*European Journal of Personality*, *Eur. J. Pers.* (2020)
Published online in Wiley Online Library (wileyonlinelibrary.com) **DOI**: 10.1002/per.2240

# Searching for Prosociality in Qualitative Data: Comparing Manual, Closed-Vocabulary, and Open-Vocabulary Methods

WILLIAM H.B. MCAULIFFE[1,2]*, HANNAH MOSHONTZ[3], THOMAS G. MCCAULEY[1,4] and MICHAEL E. MCCULLOUGH[1,4]

[1]*Department of Psychology, University of Miami, Miami, FL USA*
[2]*Department of Health Care Policy, Harvard Medical School, Harvard University, Boston, MA USA*
[3]*Department of Psychology and Neuroscience, Duke University, Durham, NC USA*
[4]*Department of Psychology, University of California, San Diego, La Jolla, CA USA*

*Abstract: Although most people present themselves as possessing prosocial traits, people differ in the extent to which they actually act prosocially in everyday life. Qualitative data that were not ostensibly collected to measure prosociality might contain information about prosocial dispositions that is not distorted by self-presentation concerns. This paper seeks to characterise charitable donors from qualitative data. We compared a manual approach of extracting predictors from participants' self-described personal strivings to two automated approaches: A summation of words predefined as prosocial and a support vector machine classifier. Although variables extracted by the support vector machine predicted donation behaviour well in the training sample (N = 984), virtually, no variables from any method significantly predicted donations in a holdout sample (N = 496). Raters' attempts to predict donations to charity based on reading participants' personal strivings were also unsuccessful. However, raters' predictions were associated with past charitable involvement. In sum, predictors derived from personal strivings did not robustly explain variation in charitable behaviour, but personal strivings may nevertheless contain some information about trait prosociality. The sparseness of personal strivings data, rather than the irrelevance of open-ended text or individual differences in goal pursuit, likely explains their limited value in predicting prosocial behaviour. © 2020 European Association of Personality Psychology*

Key words: prosocial behaviour; personal strivings; text analysis; social desirability bias

## INTRODUCTION

Although many personality traits are best measured through self-report, people do not always provide accurate reports of their motives and behaviour (Sun & Vazire, 2019; Vazire, 2010). Prosocial traits like generosity, honesty, and fairness are particularly difficult to measure validly because people present themselves favourably (Paulhus & John, 1998). Favourable self-presentation of prosocial traits is pervasive because moral character plays a central role in social judgement, guiding decisions about cooperation and exclusion (Baumard, André, & Sperber, 2013; Goodwin, Piazza, & Rozin, 2014). Thus, even people who are not generous, honest, or fair have incentives to present as if they are.

*Correspondence to: William H. B. McAuliffe, Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA, USA 02115-5899.
E-mail: williamhbmcauliffe@gmail.com
William H. B. McAuliffe and Hannah Moshontz contributed equally to this work.

This article earned Open Date, Open Materials and Preregistered + Analysis Plan badges through Open Practices Disclosure from the Center for Open Science: https://osf.io/tvyxz/wiki. The data and materials are permanently and openly accessible at https://osf.io/yhxgf/, https://osf.io/f3ckm/ and https://osf.io/b254j. Author's disclosure form may also be found at the Supporting Information in the online version.

Relative to self-report measures, measures of actual prosocial behaviour can be more diagnostic of prosocial traits. Unlike declarations of prosociality, prosocial behaviour typically involves real costs (e.g. time, resources, safety, etc.). However, incentives to present as prosocial are not only present when they complete self-report questionnaires, but also when people behave in any public domain (Barclay & Willer, 2006). When the costs of behaving prosocially are low or the potential reputational benefits to helping are high, observed prosocial behaviour is perhaps just as likely as self-reported prosociality to be motivated by self-presentation concerns (Batson & Shaw, 1991).

Qualitative data—in particular open-ended text generated by participants in structured interviews, essays, sentence completion tasks, etc.—may have unique advantages in measuring trait prosociality. Open-ended tasks are less transparent than direct self-report in what they are intending to measure, which may reduce socially desirable responding because participants do not know if or how they are being evaluated. For example, when directly asked about the importance of prosocial goals, most people rate prosocial goals as important to them, but when asked to list the goals that are important to them, far fewer people spontaneously mention prosocial goals (Frimer, Schaefer, & Oakes, 2014). Furthermore, responses to open-ended questions can reveal

*Handling editor: John Rauthmann*
*Received 3 June 2019*
*Revised 20 December 2019, Accepted 20 January 2020*

new insights because they are unconstrained by researchers' preconceptions about which predictors are most important to measure.

This paper focuses on the use of qualitative data to understand individual differences in prosociality. To maximise the relevance of our inquiry to personality researchers, we use measures and analysis techniques that researchers with limited resources or familiarity with machine learning techniques could implement. The success of this approach can speak to whether personality researchers need to change how they conduct research in order to reap the advantages of open-ended text measures.

## Manual text analysis

Qualitative data have long been used to characterise extremely prosocial people and generate new insights about the nature of prosociality. *Exemplar studies* frequently use open-ended prompts and structured interviews to understand 'moral exemplars'—people who have engaged in extreme prosocial acts that are likely to be non-instrumental, such as anonymously donating a kidney to a stranger (Marsh et al., 2014). Quantitative predictors derived from participants' natural language can outperform self-report measures in characterising exemplars. Walker and Frimer (2007) found that 50 people who had won an award for helping others (e.g. risking their lives to save others) were no different in self-reported personality traits from 50 demographically matched controls. In contrast, thematic coding of qualitative data was able to distinguish moral exemplars from matched controls (Frimer, Walker, Dunlop, Lee, & Riches, 2011). Qualitative data have also facilitated theory development. Close readings of text generated by exemplars have uncovered the role of redemption narratives in generativity (Mansfield & McAdams, 1996) and the role of idealistic peers in encouraging commitment to moral causes (Colby & Damon, 1992). These insights may have eluded researchers using only theory-constrained measures.

Although suggestive, the results of exemplar studies do not conclusively show that qualitative data contain more valid information about trait prosociality than self-report questionnaires. One reason is that exemplar studies do not sample people who show typical levels (i.e. close to the population mean) of prosociality. Instead, they focus on extreme manifestations of prosociality, which may introduce selection bias (Preacher, Rucker, MacCallum, & Nicewander, 2005). In addition, exemplar studies typically have modest sample sizes (e.g. the previously described 50 exemplars in Walker & Frimer, 2007). These modest sample sizes result from practical constraints: Moral exemplars are uncommon, and recruiting them is time-consuming and often expensive. Even if difficulties in recruiting moral exemplars could be overcome, manually coding open-ended audio and text requires extensive labour, and is therefore infeasible in large sample studies. Although understandable, the small samples of exemplar studies limit their ability to accurately estimate effect sizes. Simulations in contexts typical of personality research suggest that more than 200 participants are needed to obtain accurate and stable estimates of effect sizes (Schönbrodt & Perugini, 2013). Yet, to our knowledge, only one exemplar study had more than 200 participants (Oliner & Oliner, 1988). Given the absence of evidence from large, non-exemplar samples, it is unclear whether qualitative data can be used to measure prosociality with high validity.

## The rise of automated text analysis

Deriving nomothetic measures from text is resource intensive, but recent years have seen an increase in automated methods for capturing personality differences. Such methods can be broadly grouped into 'closed-vocabulary' and 'open-vocabulary' techniques. Closed-vocabulary techniques code text data according to the presence or absence of a pre-existing set of words. Like manual coding, closed-vocabulary approaches require the researcher to choose which words to code, based on either conceptual correspondence with the outcome of interest or based on a theory about the process that generates the outcome of interest. For example, Rand and Epstein (2014) used a dictionary of inhibition-related words (e.g. 'constrain', 'stop') to characterise interviews in which Carnegie Hero Medal recipients described the thoughts they had when they risked their own lives to rescue others. The authors found that the recipients' thoughts were more similar to control descriptions of intuitive thoughts than to control descriptions of deliberative thoughts. Rand and Epstein (2014) concluded that costly helping acts are based on intuitive rather than deliberate decision-making processes.

In contrast, open-vocabulary techniques *identify* words and other aspects of language (e.g. phrases, letters, or punctuation) that are associated with a focal outcome. Open-vocabulary approaches sometimes overfit data, leveraging noise in a sample to optimise prediction in a way that will not generalise to new data. To avoid overfitting, open-vocabulary techniques are often paired with cross-validation techniques. For instance, Park et al. (2015) used an open-vocabulary approach, ridge regression, to find the words and topics in a 'training' sample of about 66 000 people's Facebook statuses that predicted scores on questionnaire measures of Big Five personality traits. To estimate how well their model would predict new data, they used its parameter estimates to predict questionnaire measures in a 'holdout' sample of 4824 people who were not included in the training sample. The model was successful in predicting Big Five traits on both the training and holdout data sets, suggesting that social media behaviours can be used as markers of stable individual differences. Furthermore, examination of the word features also generated novel insights into the nature of Big Five traits. For example, Facebook statuses that lamented rude and selfish behaviour were associated with agreeableness—the Big Five trait most related to prosocial behaviour (Graziano & Eisenberg, 1997). Although other qualitative investigations have suggested that moral exemplars have strong convictions (Damon & Colby, 2015), researchers do not typically mention condemnation as a prototypical behaviour of agreeable people.

Although both closed-vocabulary and open-vocabulary methods of coding qualitative data may be more efficient than manual methods, they may not work as well in all circumstances. Open-vocabulary approaches sometimes require



**DOI**: 10.1002/per

larger samples than a single research group can afford to collect (Yarkoni & Westfall, 2017). This is particularly true when words appear infrequently in text because sparseness makes it difficult to detect reliable associations between individual word features and outcomes. Second, many 'out-of-the-box' automated methods may not be sophisticated enough to replace manual coding. The meaning of words is context dependent, and human raters are sensitive to the overall context of open-ended text data. Although there are methods of machine learning (e.g. neural networks) that train models on large corpora of natural language and can accommodate such subtleties, the most accessible methods divorce words from the grammatical structure that determines their meaning. For instance, a dictionary of prosocial words likely would not include the word 'money'. However, a human rater would recognize when money is being used to benefit others (e.g. 'I am trying to make more money to send my daughter to college'). Automated methods developed in the context of 'big data' analysis have the potential to replace or supplement manual methods of coding qualitative data, but their utility in personality research as it is normally conducted is unknown.

### The present study

The present study aims to use qualitative data to capture individual differences in prosociality. We compare manual and automated methods of deriving quantitative predictors from people's qualitative descriptions of goals they routinely pursue. In particular, we attempted to extract information from participants' personal strivings that would predict both whether and how much of their study compensation they donated to charity. We extracted quantitative predictors using three different coding methods: (i) human raters trained to code for themes reflecting universal values using a validated manual (the 'manual' approach), (ii) a dictionary of prosocial words (the closed-vocabulary approach), and (iii) a support vector machine (SVM) classifier (the open-vocabulary approach). We also evaluated how much variability in charitable behaviour each coding method accounted for on its own in the context of logistic regression (whether or not the participant donated) and linear regression (the proportion of bonus payment that the participant donated). To quantify each method's success in relative terms, we also observed how well past charitable involvement predicted donation decisions. Finally, we used a holdout sample to assess whether significant effects that we observe emerge in new data.

## METHODS

### Recruitment

We used data from two studies that were conducted on Amazon.com's Mechanical Turk (Study 1: $N$ = 814, 410 women, $M_{age}$ = 36.6, $SD_{age}$ = 11.19; Study 2: $N$ = 778, 350 women, $M_{age}$ = 35.75, $SD_{age}$ = 11.45). All participants in both studies were from the USA. Participants in Study 1 were offered $1.00 to begin the study and $2.00 bonus payment for completing the study. Participants from Study 1 were not allowed to enrol in Study 2. Participants in Study 2 were offered $1.00 to begin the study and a $4.00 bonus payment for completing the study.

### Procedure

The full procedure of both studies were described in detail elsewhere (McAuliffe, 2019; McCauley & McCullough, 2019); here, we describe only the details that are relevant to the present analysis (see these study materials here: https://osf.io/f3ckm/). After providing informed consent, both studies began with the Personal Strivings List (Emmons, 1999). To complete the list, participants completed the sentence 'I typically try to …' with their personal strivings—that is, the goals that they are trying to achieve in their everyday lives. Strivings are idiographic in the sense that they are personalised (e.g. 'I am typically trying to gain back my spouse's trust') but are nomothetic in the sense that they can be classified into a type of motive (e.g. a desire to sustain affiliations). Strivings are thought to represent individual differences in motivation better than other personality constructs (Dunlop, 2015; McAdams, 1995). Studies suggest that nomothetic variables derived from strivings have convergent validity (Hart, McAdams, Hirsch, & Bauer, 2001) and predict prosocial behaviour in the laboratory (Magee & Langner, 2008) and in everyday life (Frimer et al., 2011).

Participants reported the 10 strivings that best characterise their prototypical motives. We used an abbreviated version of the instructions provided by Emmons (1999) that included several example strivings and encouraged participants to consider all of the goals that they are pursuing in the myriad domains of their lives. Two research assistants corrected typos and misspellings before we conducted the automated analyses. The strivings were relatively short for open-ended text (mean words in combined strivings per person = 48.25, $SD$ = 20.00). Word sparseness can make it difficult for automated methods to identify relationships between word features and behavioural outcomes (Banks, Woznyj, Wesslen, & Ross, 2018). We took steps to mitigate sparseness at each stage of our procedure.

Subsequently, participants in both studies completed several individual difference measures. These measures were beyond the scope of the present study save for a subset of items from the Self-Report Altruism Scale related to charity (Rushton,Chrisjohn, & Fekken, 1981) that was administered in Study 1. Participants completed the self-report altruism scale by reporting the number of times they have performed 29 (20 from the original scale, and 9 pilot items developed by us) different helpful acts (1 = *never*, 2 = *once*, 3 = *more than once*, 4 = *often*, 5 = *very often*). To derive a measure of past charitable involvement (omega = .79; $M$ = 2.55; $SD$ = 0.77), we averaged three items from the original scale ('I have given money to a charity', 'I have donated goods or clothes to a charity', and 'I have done volunteer work for a charity') with three additional items about charitable involvement that we created ('I have participated in a charity fundraising event (e.g. a 5k run)', 'I have performed an administrative role for a charitable cause', and 'I have made a "pledge" to make a regular contribution to a charitable cause').



**DOI**: 10.1002/per

Finally, participants were told that they would watch a video made by a charity about a social problem. In Study 1, participants watched a video made by Oxfam, documenting a family of refugees fleeing civil war in South Sudan. In Study 2, participants watched a video we created documenting recent hurricane victims, and that United Nations Children's Fund was aiming to help them. Then, participants used a sliding scale to indicate how much of their bonus payment ($2.00 in Study 1, $4.00 in Study 2) that they would like to donate (in $0.05 increments in Study 1 and $0.10 increments in Study 2) to charity (Oxfam in Study 1: *M* = $0.55, *SD* = $0.66; United Nations Children's Fund in Study 2: *M* = $0.86, *SD* = $1.11). After making a donation decision, Study 1 ended. In Study 2, after making a donation decision, participants observed and reported the outcome of a coin flip that could nullify their donation decision. Analysis of the coin flip task will be reported in another paper (McCauley & McCullough, 2019).

Charitable giving tasks are similar to economic game measures that are increasingly used by personality researchers to measure prosocial behaviour (Zhao & Smillie, 2015). Whether people give money in these tasks is associated with dispositional factors (Thielmann, Spadaro, & Balliet, 2020). Millions of people donate to charity each year (List, 2011), making charitable donations a plausible outcome of the goals people pursue in daily life.

Table 1. Universal value means, standard deviations, and interrater reliabilities in Study 1 and Study 2

| Universal value | Study 1 | | | Study 2 | | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | K | *M* | *SD* | K |
| Benevolence | 2.38 | 1.61 | .82 | 2.57 | 1.68 | .82 |
| Universalism | 0.23 | 0.56 | .63 | 0.41 | 0.69 | .66 |
| Conformity | 0.84 | 0.96 | .65 | 0.89 | 1.01 | .70 |
| Tradition | 0.44 | 0.73 | .56 | 0.59 | 0.83 | .68 |
| Security | 2.25 | 1.74 | .78 | 2.23 | 1.74 | .83 |
| Hedonism | 0.84 | 1.00 | .61 | 0.77 | 1.00 | .82 |
| Self-direction | 2.22 | 2.19 | .87 | 2.18 | 2.12 | .86 |
| Stimulation | 0.63 | 0.87 | .40 | 0.61 | 0.83 | .65 |
| Achievement | 2.16 | 2.00 | .49 | 2.54 | 2.07 | .84 |
| Power | 1.16 | 1.17 | .70 | 0.88 | 1.13 | .79 |

*Note:* *M* and *SD* are used to represent mean and standard deviation, respectively. K is used to represent Kappa. In Study 1, $N = 778$. In Study 2, $N = 726$.

## RESULTS

### Deriving predictors from qualitative text

*Manual approach*

In the manual approach, we had research assistants code personal strivings using the Values Embedded in Narrative (VEiNs) manual (Frimer, Walker, & Dunlop, 2009; see manual at http://www.jeremyfrimer.com/research-downloads.html) that describes how to code for 10 values that Schwartz (1994) identified as pancultural (Table 1). These values fall into two higher order dimensions: self-transcendence (benevolence and universalism) versus self-enhancement (power, achievement, and hedonism) and openness to change (self-direction, stimulation, and hedonism) versus conservation (tradition and conformity). Frimer et al. (2011) found that moral exemplars reported not only more self-transcendence strivings than matched controls, but also more self-enhancement strivings.

Each striving was coded by two raters who had no information about participants (including their donation status). Raters went through a thorough training process: For each value, they completed half of the manual's practice set of 400 strivings, referred to the 'correct' answers and discussed confusions with the first author, and then completed the remaining 200 practice strivings. Raters coded no more than two values at a time. The first author calculated interrater reliabilities (Table 1) and resolved coding disagreements. Each participant received a sum score for each of the 10 values (Table 1).

The VEiNs manual recommends interrater reliability of $\kappa \geq .60$, which is considered 'substantial agreement' according to conventional benchmarks (Landis & Koch, 1977). We

Table 2. Words with largest average weight across 10-fold cross-validation of support vector machine classifier

| Word | *M* weight | Weight rank | | | Runs ranked in top 10 | % of donors | % of non-donors |
|---|---|---|---|---|---|---|---|
| | | *M* | Min | Max | | | |
| Easy | 0.4496 | 3.6 | 1 | 19 | 9 | 1.19 | 0.00 |
| God | 0.4150 | 6.2 | 1 | 27 | 8 | 1.71 | 1.51 |
| Shows | −0.3736 | 11.5 | 2 | 55 | 8 | 0.17 | 1.00 |
| Timely | 0.3304 | 15.9 | 4 | 32 | 4 | 2.05 | 0.25 |
| Career | 0.3208 | 18.8 | 5 | 31 | 3 | 3.75 | 0.25 |
| Mind | −0.3438 | 21 | 3 | 90 | 6 | 2.39 | 4.27 |
| Sleep | −0.3164 | 22.4 | 5 | 39 | 2 | 9.04 | 11.31 |
| Sister | −0.3193 | 26.4 | 1 | 94 | 2 | 0.34 | 2.01 |
| Causing | −0.3058 | 30.4 | 4 | 88 | 2 | 0.00 | 0.75 |
| Thrifty | −0.3233 | 31.3 | 3 | 170 | 2 | 0.00 | 0.75 |
| Safely | −0.2946 | 31.9 | 9 | 75 | 1 | 1.00 | 1.19 |

*Note:* Weights provide information about importance of the word relative to all others in a particular model. Mean weight is the average weight across 10 runs of 10-fold CV. % of Donors is the proportion of participants who donated at least $0.01 who used the word at least once in their personal strivings. % of non-donors is the proportion of participants who donated no money and used the word at least once in their personal strivings. *N* in the training set was 984.



**DOI**: 10.1002/per

investigated why agreement fell below this threshold for tradition, achievement, and stimulation in Study 1. For achievement and stimulation, the first author's corrections showed only 'fair' agreement with this rater, but 'almost perfect' agreement with the other rater. For tradition, raters disagreed on how to code a few common strivings, perhaps because they were rare in the practice strivings set. Thus, the first author acted as a third rater for tradition and, at times, overruled agreement among both raters. Interrater reliability was more consistent for Study 2, with all kappas greater than .6.

*Closed-vocabulary approach*

We used a dictionary of 127 word stems that have been used to characterise the prevalence of prosocial language in study settings (Frimer et al., 2014) and transcripts of public proceedings (Frimer, Aquino, Gebauer, Zhu, & Oakes, 2015). Because many applied researchers are familiar with the Linguistic Inquiry and Word Count (LIWC) programme (Pennebaker, Booth, & Francis, 2007), we used TACIT's (Dehghani et al., 2017) LIWC-style word count plugin to count how many times participants used each of the word stems in their strivings. TACIT is a free, online tool that was designed to be user-friendly for psychologists who have only a rudimentary background in conducting text analysis.

Prosocial dictionary word counts were generated by creating a dictionary file with words from Frimer et al. (2015). Example words included charit*, good, and volunteer (see http://www.jeremyfrimer.com/research-downloads.html for a complete list). Individual text files of all 10 strivings were created for each participant. Using the uploaded dictionary file, TACIT created a file for every participant who used any word in the prosocial dictionary in any of their strivings. These files were used to count the number of times each of the 127 words appeared in each participant's combined strivings.

*Open-vocabulary approach*

There were two phases to our open-vocabulary approach. First, we used TACIT's SVM classifier to identify the linguistic features most predictive of whether participants donated. Like other machine-learning approaches, SVM optimises classification accuracy. Optimisation on one data set can cause overfitting, resulting in a set of features that does not generalise to new samples. To identify features that were more likely to generalise to new samples, we used 10-fold cross validation, which means that TACIT iterated the model training process 10 times, with a given 10% subset of the data serving as the validation set once and comprising part of the training sample in the nine other iterations. Then, TACIT selects the best model based on its performance in each validation set, producing 10 models. This allowed us to eliminate features that predict donations only because of sampling error in the held-in training set. Furthermore, we used a separate holdout sample to assess the performance of all features, including those derived from SVM. That is, before we conducted the SVM, we randomly selected 33% of observations in our sample (stratified on donation status) to serve as a validation set. These observations were not included in the SVM analysis and allowed us to estimate how well models using these features would perform in new samples. The SVM approach was not compellingly useful; across the 10 runs, model accuracy ranged from 50.5% to 63.6% (average = 53.7%), and nine of the 10 $p$-values were larger than 0.05.

In the second phase of our open-vocabulary approach, we used the results of the SVM analyses (i.e. the weights associated with extracted feature in each SVM) to select the final set of SVM-derived features. We ranked the absolute value of the weight of each feature in each run. We selected features that (i) had consistent weight direction and class association across runs, (ii) were in the top 127 of at least three runs, and (iii) were words (e.g. we dropped 'M'). Then, we used TACIT's LIWC-style dictionary tool to count the frequency of each word in participants' combined strivings. We counted the presence of SVM-identified words in the entire data set (but, again, the SVM was run using only the training set). The procedure we used to select SVM features was similar, but not identical to, the procedure we described in our preregistration. The preregistered procedure resulted in features (e.g. M) that were not meaningful words or that appeared only once in the entire data set.

## Analytic approach

Our preregistration explains both our prior knowledge of the data sets, our planned analyses for the current study, and our hypotheses (https://osf.io/b254j).

*Data exclusions*

We excluded data from participants who did not make a decision about whether to donate or did not complete all 10 strivings. Participants who clearly did not take the strivings task seriously (e.g. wrote 'blablabla' for all 10 strivings) were also excluded. These exclusions left 1480 usable cases. This final $N$ is one case smaller than the number of cases we expected to be able to use in our preregistration; after the preregistration, we identified an additional case in which the participant did not take the strivings task seriously. In a deviation from our preregistration, we split the data set into a training set with two thirds ($n$ = 984) of usable cases and a holdout set of one third ($n$ = 496) instead of our originally planned seven tenths training–three tenths holdout split. We found it was easier to use a two thirds–one third split to create equal-sized folds that had the same percentage of donors versus non-donors. Power analyses using G*Power (Faul, Erdfelder, Buchner, & Lang, 2009) revealed that we have 80% power to explain as little as 1.64% of variance in the training sample and 3.23% of variance explained in the holdout sample (based on 10 predictors and an alpha of .05).

The training and holdout sets included data from both Study 1 and Study 2 in order to prevent optimising the results to idiosyncratic features of either study. Because Study 1 and Study 2 involved different amounts of money, we transformed donation amounts to proportions for linear regression analyses. Of course, putting the raw donation amounts from both studies on a common metric does not necessarily make donation decisions across studies directly comparable.



**DOI**: 10.1002/per

Indeed, participants from Study 2 donated a significantly smaller proportion of their bonus payment in the training set, $b = -.06$, se = .02, $t = -3.05$, $p = .002$, 95% confidence interval (CI) [−0.10, −0.02], and in the holdout set, $b = -.08$, se = .03, $t = -2.93$, $p = .003$, 95% CI [−0.13, −0.03]. Moreover, combining data across studies could introduce correlated residuals, causing us to underestimate standard errors. To correct for these biases, we included a study dummy variable (0 = Study 1, 1 = Study 2) as a covariate in all regression models that included data from both studies.

*Identifying the top 10 predictors*

Because our primary goal was to identify word features that may yield new insight into charitable donors, we focused on finding a small set of predictors that maximise prediction of prosocial behaviour. We settled on 10 predictors on the premise that researchers may want to know whether they should manually code for all ten universal values or instead code some other set of ten features. We created two 'top 10' predictor lists: One for the decision to make a donation at all and one for donation proportion.

Our candidate set of predictors came from each of our three methods. We considered all 10 universal values from our manual coding approach. From our closed-dictionary approach, all 127 word stems were candidates. For our open-vocabulary approach, the 127 word features that had the highest weights[1] across SVM runs were candidates. Variables from both the prosocial dictionary words and the SVM method were sparse (see supplement: https://osf.io/wajxq/).

We identified the top 10 predictors from our candidate set using stochastic search variable selection (SSVS). SSVS uses Gibbs sampling, a Markov chain Monte Carlo procedure that simulates the joint density distribution for an outcome and a set of predictors, to determine which predictors consistently predict the outcome, even after accounting for variability in the other predictors included in the model (George & McCulloch, 1993). Gibbs sampling is useful in the context of variable selection because, after observing many 'warm-up' samples, it primarily simulates probability distributions that have the highest posterior probability values, iteratively ranking the posteriors produced in each run of the simulation. Thus, as the number of runs increases, models with higher ranking posteriors are selected more frequently, and models that possess strong predictors of the outcome are more likely to be estimated.

[1]Many features were identified by SVM runs, but in our preregistered analysis plan, we limited ourselves to 127 features for parity with the prosocial word dictionary. After looking at these 127 words, we deviated from our preregistration to avoid overfitting by further restricting this variable set as follows: (i) we removed both SVM and prosocial dictionary words that appeared in fewer than 10 participants' data; (ii) we combined similar SVM words into one stem (e.g. 'listen' and 'listener'); (iii) we stemmed all SVM words (e.g. one of the SVM words was 'social' and we counted 'socially', 'socializing', and 'socialize'). After these steps, there were 102 predictors available for Stochastic Variable Search Selection (SSVS) analyses (with some overlap in the predictors contributed by SVM and the prosocial dictionary).

We used the R code available from Bainter, McCauley, Wager, and Losin (n.d., in press) to run the SSVS models. We based results on 20 000 regression models (the first 5000 were discarded warm-ups) that sample different combinations of the 102 predictors. All predictors had prior probabilities of .50. We chose the 10 variables that had the highest 'marginal inclusion probability'(MIP) (i.e. the proportion of times a predictor is included in the most predictive models) to represent the top 10 predictors.

Although we lacked predictions about which specific word features would be most strongly associated with donation decisions, we hypothesised that words from the SVM would be more likely to appear in the top 10 features that predict decisions to make a donation because it is designed to maximise prediction in the training set. We did not extend this hypothesis to the SSVS for predicting donation proportion because the SVM was trained only to predict whether participants would donate at all.

Our second hypothesis was that, if any manually coded universal values appeared in the top 10 models, they would be benevolence, universalism, power, or achievement. These four values have been found to distinguish extraordinarily prosocial individuals from matched controls (Frimer et al., 2011). In contrast, there is no evidence yet that VEiNs coding of the other six values are useful in predicting prosocial behaviour.

The SSVS does not estimate multicollinearity among predictive variables, and text analysis methods often produce multicollinear predictors (Firmin, Bonfils, Luther, Minor, & Salyers, 2017). To characterise multicollinearity among each of our top 10 predictors, we used dominance analysis (DA; Azen & Budescu, 2003) and its extension to logistic regression (Azen & Traxel, 2009). DA assesses the variance explained by all possible subsets of a set of predictors. For each predictor, it provides the average variance explained and the proportion of models in which it explains more variance than a second predictor. We used the DA package in R to perform analyses on both the training and holdout set (Bustos & Soares, 2019). Because multicollinearity was low and prediction was generally poor, the DA analyses were uninformative and we relegated them to supplemental materials: https://osf.io/379ku/.

## Preregistered analyses: Which word features best predict charitable donations?

The analyses reported below were all conducted in R version 3.6 (R Core Team, 2019). The syntax can be found here: https://osf.io/5yac4/; the training data set can be found here: https://osf.io/x3ems/; and the holdout data set can be found here: https://osf.io/8xpqh/. All tests were two-tailed with an alpha of .05.

### Stochastic search variable selection

The SSVS on the training set yielded discouraging results for both decisions to make a donation and donation proportion (see Figures 1 and 2 for full results). For the donation decision (i.e. binary logistic) SSVS, only 'career', had an MIP greater than .50. 'Gain*' had an MIP of .45, and the other eight features ('sister*', 'easy*', 'everyone', 'visit', 'me',



**DOI**: 10.1002/per

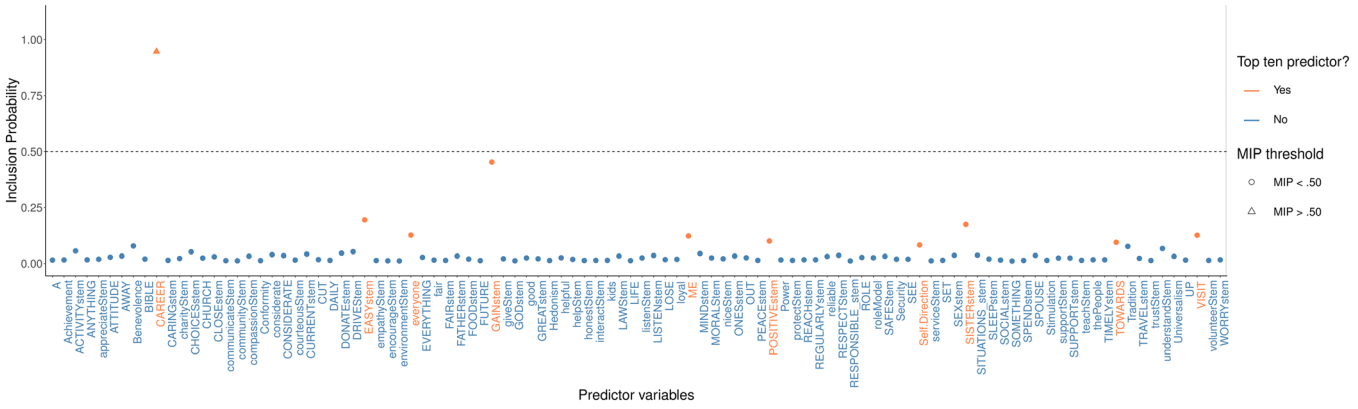


Figure 1. Results from the stochastic search variable selection model predicting decisions to make a non-zero donation to charity. MIP, marginal inclusion probability. [Colour figure can be viewed at wileyonlinelibrary.com]

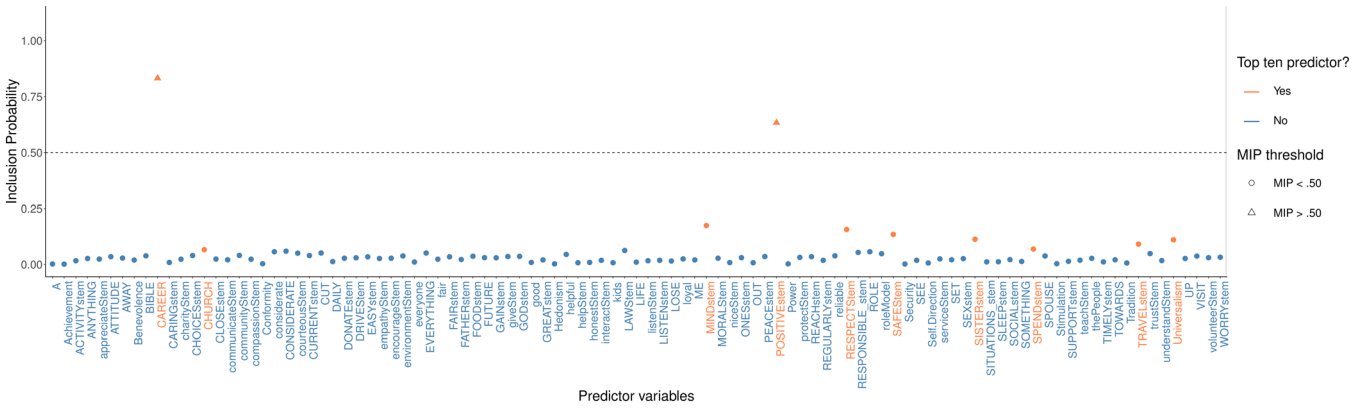


Figure 2. Results from the stochastic search variable selection model predicting proportion of bonus payment donated to charity. MIP, marginal inclusion probability. [Colour figure can be viewed at wileyonlinelibrary.com]

'positive*', 'towards', and self-direction values) all had MIPs lower than .20, indicating that consideration of the data *reduced* their probability of inclusion in models with high posterior probability. One feature was from the prosocial dictionary ('everyone'), one was from manual coding (self-direction values), and the other eight were from SVM.

For the donation proportion SSVS, 'career' and 'sister*' again appeared among the features with the 10 highest MIPs. Only two words, 'career' and 'positive*', had MIPs greater than .5. The other eight top 10 features, 'mind*', 'respect*', 'safe*', 'sister*', 'travel*', 'spend*', 'church', and universalism values, all had MIPs lower than .18. Besides universalism values, all top features came from SVM, although 'respect*' is both an SVM word and a prosocial dictionary word.

## Regression models featuring the top 10 predictors

We began by entering the study dummy and all top 10 predictors from the logistic SSVS as predictors of making a donation, using the training data set (Table 3). The model explained 4.55% of overall variance (based on McFadden's $R^2$) and significantly improved on an intercept-only model, $\chi^2$ (11) = 60.36, $p < .001$. 'Career,' 'gain*,' and 'easy*' had significant, positive effects, while 'sister' had a

Table 3. Logistic regression of "top 10" word features on donation decision

| | Data set | OR | *p* | 95% CI |
|---|---|---|---|---|
| Career | Training | 12.31 | .014 | [2.58, 221.21] |
| | Holdout | 1.48 | .522 | [0.47, 5.53] |
| Positive* | Training | 1.57 | .059 | [0.99, 2.55] |
| | Holdout | 0.65 | .187 | [0.34, 1.23] |
| Gain* | Training | 3.85 | .029 | [1.34, 16.32] |
| | Holdout | 1.14 | .823 | [0.34, 4.53] |
| Easy* | Training | 4.70 | .046 | [1.25, 30.72] |
| | Holdout | 1.70 | .595 | [0.30, 29.28] |
| Everyone | Training | 1.51 | .070 | [0.98, 2.40] |
| | Holdout | 1.29 | .301 | [0.81, 2.17] |
| Sister* | Training | 0.26 | .032 | [0.07, 0.83] |
| | Holdout | 0.63 | .654 | [0.07, 5.43] |
| Visit | Training | 3.37 | .106 | [0.96, 21.39] |
| | Holdout | 1.22 | .790 | [0.28, 6.25] |
| Me | Training | 1.32 | .066 | [0.99, 1.80] |
| | Holdout | 0.86 | .353 | [0.63, 1.18] |
| Towards | Training | 1.92 | .121 | [0.90, 4.83] |
| | Holdout | 1.40 | .527 | [0.51, 4.44] |
| Self-direction | Training | 1.05 | .109 | [0.99, 1.12] |
| | Holdout | 1.10 | .029 | [1.01, 1.21] |
| Study dummy | Training | 0.98 | .909 | [0.76, 1.28] |
| | Holdout | 0.85 | .397 | [0.59, 1.23] |

*Note:* * = word stem. Study dummy = the effect of participating in Study 2 rather than Study 1. OR = odds ratio for making a donation. Confidence intervals are exponentiated. All tests were two-tailed with an alpha of .05.





**DOI**: 10.1002/per

significant negative effect. The other words had non-significant effects. The variance inflation factors (VIFs) were all less than 1.04, suggesting that the estimation and interpretation of coefficients were not obscured by overlapping variance.

To determine whether the associations we observed would generalise to data that the predictors were not trained on, we then conducted the same logistic regression model on the holdout set (Table 3). The model explained 1.52% of variance and was *not* a significant improvement on an intercept-only model, $\chi^2(11) = 10.17$, $p = .516$. There was a significant, positive effect of self-direction; no other effects were significant. Multicollinearity was again too low to provide a ready explanation for non-significant effects (all VIFs $< 1.06$).

Next, we entered all top 10 predictors from the donation proportion SSVS and the study dummy variable as predictors of the proportion of bonus payment donated to charity, using the training data set (Table 4). The model was significant, $F(11, 972) = 5.87$, $p < .001$, adjusted $R^2 = .051$. Universalism was, contrary to our expectations, negatively associated with donations. Multicollinearity was again low (all VIFs $< 1.05$).

Next, we created a linear multiple regression model with the same 11 predictors and the same outcome, but this time on the holdout data set. The model was just barely significant, $F(11, 484) = 1.81$, $p = .050$, adjusted $R^2 = .018$, and 'respect*' was the only significant predictor other than the study dummy variable. Multicollinearity was again low (all VIFs $< 1.07$).

Table 4. Linear regression of "top 10" word features on donation proportion

| | Data set | *b* | *p* | 95% CI |
|---|---|---|---|---|
| Career | Training | 0.19 | .002 | [0.07, 0.31] |
| | Holdout | −0.06 | .467 | [−0.22, 0.10] |
| Positive* | Training | 0.11 | <.001 | [0.05, 0.17] |
| | Holdout | −0.07 | .144 | [−0.16, 0.02] |
| Mind* | Training | −0.10 | .020 | [−0.17, −0.01] |
| | Holdout | −0.02 | .628 | [−0.12, 0.07] |
| Respect* | Training | −0.13 | .047 | [−0.24, 0.01] |
| | Holdout | −0.12 | .045 | [−0.24, −0.00] |
| Safe* | Training | −0.14 | .032 | [−0.26, −0.02] |
| | Holdout | −0.08 | .520 | [−0.32, 0.16] |
| Sister* | Training | −0.16 | .053 | [−0.32, 0.00] |
| | Holdout | −0.08 | .581 | [−0.38, 0.21] |
| Universalism | Training | −0.03 | .020 | [−0.06, −0.00] |
| | Holdout | −0.03 | .197 | [−0.07, 0.02] |
| Travel* | Training | 0.11 | .088 | [−0.01, 0.24] |
| | Holdout | −0.13 | .242 | [−0.36, 0.09] |
| Spend* | Training | 0.05 | .015 | [0.01, 0.09] |
| | Holdout | −0.03 | .432 | [−0.09. 0.04] |
| Church | Training | −0.10 | .164 | [−0.23, 0.04] |
| | Holdout | −0.14 | .231 | [−0.36, 0.09] |
| Study dummy | Training | −0.05 | .009 | [−0.09, −0.01] |
| | Holdout | −0.08 | .006 | [−0.13, −0.02] |

*Note:* * = word stem. Study dummy = the effect of participating in Study 2 rather than Study 1. Coefficients and confidence intervals represent proportion of bonus payment transferred to charity. All tests were two-tailed with an alpha of .05.

## Exploratory analyses: Which coding method best predicts charitable giving?

We were unsure whether to interpret the null results from our preregistered analyses as an artefact of only using the top 10 word features as predictors. To the extent that SVM predictors 'won' in the SSVS simply because the SVM was designed to maximise prediction in the training set, we may have overlooked word features from other coding methods that have reliable associations with charitable giving. Moreover, it is arguably unreasonable to expect any one variable, especially when sparse, to evince a strong association with a behavioural outcome. Using many predictors, either aggregated together or used separately but simultaneously, is more likely to do justice to the myriad causes that come together to cause any one action (Seeboth & Mõttus, 2018).

We decided to conduct exploratory regressions that used all variables from a single method to predict donation behaviour. For these analyses, we used all potential variables, meaning that we did not exclude based on low frequencies (we did not stem previously excluded words, however). With the expectation that our new analyses may yield more positive results, we also examined aspects of construct validity beyond convergent validity. Bleidorn and Hopwood (2019) note that machine learning techniques often identify variables that have convergent validity but not necessarily other aspects of construct validity. For example, word features that predict charitable giving in the context of one study may be of limited use if they cannot predict charitable giving enacted in different types of situations (*generalisability*) and if they cannot predict charitable giving over and above convenience measures like self-report (*incremental validity*). Consequently, we assessed the generalisability and incremental validity of variables from each method by (i) examining the extent to which they predict past charitable involvement (generalisability) and (ii) observing whether they significantly predict charitable giving over and above self-reported past charitable involvement (incremental validity). Past charitable involvement is strongly correlated with verified donation behaviour (Bekkers & Wiepking, 2011), which in turn strongly predicts future donations (Prospect Research and Wealth Screening Statistics, 2015). In our sample, past charitable behaviour was correlated with actual donation behaviour (proportion of bonus payment donated) in both the training set ($r = .19$, 95% [.10, .27], $p < .001$) and the holdout set ($r = .16$, 95% CI [0.05, 0.28], $p = .007$). Because we only measured past charitable involvement in Study 1, models involving this variable had a smaller sample size and did not include the study dummy variable.

## Manual ratings of universal values

We created a logistic regression model that contained all 10 universal values and the study dummy variable as predictors of making a donation (Table 5). The model was not a significant improvement on an intercept-only model, $\chi^2(11) = 15.87$, $p = .146$. Self-direction was the only significant predictor, $b = .11$, se = .05, odds ratio (OR) = 1.12,




**DOI**: 10.1002/per

Table 5. Logistic regression of universal values on donation decision

| | Data set | OR | *p* | 95% CI |
|---|---|---|---|---|
| Benevolence | Training | 0.94 | .143 | [0.87, 1.02] |
| | Holdout | 0.96 | .476 | [0.84, 1.08] |
| Universalism | Training | 0.90 | .329 | [0.74, 1.11] |
| | Holdout | 0.93 | .639 | [0.68, 1.27] |
| Self-direction | Training | 1.12 | .020 | [1.02, 1.23] |
| | Holdout | 1.16 | .027 | [1.02, 1.32] |
| Stimulation | Training | 1.00 | .973 | [0.85, 1.18] |
| | Holdout | 0.83 | .121 | [0.66, 1.05] |
| Security | Training | 0.98 | .585 | [0.90, 1.06] |
| | Holdout | 1.04 | .506 | [0.93, 1.17] |
| Power | Training | 1.03 | .597 | [0.92, 1.16] |
| | Holdout | 1.05 | .530 | [0.90, 1.25] |
| Achievement | Training | 0.94 | .230 | [0.86, 1.04] |
| | Holdout | 0.94 | .410 | [0.82, 1.08] |
| Hedonism | Training | 0.97 | .643 | [0.85, 1.11] |
| | Holdout | 1.06 | .605 | [0.86, 1.31] |
| Tradition | Training | 0.86 | .095 | [0.73, 1.03] |
| | Holdout | 0.88 | .294 | [0.69, 1.12] |
| Conformity | Training | 1.04 | .614 | [0.90, 1.19] |
| | Holdout | 0.92 | .380 | [0.75, 1.11] |
| Study dummy | Training | 1.07 | .609 | [0.82, 1.41] |
| | Holdout | 0.95 | .797 | [0.65, 1.39] |

*Note:* OR = odds ratio for making a donation. Study dummy = the effect of participating in Study 2 rather than Study 1. Confidence intervals are exponentiated. All tests were two-tailed with an alpha of .05.

Table 6. Linear regression of universal values on donation proportion

| | Data set | *b* | *p* | 95% CI |
|---|---|---|---|---|
| Benevolence | Training | −0.01 | .207 | [−0.02, 0.00] |
| | Holdout | −0.00 | .901 | [−0.02, 0.02] |
| Universalism | Training | −0.03 | .063 | [−0.06, −0.00] |
| | Holdout | −0.02 | .279 | [−0.07, 0.02] |
| Self-direction | Training | 0.02 | .049 | [0.00, 0.03] |
| | Holdout | −0.00 | .970 | [−0.02, 0.02] |
| Stimulation | Training | 0.00 | .843 | [−0.02, 0.03] |
| | Holdout | −0.01 | .457 | [−0.05, 0.02] |
| Security | Training | 0.00 | .612 | [−0.01, 0.02] |
| | Holdout | 0.00 | .923 | [−0.02, 0.02] |
| Power | Training | 0.00 | .883 | [−0.02, 0.02] |
| | Holdout | 0.00 | .864 | [−0.02, 0.03] |
| Achievement | Training | −0.01 | .308 | [−0.02, 0.01] |
| | Holdout | 0.01 | .294 | [−0.01, 0.03] |
| Hedonism | Training | 0.00 | .344 | [−0.02, 0.02] |
| | Holdout | −0.01 | .703 | [−0.04, 0.02] |
| Tradition | Training | −0.01 | .530 | [−0.03, 0.02] |
| | Holdout | −0.01 | .596 | [−0.04, 0.03] |
| Conformity | Training | 0.00 | .658 | [−0.02, 0.03] |
| | Holdout | −0.02 | .304 | [−0.04, 0.01] |
| Study dummy | Training | −0.05 | .023 | [−0.09, −0.01] |
| | Holdout | −0.08 | .007 | [−0.13, −0.02] |

*Note:* Coefficients and confidence intervals represent proportion of bonus payment transferred to charity. Study dummy = the effect of participating in Study 2 rather than Study 1. All tests were two-tailed with an alpha of .05.

$Z = 2.32$, $p = .020$, 95% CI [0.02, 0.21]. Multicollinearity was higher in this model than in previous models, but still modest (all VIFS $< 2.42$). We conducted the same model on the holdout set and observed the same pattern: The model was not a significant improvement on an intercept-only model, $\chi^2(11) = 13.75$, $p = .247$, and self-direction was the only significant predictor, $b = .14$, se = .07, OR = 2.21, $Z = 2.27$, $p = .027$, 95% CI [0.02, 0.27].

Using the same set of 10 values and the study dummy variable, we created a linear regression model predicting donation proportion in the training set (Table 6). The overall model was significant, $F(11, 972) = 1.68$, $p = .024$, adjusted $R^2 = .011$. Self-direction was the only value that significantly predicted donations, $b = .01$, se = .01, $t = 2.16$, $p = .049$, 95% CI [0.00, 0.03]. The same model was not significant in the holdout set, $F(11, 484) = 1.37$, $p = .185$, adjusted $R^2 = .009$, and self-direction was no longer a significant predictor, $b = -0.00$, se = .01, $t = -0.04$, $p = .970$, 95% CI [−0.02, 0.02].

Lastly, we entered the 10 values as predictors of self-reported charitable acts on both the training set and holdout set. The model for the training set was non-significant, $F(10, 501) = 1.51$, $p = .132$, adjusted $R^2 = .010$. Hedonism had a significant, negative effect, $b = -.09$, se = .03, $t = -2.68$, $p = .008$, 95% CI [−0.16, −0.02]; no other effects were significant. The model for the holdout set was also not significant, $F(10, 255) = 1.39$, $p = .186$, adjusted $R^2 = .014$. Self-direction was *negatively* associated with past charitable acts, $b = -.06$, se = .03, $t = -2.05$, $p = .042$, 95% CI [−0.13, −0.00], opposite of its relationship with the decision to make a donation during the study. No other effects were significant.

### Prosocial dictionary

We fit models in which making a donation at all was regressed on the number of prosocial dictionary words and the study dummy variable. Using more prosocial dictionary words was not a significant predictor in the training set, $b = -0.03$, se = .03, OR = 0.97, $Z = -0.84$, $p = .403$, 95% CI [−0.09, 0.04], or in the holdout set, $b = 0.02$, se = .05, OR = 1.02, $Z = 0.47$, $p = .640$, 95% CI [−0.07, 0.12]. We then entered the number of prosocial dictionary word features participants used and the study dummy variable as predictors of donation proportion on the training set and holdout set. The number of prosocial words had a non-significant effect in the training set, $b = -0.01$, se = .01, $t(981) = -1.03$, $p = .304$, 95% CI [−0.02, 0.00] and in the holdout set, $b = 0.00$, se = .01, $t(493) = -0.05$, $p = .958$, 95% CI [−0.01, 0.01].

Lastly, we correlated the number of prosocial dictionary words as a predictor of past charitable involvement. The association was significant in the training set, $r(510) = .11$, 95% CI [0.02, 0.19], $p = .014$, but not in the holdout set $r(264) = .08$, 95% CI [−0.04, 0.20], $p = .189$. However, the effect sizes were similar across both data sets and had highly overlapping CIs. Overall, a composite of prosocial dictionary words may be linked to past charitable involvement, but there was no evidence of a link with charitable behaviour during the study.

### Support vector machine

We created a composite of all SVM words with each word weighted by its average weight across the 10 folds. A binary





**DOI**: 10.1002/per

logistic regression that included the study dummy variable as a covariate revealed that the SVM composite was positively associated with making a donation in the training set, $b = 0.79$, se = .12, OR = 2.20, $Z = 6.72$, $p < .001$, 95% CI [0.56, 1.02]. This association was still significant and of nearly identical magnitude after replacing the study dummy variable with self-reported charitable acts. On the holdout set, the SVM composite did not significantly predict donation decisions, $b = 0.27$, se = .15, OR = 1.31, $Z = 1.81$, $p = .070$, 95% CI [0.56, 1.02]. The fact that the holdout regression coefficient was lower than the lower confidence limit of the same effect in the training data set is consistent with overfitting.

A linear regression demonstrated that the SVM composite positively predicted donation proportion, $b = 0.07$, se = .02, $t(982) = 4.48$, $p < .001$, 95% CI [0.04, 0.10]. The size and significance of the effect was qualitatively unchanged after adding self-reported charitable acts to the model. But the SVM composite did not significantly predict donation proportion on the holdout set, $b = 0.02$, se = .02, $t(493) = 1.14$, $p = .253$, 95% CI [−0.02, 0.07], again consistent with overfitting.

Lastly, we computed the correlation between the SVM composite and self-reported charitable acts. The association was positive but only marginally significant in the training set, $r(494) = .08$, 95% CI [−0.01, 0.17], $p = .073$, and nil in the holdout set, $r(264) = .00$, 95% CI [−0.11, 0.12], $p = .988$. Overall, the SVM was successful only in detecting word features that predicted study donation behaviour in the training set.

### Exploratory analyses: Can humans extract information about prosociality from personal strivings?

We found that neither entering all 10 universal values as simultaneous predictors, summing together all prosocial words from a participant's strivings, nor using a weighted composite of all SVM words yielded significant effects in both the training and holdout data sets. These null effects imply that lists of personal strivings are not a rich source of information about how prosocial people are. However, it is possible that personal strivings do contain information relevant to donation behaviour that other coding methods could extract.

We speculated that with sparse text, it is necessary to use a coding method that (i) can understand the contextual meaning of words and (ii) is designed to predict the outcome of interest. None of the three coding methods we used possess both features. The manual coding approach met the first criterion, but SVM and the prosocial word dictionary methods did not. The SVM met the second criterion, but the manual coding and prosocial word dictionary approaches made more abstract characterisations and were thereby one step removed from predicting donations.

One coding method that meets both criteria is having human raters familiar with the design of the original studies read the personal strivings and guess how much people donated to charity. Not only can human raters understand language in context, but they can also accurately perceive how cooperative a target is from a brief face-to-face interaction, even when the target has a monetary incentive to convince others that he or she is cooperative (Sparks, Burleigh, & Barclay, 2016). It is plausible then that human raters could detect evidence of a desire to help needy others from participants' strivings, even if some participants tried to present themselves as pursuing noble goals. We therefore had human coders guess participants' donation amounts and used their guesses to predict donation behaviour (convergent validity), previous charitable acts (generalisability), and—for significant predictors—donation behaviour over and above self-reported previous charitable acts (incremental validity).

Before making predictions, each rater read detailed instructions (https://osf.io/qmjyp/) describing the protocols that participants completed, the purpose of the studies, and how to complete the task. In addition to guessing how many cents a given participant gave out of the total number of cents he or she had available, raters were also instructed to articulate the primary basis for their prediction by cutting and pasting words or describing relevant themes they noticed across a participant's strivings. We had raters articulate their decisions to encourage them to make evidence-based predictions.

There were five raters overall. Each participant was evaluated by two raters, but we did not always have the same two raters evaluate the same strivings; thus, we regarded the raters as random effects. We observed poor interrater reliability (intraclass correlation coefficient = .47; Koo & Li, 2016), which is consistent with sparseness affecting the ability of judges to detect diagnostic information. We averaged the two raters' guesses, and then divided this average by the amount of money participants had available to donate. This proportion was used to represent raters' guesses in our regression models.

We first created logistic regression models in which donation decisions were regressed upon rater guesses and the study dummy variable. In the training set, rater guess was a non-significant *negative* predictor of donation decisions, $b = -0.06$, se = 0.22, OR = 0.95, $Z = -0.25$, $p = .801$, 95% CI [−0.49, 0.38]. The same model yielded similar results in the holdout set, $b = -0.46$, se = 0.30, OR = 0.63, $Z = -1.53$, $p = .127$, 95% CI [−1.06, 0.13]. We then fit the same predictor variables to linear regression models in which donation proportion was the outcome. Rater guess was a non-significant predictor in the training set, $b = 0.01$, se = 0.03, $t(981) = 0.35$, $p = .728$, 95% CI [−0.05, 0.08], and in the holdout set, $b = -0.06$, se = 0.04, $t(981) = -1.35$, $p = .176$, 95% CI [−0.15, 0.03]. Because the effect of guesses on donation behaviour were not significant, we did not examine incremental validity.

We then examined how well guesses predicted past charitable involvement. Guessing a higher donation amount was moderately, positively associated with past charitable involvement in both the training set, $r(510) = .17$, 95% CI [0.09, 26], $p < .001$, and in the holdout set, $r(264) = .19$, 95% CI [0.08, 31], $p < .001$. Ironically, rater guesses were only predictive of past charitable involvement even though raters were guessing how much participants would donate within in the study situation.



**DOI**: 10.1002/per

## DISCUSSION

Individual differences in prosociality are hard to measure because directly asking people about sociality desirable traits evokes self-presentation concerns. Qualitative data offer a potentially more honest source of information about prosocial traits. Researchers have long relied on manual methods of coding open-ended text data generated by small samples of moral exemplars. In a large sample of individuals who were not selected based on past prosocial acts, we used the VEiNs manual (Frimer et al., 2009), a prosocial dictionary (Frimer et al., 2014), and an SVM algorithm (Dehghani et al., 2017) to extract variables from participants' personal strivings. The extracted variables were compared in their ability to predict charitable giving on a training set and holdout set.

Although none of these three methods yielded reliable predictors of donation behaviour, there were two possibly notable effects. First, self-direction values predicted whether participants made a non-zero donation in both the initial analyses and in the holdout analysis. This finding is consistent with Schwartz's (2010) suggestion that those high in self-direction may help others more because they are less worried about threats to their own well-being. However, the effects were small and just crossed the thresholds of significance in both the training and holdout sets. Moreover, self-direction did not predict donation amount in the holdout set, was not significantly associated with self-reported charitable behaviour in the training set, and was negatively associated with self-reported charitable behaviour in the holdout set. Second, the number of prosocial dictionary words in peoples' strivings predicted participants' reported history of charitable giving in the training set. Although the effect size was comparable, this effect did not reach significance in the holdout analyses.

Upon finding that none of these methods was clearly useful in explaining variation in charitable giving, we had human raters try to guess how much participants donated to charity based on reading their personal strivings. These guesses were also no better than chance. However, rater guesses were highly significant predictors of past charitable involvement in both the training and holdout data sets. Because this analysis was unplanned and was one of a large number of tests we conducted, we encourage future researchers to try to replicate this result on new data sets. If our results are robust, they suggest that past charitable involvement correlates with raters' guesses and with donation behaviour in the research studies for different reasons.

We speculate that individual differences in charitable giving are in part due to trait prosociality and in part due to stable tendencies related to charitable giving specifically. That is, donations to charity can reflect a genuine desire to help other people (which would manifest in other behaviours and goals beyond charity), as well as charity-specific tendencies, such as those related to religious duties (which would not manifest in contexts unrelated to charitable organisations). Raters could likely glean evidence of a general disposition towards prosociality from people's strivings (e.g. based on the number of strivings that referred to helping other people), and so rater guesses about charitable donation could have been associated with the portion of past charitable involvement that reflects trait prosociality. In contrast, donation behaviour during the studies could relate to tendencies to give to charity specifically, which would not necessarily manifest in personal strivings. Thus, donation behaviour during the study was associated with the portion of past charitable involvement that reflects tendencies to give to charity, but not with rater guesses.

A non-exclusive alternative possibility is that self-presentation concerns affected reports of past charitable involvement and donations during the study, but not descriptions of personal strivings. Being explicitly asked to donate to charity elicits concerns about appearing callous (Andreoni, Rao, & Trachtman, 2017). In contrast, the personal strivings task ostensibly has nothing to do with prosocial behaviour and so does not evoke the same self-presentation concerns (Frimer et al., 2014).

### Do personal strivings contain information about prosociality?

The overall pattern of results imply that strivings do contain some information about individual differences in charitable giving, but the method for coding personal strivings and the outcome measures used matter. We found that only human raters' guesses (and possibly the prosocial dictionary) could predict past charitable involvement, and even human raters could not guess charitable behaviour during the study. Using a single behavioural outcome to assess individual differences in charitable giving may be too noisy and consequently understate the predictive power of strivings (Epstein & O'Brien, 1985). For example, someone who strives to donate to worthy causes may have refrained from donating in our studies simply because they did not like the charities that we exposed them to, or because they had already committed their discretionary funds to another charitable cause. The measure of self-reported past charitable acts is a fairer criterion in the sense that it was measured at a higher level of generality.

We also may have had more success predicting donations had we coded for different content. For instance, many researchers code textual data for *generativity*, which reflects a preoccupation with making a positive impact on the next generation (Emmons, 1999; Mansfield & McAdams, 1996) and is linked to charitable donations (Sikkel & Schoenmakers, 2012). Alternatively, we may have uncovered stronger predictors had we measured participants' attitudes towards their own strivings rather than coding their content. Goodman, Kashdan, Stiksma, and Blalock (2019) found that although people with social anxiety disorder did not differ much in the content of their strivings from healthy people, they were much more likely to report pursuing their strivings out of controlled motives (e.g. avoiding guilt, pleasing others) rather than autonomous motives (e.g. upholding a cherished value, pleasure; Deci & Ryan, 2000). Of course, measuring striving properties using questionnaire methods



**DOI**: 10.1002/per

moves away from the idiographic approach that has historically made open-ended text data appealing to prosocial behaviour researchers.

Lastly, it is possible different elicitations of personal strivings could have enabled even our original three coding methods to be successful. Perhaps Emmons' (1999) popular sentence completion task simply does not yield enough text to reliably capture relevant content. The prevalence of dictionary and SVM words was low even after removing especially rare word features and combining together words together that have the same stem. Even aggregating across many word features may not be enough to overcome sparseness if the relationships between individual word features and charitable behaviour are small.

In order to extract information about prosociality from open-ended text, researchers will likely need to use tasks that yield either a lot of data or data directly relevant to prosociality. Increasing the quantity of relevant text will require personality psychologists to move beyond well-established measures like the Personal Strivings List that are simple and cheap to administer. For instance, life story narratives have shown promise in differentiating moral exemplars from matched controls (Frimer et al., 2011), but take over an hour to complete and usually require an in-person interview. When such administrative costs prove too burdensome, researchers could analyse publicly available text data generated by individuals that are known to have engaged in above-average levels of prosocial behaviour (e.g. Rand & Epstein, 2014). A third alternative would be to develop essay tasks in which participants generate a large amount of text. Regardless of which alternative researchers choose, they must strike a balance between directing participants to provide text that reflects their trait prosociality while but not being so direct as to elicit socially desirable responding.

Although the problem caused by the brevity of personal strivings is most apparent for automated methods, they can also affect the validity of manual methods. Although the least common universal values were more common than most SVM and prosocial dictionary words (Table S1), many coding decisions were educated guesses rather than confident judgements. Most raters found coding the strivings difficult because they contained little text, few participants gave any indication why they pursue the goals that they described, and the VEiNs manual did not give guidance on how to code many strivings that were common in our data sets (likely because the manual was originally applied to exemplar studies rather than to studies sampling from the Mechanical Turk population). Achieving minimal standards for interrater reliability ($\kappa \geq .60$) partially depended on agreeing to coding rules that were plausible but of unverifiable validity (see https://osf.io/q85d4/).

## CONCLUSION

We set out to discover whether automated approaches could rival or even improve upon manual coding of qualitative data. We used prediction of donation behaviour as a yardstick of performance, given that researchers have historically used strivings and other open-ended text measures to gain insight into prosocial behaviour. Expecting an embarrassment of riches, we found that neither manual nor automated methods extracted striving features that were reliably associated with either actual donation behaviour or self-reported past charitable behaviour. Instead, raters' guesses of donation amounts predicted past charitable involvement, but not donation behaviour during the study. Raters' ability to extract whatever context they believed was relevant to identifying charitable individuals may explain why their guesses were more accurate than the original coding methods we used. Although personal strivings are just one kind of qualitative data that might be used to capture individual differences in prosociality, our findings suggest that the utility of qualitative data in yielding insight into prosocial individuals can be highly dependent on measurement decisions.

Our results serve as a cautionary tale that big data analysis strategies cannot overcome inherent weaknesses in data such as text sparseness. On the other hand, a routine feature of big data analysis—cross-validation—is precisely what enabled us to conclude that ostensible predictors of charitable giving were unlikely to prove useful in new samples. Without our holdout analyses, we may have come to more positive conclusions about using personal strivings to understand or predict prosocial behaviour. In doing so, we may have inspired other researchers to squander their resources. Our use of a holdout also gave us licence to conduct a wide range of exploratory tests without great risk of basing our conclusions on false positives: Any chance associations that we might have 'uncovered' in the training set presumably would not have replicated in the holdout. It is for this reason that we can have some confidence that the association between rater guesses and self-reported charitable acts is not merely the product of chance. We conclude that regardless of whether personality psychologists use data-driven techniques to better understand behaviour, they could shorten the cycle of self-correction in science by borrowing big data techniques designed to reduce overfitting.

## ACKNOWLEDGEMENTS

We would like to thank Brooke Donner, Kendall Mather, Lindsey Hoshaw, Erika Boone, Michele Marenus, Abdulrahman Bindamnan, Andrea Rosales, Mia Graham, Kelis Johnson, and Sonia He for their painstaking effort in coding personal strivings. Study 1 was funded by a University of Miami Department of Psychology Dissertation Award given to W. H. B. McAuliffe and a grant from the John Templeton Foundation (award 29165) awarded to M. E. McCullough. Study 2 was funded by a grant from the University of Miami College of Arts and Sciences awarded to M.E. McCullough to investigate the impacts of Hurricane Irma. While writing this manuscript, W. H. B. McAuliffe received support from a University of Miami College of Arts and Sciences Dissertation Award.





**DOI**: 10.1002/per

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

## REFERENCES


Andreoni, J., Rao, J. M., & Trachtman, H. (2017). Avoiding the ask: A field experiment on altruism, empathy, and charitable giving. *Journal of Political Economy*, *125*, 625–653.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods*, *8*, 129–148. https://doi.org/10.1037/1082-989x.8.2.129.

Azen, R., & Traxel, N. (2009). Using dominance analysis to determine predictor importance in logistic regression. *Journal of Educational and Behavioral Statistics*, *34*, 319–347.

Bainter, S., McCauley, T. G., Wager, T., & Losin, E. A. R.. (in press). Improving practices for selecting a subset of important predictors in psychology: An application to predicting pain. *Advances in Methods and Practices in Psychological Science.*. https://doi.org/10.31234/osf.io/j8t7s.

Banks, G. C., Woznyj, H. M., Wesslen, R. S., & Ross, R. L. (2018). A review of best practice recommendations for text analysis in R (and a user-friendly app). *Journal of Business and Psychology*, *33*, 445–459.

Barclay, P., & Willer, R. (2006). Partner choice creates competitive altruism in humans. *Proceedings of the Royal Society B: Biological Sciences*, *274*, 749–753.

Batson, C. D., & Shaw, L. L. (1991). Evidence for altruism: Toward a pluralism of prosocial motives. *Psychological Inquiry*, *2*, 107–122.

Baumard, N., André, J. B., & Sperber, D. (2013). A mutualistic approach to morality: The evolution of fairness by partner choice. *Behavioral and Brain Sciences*, *36*, 59–78.

Bekkers, R., & Wiepking, P. (2011). Accuracy of self-reports on donations to charitable organizations. *Quality & Quantity*, *45*, 1369–1383.

Bleidorn, W., & Hopwood, C. J. (2019). Using machine learning to advance personality assessment and theory. *Personality and Social Psychology Review*, *23*, 190–203. https://doi.org/10.1177/1088868318772990.

Bustos, C. & Soares, F.C. (2019) *The dominance analysis package*. https://CRAN.R-project.org/package=dominanceanalysis

Colby, A., & Damon, W. (1992). *Some do care*. Free press.

Damon, W., & Colby, A. (2015). *The power of ideals: The real story of moral choice*. USA: Oxford University Press.

Deci, E. L., & Ryan, R. M. (2000). The "what" and" why" of goal pursuits: Human needs and the self-determination of behavior. *Psychological Inquiry*, *11*, 227–268.

Dehghani, M., Johnson, K. M., Garten, J., Boghrati, R., Hoover, J., Balasubramanian, V., … Parmar, N. J. (2017). TACIT: An open-source text analysis, crawling, and interpretation tool. *Behavior Research Methods*, *49*, 538–547. https://doi.org/10.3758/s13428-016-0722-4.

Dunlop, W. L. (2015). Contextualized personality, beyond traits. *European Journal of Personality*, *29*, 310–325.

Emmons, R. A. (1999). *The psychology of ultimate concerns: Motivation and spirituality in personality*. New York, NY: Guilford Press.

Epstein, S., & O'Brien, E. J. (1985). The person–situation debate in historical and current perspective. *Psychological Bulletin*, *98*, 513–537.

Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*, 1149–1160.

Firmin, R. L., Bonfils, K. A., Luther, L., Minor, K. S., & Salyers, M. P. (2017). Using text-analysis computer software and thematic analysis on the same qualitative data: A case example. *Qualitative Psychology*, *4*, 201.

Frimer, J. A., Aquino, K., Gebauer, J. E., Zhu, L. L., & Oakes, H. (2015). A decline in prosocial language helps explain public disapproval of the US Congress. *Proceedings of the National Academy of Sciences*, *112*, 6591–6594.

Frimer, J. A., Schaefer, N. K., & Oakes, H. (2014). Moral actor, selfish agent. *Journal of Personality and Social Psychology*, *106*, 790–802.

Frimer, J. A., Walker, L. J., & Dunlop, W. L. (2009). *Values embedded in Narrative (VEiN) coding manual*. Department of Psychology, University of British Columbia, Vancouver, Canada: Unpublished manuscript.

Frimer, J. A., Walker, L. J., Dunlop, W. L., Lee, B. H., & Riches, A. (2011). The integration of agency and communion in moral personality: Evidence of enlightened self-interest. *Journal of Personality and Social Psychology*, *101*, 149–163. https://doi.org/10.1037/a0023780.

George, E. I., & McCulloch, R. E. (1993). Variable selection via Gibbs sampling. *Journal of the American Statistical Association*, *88*, 881–889.

Goodman, F. R., Kashdan, T. B., Stiksma, M. C., & Blalock, D. V. (2019). Personal strivings to understand anxiety disorders: Social anxiety as an exemplar. *Clinical Psychological Science*, *7*, 283–301.

Goodwin, G. P., Piazza, J., & Rozin, P. (2014). Moral character predominates in person perception and evaluation. *Journal of Personality and Social Psychology*, *106*, 148–168. https://doi.org/10.1037/a0034726.

Graziano, W. G., & Eisenberg, N. (1997). Agreeableness: A dimension of personality. In *Handbook of personality psychology* (pp. 795–824). Academic Press.

Hart, H. M., McAdams, D. P., Hirsch, B. J., & Bauer, J. J. (2001). Generativity and social involvement among African Americans and White adults. *Journal of Research in Personality*, *35*, 208–230.

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine*, *15*, 155–163. https://doi.org/10.1016/j.jcm.2016.02.012.

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 159–174.

List, J. A. (2011). The market for charitable giving. *Journal of Economic Perspectives*, *25*, 157–180.

Magee, J. C., & Langner, C. A. (2008). How personalized and socialized power motivation facilitate antisocial and prosocial decision-making. *Journal of Research in Personality*, *42*, 1547–1559.

Mansfield, E. D., & McAdams, D. P. (1996). Generativity and themes of agency and communion in adult autobiography. *Personality and Social Psychology Bulletin*, *22*, 721–731.

Marsh, A. A., Stoycos, S. A., Brethel-Haurwitz, K. M., Robinson, P., VanMeter, J. W., & Cardinale, E. M. (2014). Neural and cognitive characteristics of extraordinary altruists. *Proceedings of the National Academy of Sciences*, *111*, 15036–15041.

McAdams, D. P. (1995). What do we know when we know a person? *Journal of Personality*, *63*, 365–396.

McAuliffe, W. H. B. (2019). Can studies of trait altruism be trusted? *Open Access Dissertations.*, *2252*. https://scholarlyrepository.miami.edu/oa_dissertations/2252.

McCauley, T.G., and McCullough, M.E. (2019). *Determined donors: An experimental investigation of moral excellence*. Working paper.

Oliner, S. P., & Oliner, P. M. (1988). *The altruistic personality*. New York: Free Press.

Park, G., Schwartz, H. A., Eichstaedt, J. C., Kern, M. L., Kosinski, M., Stillwell, D. J., … Seligman, M. E. (2015). Automatic personality assessment through social media






**DOI**: 10.1002/per

language. *Journal of Personality and Social Psychology*, *108*, 934–952.
Paulhus, D. L., & John, O. P. (1998). Egoistic and moralistic biases in self-perception: The interplay of self-deceptive styles with basic traits and motives. *Journal of Personality*, *66*, 1025–1060.
Pennebaker, J. W., Booth, R. J., & Francis, M. E. (2007). *Linguistic inquiry and word count: LIWC [computer software]*. Austin, TX: LIWC.net.
Preacher, K. J., Rucker, D. D., MacCallum, R. C., & Nicewander, W. A. (2005). Use of the extreme groups approach: a critical reexa mination and new recommendations. *Psychological Methods*, *10*(2), 178-192.
Prospect Research and Wealth Screening Statistics (2015, *January 7)*. Retrieved from https://www.donorsearch.net/prospect-research-statistics/
R Core Team (2019). R: A language and environment for statistical computing. *R Foundation for Statistical Computing*, Vienna, Austria. Retrieved from https://www.r-project.org.
Rand, D. G., & Epstein, Z. G. (2014). Risking your life without a second thought: Intuitive decision-making and extreme altruism. *PLoS ONE*, *9*, e109687. https://doi.org/10.1371/journal.pone.0109687.
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.
Schönbrodt, F. D., & Perugini, M. (2013). At what sample size do correlations stabilize? *Journal of Research in Personality*, *47*, 609–612.
Schwartz, S. H. (1994). Are there universal aspects in the structure and contents of human values? *Journal of Social Issues*, *50*, 19–45.
Schwartz, S. H. (2010). Basic values: How they motivate and inhibit prosocial behavior. *Prosocial motives, emotions, and behavior: The better angels of our nature*, *14*, 221–241.
Seeboth, A., & Mõttus, R. (2018). Successful explanations start with accurate descriptions: Questionnaire items as personality markers for more accurate predictions. *European Journal of Personality*, *32*(3), 186-201.
Sikkel, D., & Schoenmakers, E. (2012). Bequests to health-related charitable organisations: A structural model. *International Journal of Nonprofit and Voluntary Sector Marketing*, *17*, 183–197.
Sparks, A., Burleigh, T., & Barclay, P. (2016). We can see inside: Accurate prediction of Prisoner's Dilemma decisions in announced games following a face-to-face interaction. *Evolution and Human Behavior*, *37*, 210–216.
Sun, J., & Vazire, S. (2019). Do people know what they're like in the moment? *Psychological Science*, *30*, 405–414. https://doi.org/10.1177/0956797618818476.
Thielmann, I., Spadaro, G., & Balliet, D. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*, 30–90.
Vazire, S. (2010). Who knows what about a person? The self–other knowledge asymmetry (SOKA) model. *Journal of Personality and Social Psychology*, *98*, 281–300. https://doi.org/10.1037/a0017908.
Walker, L. J., & Frimer, J. A. (2007). Moral personality of brave and caring exemplars. *Journal of Personality and Social Psychology*, *93*, 845–860. https://doi.org/10.1037/0022-3514.93.5.845.
Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, *12*, 1100–1122.
Zhao, K., & Smillie, L. D. (2015). The role of interpersonal traits in social decision making: Exploring sources of behavioral heterogeneity in economic games. *Personality and Social Psychology Review*, *19*, 277–302.



**DOI**: 10.1002/per