
2	35 (11.4%)	22 (7.1%)

4 Survey module creation and replication

We begin our analysis of what predicts donations by trying to replicate (Falk et al., 2016) using the standard dictator game donation (*A14*), an expanded set of candidate questions and a representative sample from the U.S. Our major contribution, however, is the creation of a new survey module to predict warm glow giving. In both instances, we examine the intensive (Donation Amount) and extensive (Donated?) margins. The former approach treats the donation amount as a quantitative, variable and the latter approach classifies a donation as an indicator. Regardless of approach, the analyses lead to largely similar results, with only small differences between which survey questions are most important in predicting the given donation variable. In what follows, we describe our approach to both analyses. In addition, we provide easy-to-use scoring rubrics to predict giving.

4.1 Raw correlations

To begin our analysis, we report the raw correlations between the donation experiments and the survey questions in Tables 2 and 3. Considering general altruism, we see that 10 of the 13 questions are highly correlated with both the amount participants donated in the standard dictator game and an indicator for donating any amount. Of particular note are the strong correlations with *A38*, *A33* and *A31*, all hypothetical versions of making an actual donation. Interestingly, we find similar results for warm glow in Table 3. Here, the hypothetical lottery donations in *W32* and *W34* are strongly correlated with actual warm glow donations (and having donated) and the hypothetical experiment *WG37* is very highly correlated with both measures of giving.

4.2 Donation amount (intensive margin)

We utilized three different techniques to model the relationship between respondents' answers to the survey questions and the amount that they chose to donate in *A14* and *WG9*: multiple regression, regression trees, and random forests. A regression tree is a type of model that makes predictions about an outcome based on a

Table 2 Raw correlations between dictator donations (*A*14) and altruism survey responses

Question	$A14 > 0$	*A*14
A21: total amount given by everyone	0.08	0.04
A22: acting for the good of others	−0.05	−0.10*
A23: acting for the good of others	0.16***	0.16***
A24: tick box and donate	0.11**	0.10*
A25: very willing	0.20***	0.15***
A26: yes	0.32***	0.23***
A27: yes	0.18***	0.16***
A28: yes	0.32***	0.23***
A31: bonus donation	0.44***	0.31***
A33: lottery donation	0.46***	0.31***
A35: my ability to help or donate	0.07	0.10*
A36: salad dressing WTP	0.23***	0.18***
A38: hypothetical experiment	0.58***	0.37***

Correlation coefficients; $*p < 0.10$, $**p < 0.05$, $***p < 0.01$

Table 3 Raw correlations between warm glow donations (*WG*9) and warm glow survey responses

Question	$WG9 > 0$	*WG*9
WG21: amount that you personally gave	0.08	0.05
WG22: donating makes me happy	0.05	0.05
WG23: donating makes them happy	−0.11*	−0.07
WG29: yes	0.09	0.05
WG30: yes	−0.02	−0.01
WG32: bonus donation	0.33***	0.46***
WG34: lottery donation	0.31***	0.40***
WG35: how guilty I will feel if I do not donate	−0.11**	−0.06
WG37: hypothetical experiment	0.52***	0.64***

Correlation coefficients; $*p < 0.10$, $**p < 0.05$, $***p < 0.01$

series of binary decisions. Here, we give a brief introduction to regression trees and random forests, but more in-depth descriptions can be found in Hastie et al. (2001).

Although specific tree-building algorithms vary, a regression tree generally begins by iterating through all possible predictor variables (in our case, survey questions) and identifying the variable that can best split the respondents into two homogenous groups—in our case, one with large average donation amounts and one with small average donation amounts. This approach uses search algorithms to select a breakpoint for a variable, determine how good the split is, and then iterate through all possible breakpoints and variables (Hastie et al., 2001). Figure 1 shows an example one-step regression tree for these survey data. This tree uses the variable *WG34*, the question that reads:

WG34: Suppose that you won $1000 in a lottery. Considering your current situation, how much would you donate to charity if your friend told you that they already

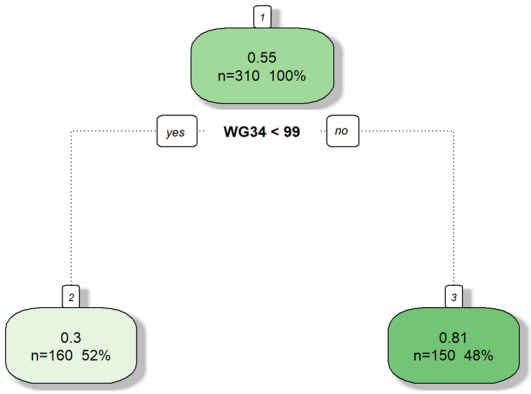

Fig. 1 A one-step regression tree for survey data

planned to give $1000 to charity and they would reduce the amount they would give by whatever you give?

The algorithm found that the best breakpoint for this variable is $99. Respondents who answered that they would donate less than $99 in *WG34* donated an average of $0.30 in the warm glow experiment (*WG14*) while respondents who said they would donate equal to or more than $99 in *WG34* donated an average of $0.81 in *WG14*. A total of 52% of respondents ($n = 160$) said they would donate less than $99, and 48% of respondents ($n = 150$) said they would donate $99 or more.

Thus, if we were to use this regression tree to make a prediction about a new participant's donation, we would predict that they would donate $0.30 if they answered *WG34* with a value less than $99, and we would predict they would donate $0.81 if they answered *WG34* with a value greater than or equal to $99. This *WG34* variable and corresponding breakpoint of $99 were deemed by the tree-building algorithm to be the best possible split into a group of respondents who donated little and those who donated a lot.

A regression tree, however, doesn't typically end after just one split or use a single variable to make predictions. The same process described above is repeated in order to determine if the remaining data can be further split into sub-groups of participants who donated more (on average) and participants who donated less (on average) using all combinations of variables and breakpoints. This splitting process is repeated until a set of stopping criteria are reached, typically related to the overall number of respondents in each terminal node (if there are too few respondents in a given node, then the regression tree might overfit the data). Figure 2 shows a regression tree predicting the donation amount in *WG9* using two variables and many nodes. This regression tree predicts donation amounts as low as $0.15 and as high as $1.40 depending on responses to questions *WG34* and *WG23*. The full regression tree using all variables is too large to meaningfully visualize, but its results are found in Tables 4 and 5.

A random forest is a compilation of hundreds of regression trees, such as the one in Fig. 2. In this analysis, our random forest contains 500 regression trees. Each of the 500 regression trees is built in the same manner as described above with one

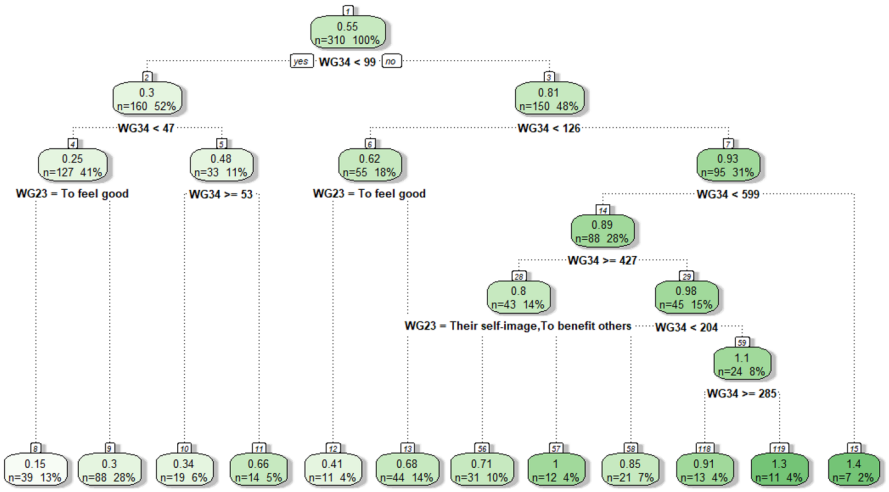

Fig. 2 Large regression tree for predicting *WG9*

Table 4 Model metrics for predicting *Donation Amount* in *A14*

Model	MSE	Adjusted R^2
Multiple regression	0.187	0.535
Regression tree	0.208	0.503
Random forest	0.189	0.538

Table 5 Model metrics for predicting *Donation Amount* in *WG9*

Model	MSE	Adjusted R^2
Multiple regression	0.147	0.576
Regression tree	0.193	0.548
Random forest	0.159	0.601

caveat: each tree only has access to a random subset of data, and each split in the tree only has access to a random subset of all the questions. Each of the 500 regression trees is therefore created independently, and each tree makes a prediction about the total amount a participant will donate based only on the subset of data and variables to which it has access. The prediction of the random forest is the average of all 500 trees' predictions. Each of the 500 regression trees built in this manner is weaker than the regression tree built on the entire data set with all variables included, but the averaging of hundreds of these trees often leads to better overall prediction accuracy. This fact is not necessarily intuitive, but it is proven in Dzeroski and Zenko (2004) and is corroborated by our data. Tables 4 and 5 show model metrics for each of the three approaches using all survey responses (both altruism and warm glow) as predictors variables and donation amounts in *A14* and *WG9* as responses variables,

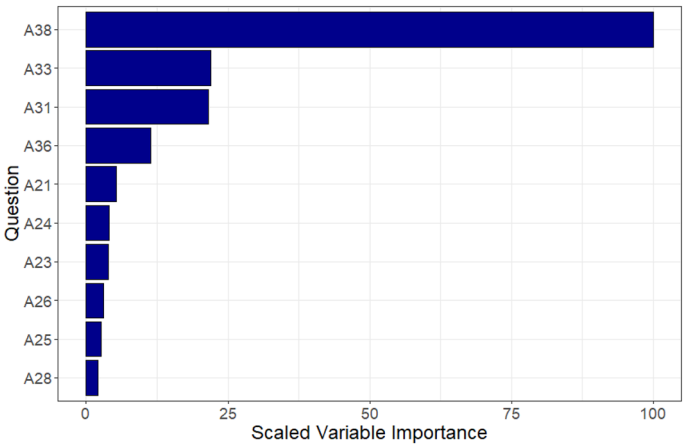

Fig. 3 Variable importance plot for the random forest modeling *Donation Amount* in *A14*

respectively. We used 5-fold cross-validation to estimate the mean squared error (MSE) and adjusted R^2 values of each of the models.

Each of the three techniques yielded similar results across predictions of *A14* and *WG9*–measures of altruistic and warm glow giving. Random forests slightly outperformed the other two techniques based on *MSE* and adjusted-R^2 metrics, significantly more so in the *WG9* model. Using the random forest model, knowing a respondent's answers to all survey questions explained approximately 60% of the total variation in donation amount in *WG9* and 54% of the total variation in donation amount in *A14*.

We next sought to identify a more parsimonious subset of the survey questions with similar predictive power to the full model. All three approaches allow for the quantification of each variable's unique importance in the predictive power of the model. Figures 3 and 4 show the Variable Importance Plots for each of the full random forest models highlighting the most important questions that predict warm glow giving (*WG9*) and altruistic giving (*A14*). The *y*-axis lists each of the variables (survey questions), and the *x*-axis—*Scaled Variable Importance*—is a scaled measure of how much predictive power would be lost if that variable were removed from the model. Specifically, *Scaled Variable Importance* approximates the average decrease in the predictive power of the model if the given variable were removed from the random forest. The most important variable has a scaled importance of 100, and a scaled importance of 50 indicates that variable is half as important as the most important variable.

Each graph shows that the hypothetical dictator game questions about general altruism (*A38*) and warm glow giving (*WG37*) are most important when predicting our measure of altruistic giving (*A14*—Fig. 3) and warm glow giving (*WG9*—Fig. 4). These variables make up the bulk of the predictive power for their corresponding models. Other variables such as *A33*, *A31*, and *A36* were also important in predicting donation amount in *A14*, whereas *WG32* and *WG34* were important in

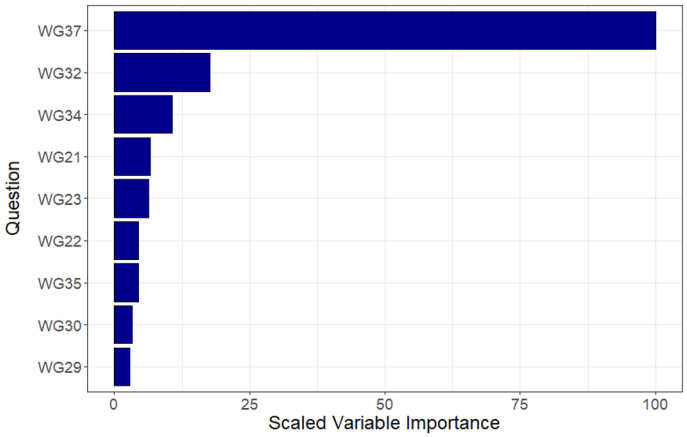

Fig. 4 Variable importance plot for the random forest modeling *Donation Amount* in *WG9*

Table 6 Metrics for models predicting *Donation Amount* in *A14*

Number of variables	Model	MSE	Adjusted R^2
3	Multiple regression	0.268	0.347
3	Regression tree	0.215	0.501
3	Random forest	0.187	0.559
62	Multiple regression	0.187	0.535
62	Regression tree	0.208	0.503
62	Random forest	0.189	0.538

predicting *WG9*.[1] After identifying these important variables, we decided to compare the accuracy of the three original methods: multiple regression, regression trees, and random forests when using a 3-variable subset for predicting *A14* and *WG9*. In other words, we evaluated the predictive cost of using more parsimonious scales.

Tables 6 and 7 show the results of the best 3-variable-subset and full models for each of the three approaches to predicting *A14* and *WG9*, respectively. In the 3-variable subset models, regression trees and random forests perform nearly identically, with slight decreases in R^2 for the 3-variable model. The multiple regression model performs worse than the other two models, and its predictive power drops significantly from the full model to the 3-variable model. This decreased performance for the regression model is due to the complexity of the interactions between individual survey responses and donation amount. For example, respondents indicating that they would donate large amounts in a hypothetical warm glow scenario—such as

[1] It is also reassuring to see that question *WG21* which has been used in Carpenter (2021) also performs relatively well here.

Table 7 Metrics for models predicting *Donation Amount* in *WG9*

Number of variables	Model	MSE	Adjusted R^2
3	Multiple regression	0.246	0.417
3	Regression tree	0.196	0.546
3	Random forest	0.142	0.583
62	Multiple regression	0.147	0.576
62	Regression tree	0.193	0.548
62	Random forest	0.159	0.601

WG32, *WG34*, or *WG37*—donated more on average in *WG9*. However, respondents who donated a lot in one or two of these scenarios and then donated little in the remaining scenarios were more likely to donate a medium-to-low amount in *WG9*. In other words, inconsistencies in donation amounts for the hypothetical scenarios often yielded lower-than-expected actual donations in *WG9*, even if most of the hypothetical donations were large. This complex structure was better captured by the tree-based algorithms than the standard regression model.

For regression trees and random forests, using a 3-variable subset yields a similar predictive power as using all 62 variables (i.e., including both altruism and warm glow variables), but with many fewer survey questions required. Further, because the regression tree and random forest approaches performed roughly equally for the subset models, we based our scoring rubric off of the regression tree model, as its output is more easily interpretable (making predictions with a random forest requires access to each of the 500 regression trees).

Based on our original survey, the best 3-question subset to identify the amount of altruism giving (*A14*) using a regression tree is:

A31) Suppose that at your job, you earned a bonus that exceeded your expectations by $1000. Considering your current situation, how much would you donate to charity?

A33) Suppose that you won $1000 in a lottery. Considering your current situation, how much would you donate to charity?

A38) Imagine participating in the following experiment. You are given $100 and told that you can donate as much of the $100 as you like to a charity of your choice and keep the rest for yourself. You can donate any amount between $0 and $100 to your selected charity.

Figure 5 shows the regression tree-based rubric for predicting altruistic giving. Each node in the tree contains three values: the top value is the predicted donation amount for all respondents belonging to that node, the bottom-left value is the

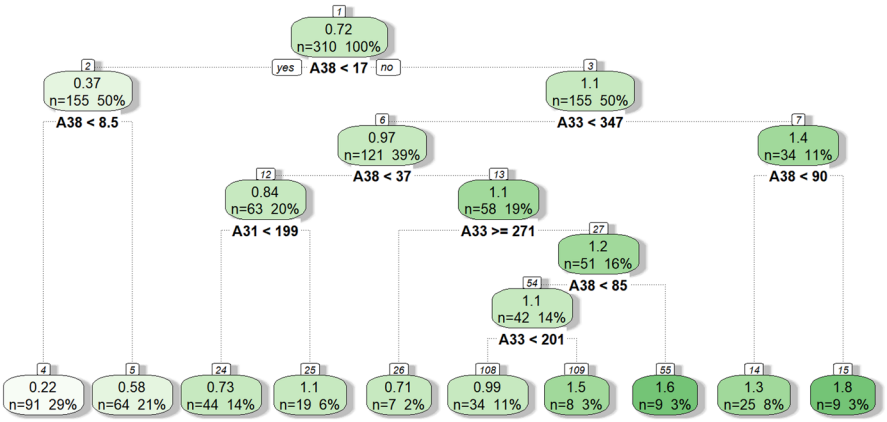

Fig. 5 Rubric for administering 3-question survey for warm glow giving

total number of respondents belonging to the given node, and the bottom-right value is the percentage of respondents belonging to the given node. Nodes that are more green indicate higher predicted donations, and nodes that are more white indicate lower predicted donations. The statements in bold below each node represent the splitting rule for moving to one of the next two nodes.

The rubric demonstrates that those answering with proportionally small values for *A38* donate little in *A14*. Those answering with large values are likely to donate most of their money in *A14*. The middle of the regression tree highlights how to predict donations in *A14* for respondents who mix and match low and high donations in *A31*, *A33*, and *A38*. To see how to use this rubric, suppose a respondent answered the survey in the following manner: a \$500 donation to charity in *A31*, a \$100 donation in *A33*, and a \$50 donation in *A38*. This path through the rubric in Fig. 5 would start by going right at node 1 because they donated more than \$17 in *A38*. At node 3, one would follow the rubric to the right because they donated more than \$347 in *A33*. Finally, at node 7, one would follow the rubric to the left the respondent donated less than \$90 in *A38*. Thus, our rubric predicts that this sample participant would donate \$1.30 in *A14*.

Based on our original survey, the best 3-question subset to identify the amount of warm glow giving (*WG9*) using a regression tree is:

WG32) Suppose that at your job, you earned a bonus that exceeded your expectations by \$1000. Considering your current situation, how much would you donate to charity if your boss told you that the company already planned to give \$1000 to charity and it would reduce the amount it gives by whatever you give?

WG34) Suppose that you won \$1000 in a lottery. Considering your current situation, how much would you donate to charity if your friend told you that

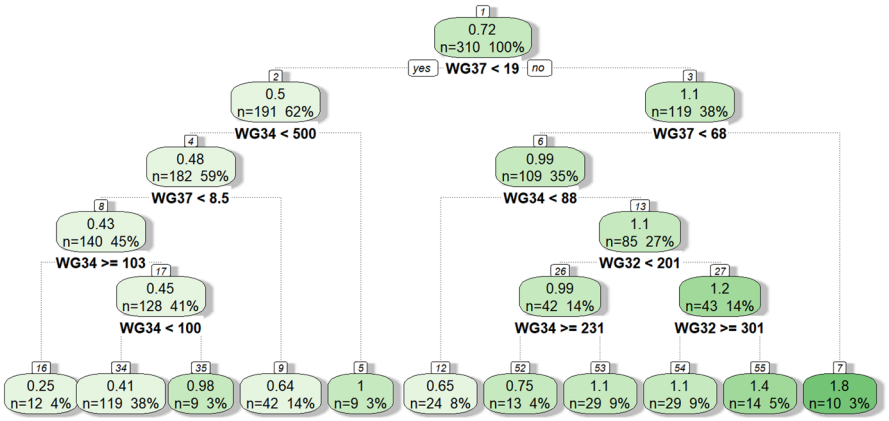

Fig. 6 Rubric for administering 3-question survey for warm glow giving

they already planned to give $1000 to charity and they would reduce the amount they would give by whatever you give?

WG37) Imagine participating in the following experiment. You are given $100 and told that you can donate as much of the $100 as you like to a charity of your choice and keep the rest for yourself. You can donate any amount between $0 and $100 to your selected charity. The proctor of the experiment will also donate up to $100 to your selected charity. However, the amount contributed by the proctor will be reduced by the amount you donate. In other words, your selected charity will always receive exactly $100. For example, if you donate $0, the proctor will donate $100. If you donate $50, the proctor will donate $50 and if you donate $100, the proctor will donate $0.

The rubric demonstrates that those answering with proportionally small values for *WG34* and *WG37* are likely to donate little in *WG9*. Those answering with large values are likely to donate most of their money in *WG9*. Those who answer with medium-sized values (or those who answer with large values for one question but small values for another) are likely to donate a medium amount of their money in *WG9*. To see how to use this rubric, suppose a respondent answered the survey in the following manner: a $200 donation for the lottery question in *WG34*, a $5 for the donation in *WG37*, and a $50 donation in *WG32*. This path through the rubric in Fig. 6 would go left at each of the splits, since the participant donated less than $19 in *WG17*, less than $500 in *WG34*, less than $8.50 in *WG37*, and more than $103 in *WG34*. Thus, our rubric predicts that the sample respondent would donate $0.25 in *WG9*.

Table 8 Model metrics for three different approaches to predicting *A14_Donated*

Model	Accuracy	Kappa	Sensitivity	Specificity
Logistic regression	0.638	0.223	0.836	0.375
Decision tree	0.725	0.414	0.892	0.503
Random forest	0.703	0.350	0.943	0.383

Table 9 Model metrics for three different approaches to predicting *WG9_Donated*

Model	Accuracy	Kappa	Sensitivity	Specificity
Logistic regression	0.664	0.315	0.706	0.609
Decision tree	0.816	0.628	0.807	0.827
Random forest	0.800	0.593	0.813	0.782

4.3 Donated (extensive margin)

We also modeled the relationship between responses to each of the survey questions and whether or not the participant donated any positive amount in *A14* and *WG9*, henceforth referred to as *A14_Donated* and *WG9_Donated*. The approach to modeling responses to *A14_Donated* and *WG9_Donated* as binary variables was similar to our approach to modeling the continuous donation amount, and our results were likewise similar. We began by again building three models: logistic regression, decision trees, and random forests. Decision trees are formed in a nearly-identical manner to regression trees, and are used when the response variable of interest is categorical instead of quantitative. Like regression trees, decision trees identify the best splits based on the homogeneity of the two resulting groups. In the case of decision trees, Gini Impurity is used to quantify homogeneity (D'Ambrosio and Tutore, 2011). Groupings with perfect homogeneity (e.g., all respondents decided not to donate) have a Gini Impurity of 0, and groupings with perfect heterogeneity (e.g., half of respondents decided to donate and half decided not to donate) have a Gini Impurity of 0.25. Thus, the decision tree algorithm attempts to minimize the Gini Impurity at each split. A random forest is a collection of decision trees, and the overall prediction of the random forest is the most common vote from each of the 500 decision trees.

We began by fitting the three models using responses to all survey questions to predict the *A14_Donated* and *WG9_Donated* variables. We calculated accuracy, Cohen's Kappa, sensitivity, and specificity to measure the goodness-of-fit of our models, and the results of the three models predicting *A14_Donated* and *WG9_Donated* can be seen in Tables 8 and 9, respectively. Cohen's Kappa is a measure of the agreement between two raters (in this case, our model and the truth) *not* due to random chance (McHugh, 2012). A Kappa value of 1 indicates perfect agreement and a Kappa value of −1 indicates perfect disagreement. A Kappa value of 0 indicates that any agreement between the model and the true value of the response variable was due to random chance. Typically, Kappa values greater than 0.6 indicate strong agreement, although this often depends on the context of the problem.

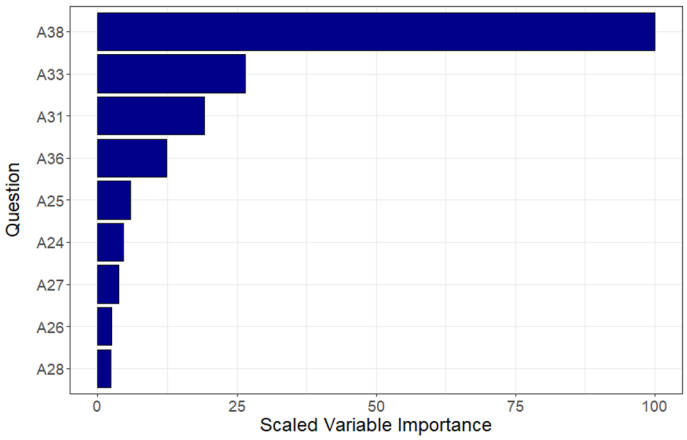

Fig. 7 Variable importance plot for the random forest modeling *A14_Donated*

For calculating sensitivity and specificity, we treated the participant donating as the positive case. We used 5-fold cross-validation to estimate our metrics for each of our models.

When predicting for both *A14_Donated* and *WG9_Donated*, the random forest and decision tree outperformed logistic regression. The sensitivity across all models was extremely high, indicating that all three model types had a relatively easy time correctly predicting when a participant would donate some amount of money in either *A14_Donated* or *WG9_Donated*. The specificity of the models, especially for those predicting *A14_Donated*, varied greatly. The specificity for the Decision Tree when predicting *WG9_Donated* was quite high, indicating that the model could very often correctly predict when a participant wouldn't donate any money in *WG9_Donated*. Conversely, the specificity of the random forest and logistic regression in predicting *A14_Donated* was low.

Like our approach with the intensive margin, we chose to utilize random forests to first investigate which survey questions best predicted whether or not a participant chose to donate. The variable importance plots for the random forest models are shown in Figs. 7 and 8. When predicting either altruistic giving or warm glow giving, the most important variable in each case was the hypothetical experimental altruistic (*A38*) or warm glow (*WG37*) scenario. Though their scaled importances are slightly different, the three most important variables for predicting whether or not a participant donated in *A14_Donated* or *WG9_Donated* are the same as the most important variables for predicting donation amount in *A14* and *WG9*.

Using the same approach as before, we identified the best 3-variable subset for each model type and compared them to the model using all 62 variables. The results of this analysis are shown in Tables 10 and 11, and they show a similar trend to the results in Tables 6 and 7. Decision Trees and Random Forests outperform Logistic Regression when using the full model, although all three models are much closer for the 3-variable-subset version. Given the consistency of the Decision Tree model (i.e., its sensitivity and specificity are similar across all trials) and its ease of use, we

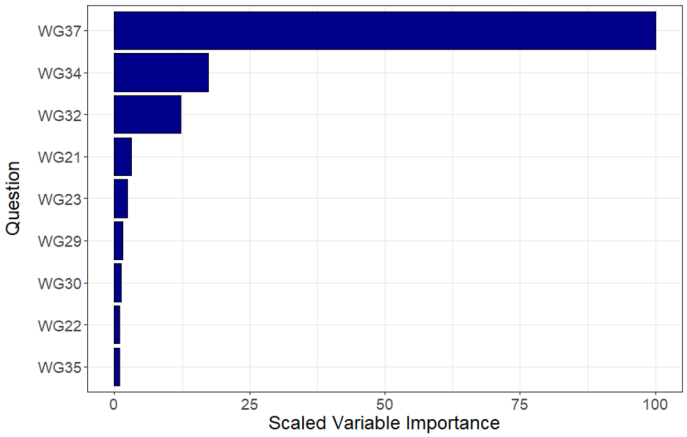

Fig. 8 Variable importance plot for the random forest modeling *WG9_Donated*

Table 10 Metrics for modeling *A14_Donated*

# of Variables	Model	Accuracy	Kappa	Sensitivity	Specificity
3	Logistic regression	0.803	0.473	0.906	0.534
3	Decision tree	0.812	0.533	0.870	0.662
3	Random forest	0.783	0.462	0.848	0.616
62	Logistic regression	0.719	0.282	0.821	0.453
62	Decision tree	0.819	0.555	0.866	0.697
62	Random forest	0.793	0.441	0.906	0.500

Table 11 Metrics for modeling *WG9_Donated*

# of Variables	Model	Accuracy	Kappa	Sensitivity	Specificity
3	Logistic regression	0.777	0.556	0.728	0.842
3	Decision tree	0.829	0.652	0.841	0.812
3	Random forest	0.796	0.586	0.813	0.774
62	Logistic regression	0.664	0.315	0.706	0.609
62	Decision tree	0.816	0.628	0.807	0.827
62	Random forest	0.800	0.593	0.813	0.782

next create rubrics for predicting *A14_Donated* and *WG9_Donated* using our best 3-variable Decision Tree.

Figures 9 and 10 show the rubrics for the 3-question survey predicting *A14_Donated* and *WG9_Donated*, respectively. The decision tree can be interpreted in a similar manner as the regression tree, with nodes predicting a binary 'Yes' or 'No' instead of a donation amount. There are four values in each node: the

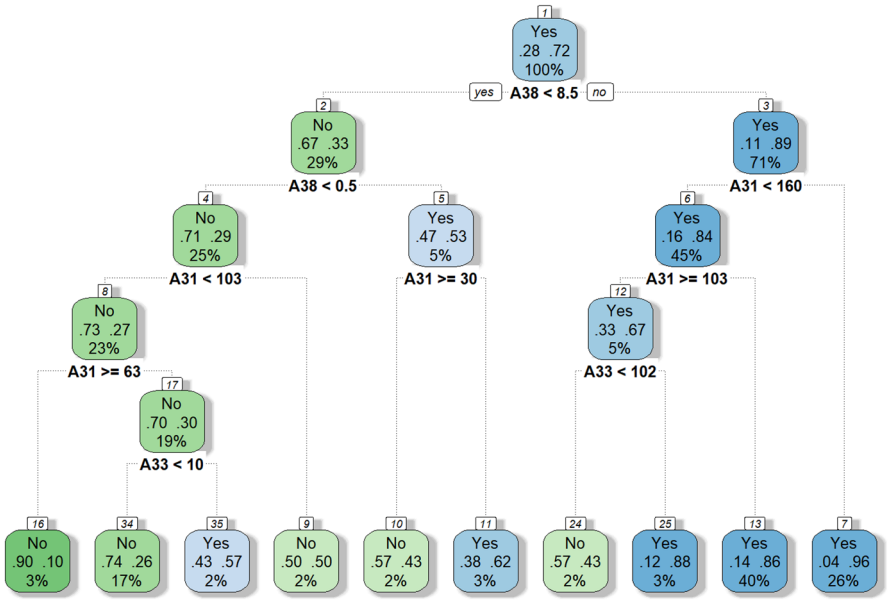

Fig. 9 Rubric for administering 3-question survey for *A14_Donated*

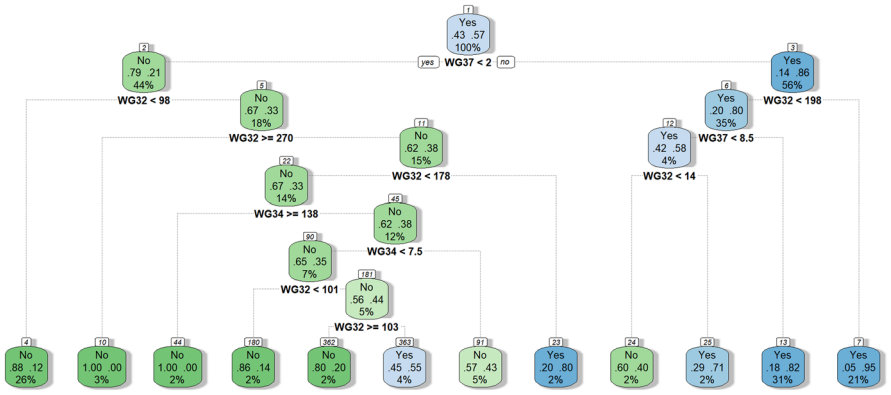

Fig. 10 Rubric for administering 3-question survey for *WG9_Donated*

top value represents the donation prediction (yes or no) for respondents in the given node, the left value represents the proportion of respondents in the given node who did not donate, the right value represents the proportion of respondents in the given node who did donate, and the bottom value represents the total proportion of respondents belonging to that node. Like before, each rubric highlights the strong relationship between donating large amounts in the hypothetical survey questions and donating some amount of money in both the altruistic and warm glow giving scenarios. The middle of each rubric again demonstrates how

to make predictions for participants who mix and match high and low donations across the three survey items.

5 Discussion

We conduct an incentivized experiment and survey with a nationally representative sample from the U.S. to construct behaviorally validated questionnaire modules that measure general altruism and warm glow. Using machine learning tools to consider all the possible combinations of survey predictors, we find that just three questions in each domain can account for as much variation in incentivized donations as using all our survey questions, predicting at both the extensive and intensive margins.

Considering altruism, our hypothetical lottery question (*A33*) correlates with donations in the standard dictator game almost exactly as well as in Falk et al. (2016). We consider this a replication of their work. However, we find that another similar question framing the endowment of money as earned instead of won does almost as well (*A31*). In the end, however, what trumps both of these questions is a hypothetical version of the experiment itself (*A38*). That said, the full 3-question module does best at capturing in intricacies of altruistic preferences.

A similar set of three questions framed in the context of warm glow giving (i.e., with potential crowding) do just as well in predicting impure altruism. Here, the lottery and bonus questions (*WG32* and *WG34*) correlate highly (especially on the intensive margin) and, together with the hypothetical version of the warm glow experiment (*WG37*), combine to form a highly predictive warm glow survey module.

As a final robustness check, we examined the "cross-correlations" to make sure that the questions designed to measure altruism did not predict warm glow giving as well as the warm glow questions and vice versa. Indeed they do not. In Supplementary Appendix Tables 12 and 13, we find that the "cross-correlations" are generally smaller, indicating two separate constructs.

The fact that hypothetical versions of the experiment (and other similar hypothetical donation decisions) are highly correlated with actual giving is encouraging and not too surprising given the seminal review of Camerer and Hogarth (1999) who first established similar results across a variety of games. Lastly, though the modules are based on hypothetical questions, linking these questions so strongly to behavior in the incentivized experiment allows us to be sure the responses represent what would happen in the much more costly experiment, were it to be conducted. In this sense, we have attenuated the traditional skepticism of hypothetical survey responses while making it essentially costless to collect validated measures of (pure and impurely) altruistic preferences.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s40881-024-00161-x.

References

Abrahams, B., & Schmitz, M. (1984). The crowding out effect of government transfers on private charitable contributions: Cross sectional evidence. *National Tax Journal, 37*, 536–38.

Andreoni, J. (2015). Warm glow and donor-advised funds: Insights from behavioral economics.

Andreoni, J. (1990). Impure altruism and donations to public goods: A theory of warm-glow giving. *The Economic Journal, 100*(401), 464–477.

Bekkers, R. (2007). Measuring altruistic behavior in surveys: The all-or-nothing dictator game. In *Survey research methods* (Vol. 1, No. 3, pp. 139–144).

Camerer, C. F., & Hogarth, R. M. (1999). The effects of financial incentives in experiments: A review and capital-labor-production framework. *Journal of Risk and Uncertainty, 19*, 7–42.

Carpenter, J. (2021). The shape of warm glow: Field experimental evidence from a fundraiser. *Journal of Economic Behavior & Organization, 191*, 555–574.

Carpenter, J., Connolly, C., & Myers, C. K. (2008). Altruistic behavior in a representative dictator experiment. *Experimental Economics, 11*(3), 282–298.

Clotfelter, C. T. (1985). *Federal tax policy and charitable giving*. University of Chicago Press.

Crumpler, H., & Grossman, P. J. (2008). An experimental test of warm glow giving. *Journal of Public Economics, 92*(5–6), 1011–1021.

D'Ambrosio, A., & Tutore, V. A. (2011). Conditional classification trees by weighting the Gini impurity measure. In *New perspectives in statistical modeling and data analysis* (pp. 273–280). Springer.

Dzeroski, S., & Zenko, B. (2004). Is combining classifiers better than selecting the best one. *Machine Learning, 2004*, 255–273.

Eckel, C. C., & Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior, 16*(2), 181–191.

Falk, A., Becker, A., Dohmen, T. J., Huffman, D., & Sunde, U. (2016). The preference survey module: A validated instrument for measuring risk, time, and social preferences.

Hastie, T., Tibshirani, R., & Friedman, J. H. (2001). *The elements of statistical learning: Data mining, inference, and prediction*. Springer.

Koschate-Fischer, N., Stefan, I. V., & Hoyer, W. D. (2012). Willingness to pay for cause-related marketing: The impact of donation amount and moderating effects. *Journal of Marketing Research, 49*(6), 910–927.

Lilley, A., & Slonim, R. (2014). The price of warm glow. *Journal of Public Economics, 114*, 58–74.

Luccasen, A., & Grossman, P. J. (2017). Warm-glow giving: Earned money and the option to take. *Economic Inquiry, 55*(2), 996–1006.

McHugh, M. (2012). Interrater reliability: The kappa statistic. *Biochemia Medica, 22*(3), 276–282. https://doi.org/10.11613/bm.2012.031

Nunes, P. A., & Schokkaert, E. (2003). Identifying the warm glow effect in contingent valuation. *Journal of Environmental Economics and Management, 45*(2), 231–245.

Payne, A. A. (1998). Does the government crowd-out private donations? New evidence from a sample of non-profit firms. *Journal of Public Economics, 69*(3), 323–345.

Ribar, D. C., & Wilhelm, M. O. (2002). Altruistic and joy-of-giving motivations in charitable behavior. *Journal of Political Economy, 110*(2), 425–457.

Rushton, J. P., Chrisjohn, R. D., & Fekken, G. C. (1981). The altruistic personality and the self-report altruism scale. *Personality and Individual Differences, 2*(4), 293–302.