Judgment and Decision Making, Vol. 17, No. 1, January 2022, pp. 31–49

Assessing the test-retest reliability of the social value orientation slider measure

Carlos A. de Matos Fernandes*† Dieko M. Bakker*‡ Jacob Dijkstra*

Abstract

Decades of research show that (*i*) social value orientation (SVO) is related to important behavioral outcomes such as cooperation and charitable giving, and (*ii*) individuals differ in terms of SVO. A prominent scale to measure SVO is the social value orientation slider measure (SVOSM). The central premise is that SVOSM captures a stable trait. But it is unknown how reliable the SVOSM is over repeated measurements more than one week apart. To fill this knowledge gap, we followed a sample of N = 495 over 6 months with monthly SVO measurements. We find that continuous SVO scores are similarly distributed (Anderson-Darling k-sample $p = 0.57$) and highly correlated ($r \geq 0.66$) across waves. The intra-class correlation coefficient of 0.78 attests to a high test-retest reliability. Using multilevel modeling and multiple visualizations, we furthermore find that one's prior SVO score is highly indicative of SVO in future waves, suggesting that the slider measure consistently captures one's SVO. Our analyses validate the slider measure as a reliable SVO scale.

Keywords: social value orientation, SVO, test-retest, slider measure, reliability

1 Introduction

"Personality traits are probabilistic descriptions of relatively stable patterns of emotion, motivation, cognition, and behavior" (DeYoung, 2015, p. 64). Social value orientation

*Department of Sociology/Interuniversity Center for Social Science Theory and Methodology (ICS), University of Groningen, the Netherlands.

†c.a.de.matos.fernandes@rug.nl. ORCID 0000-0002-3664-4989

‡ORCID 0000-0002-5451-1979

Data, materials, and analyses code are available on the Open Science Framework (OSF): https://osf.io/tw8dq/?view_only=3ccfc9eb774b49e687bacdb729e7b4a6.

The first and third authors acknowledge that this study is funded by the research program Sustainable Cooperation – Roadmaps to Resilient Societies (SCOOP) funded by NWO and the Dutch Ministry of Education, Culture, and Science (OCW) in its 2017 Gravitation Program (grant number 024.003.025).

We thank Jon Baron (editor), Ryan O. Murphy (reviewer), an anonymous reviewer, and members of the Norms and Networks Cluster (NNC) at the University of Groningen for thoughtful feedback and suggestions which enriched our paper.

(SVO) is purportedly such a personality trait and is frequently invoked as an explanation for individual variation in cooperative behavior (Van Lange et al., 2014). To qualify as a personality trait, however, SVO must be *stable* over time. In particular, empirical measures of SVO should exhibit high degrees of test-retest reliability. The empirical evidence for this is currently scant. Therefore, we test whether the SVO slider measure (SVOSM), a very popular SVO measure that has been frequently used after its introduction in 2011 (Bakker & Dijkstra, 2021; Murphy et al., 2011), captures a stable personality trait. Stability is an important psychometric property required of any measure claiming to translate to an internally valid, consistent, and reliable assessment of the studied trait. Measuring SVO reliably has long been a scientific goal (Au & Kwong, 2004; Balliet et al., 2009; Murphy & Ackermann, 2014). We contribute to reaching this goal by analyzing SVOSM panel data (N = 495) from six-monthly repeated measures and assessing test-retest reliability.

Even though several measures exist (Thielmann et al., 2020), we focus on the SVOSM because this measure is specifically designed to assess a trait related to cooperation, namely: SVO is defined by the weight individuals assign to their own and others' outcomes in situations of interdependence (Messick & McClintock, 1968). Primary reasons for researchers to rely on the SVOSM, instead of other measures, include the fact that SVOSM is not very burdensome for participants (consisting of just 6 items), has clear consistency checks, purportedly has high test-retest reliability, and yields a continuous score (Murphy & Ackermann, 2014). Categorical classifications may fail to capture individual differences in SVO (Bakker & Dijkstra, 2021) and the SVOSM allows researchers to utilize continuous scores. Even though most researchers utilize the slider measure to capture SVO as a categorical construct, the designers of the SVOSM recognized that SVO is "best represented as a continuous scale" (Murphy et al., 2011, p. 772). We move in this paper beyond treating SVO as a category and rely on SVO as a continuous construct. A similar approach is conducted by Fleeson (2001) who studied the Big Five as a distributional continuous measure rather than discrete categorical ones. Yet, we go further than Fleeson and inspect whether distributions of SVO continuous scores are alike over time.

SVO is generally taken to be a stable construct (Bogaert et al., 2008; Van Lange et al., 2014). If this is true, repeated measurements of the SVOSM should show stable and strong associations between continuous SVO scores over longer periods. Yet, the little research into the test-retest reliability of the SVOSM there is used measurements just one week apart. With our panel design of 6 measurements one *month* apart, we remedy this situation.

Assessing test-retest reliability is of the highest relevance both empirically and methodologically. Individuals with high SVO scores are shown to cooperate more than individuals low on SVO both in observational studies, e.g., volunteering (Manesi et al., 2019), and in experimental ones (Balliet et al., 2009). Establishing the test-retest reliability of SVO strengthens its case as a reliable predictor of cooperation (and other behaviors). Apart from employing a design with longer time intervals between measurements, we also advance the field by relying on a non-student sample. Van Lange et al. (2014, p. 148) posit that we

know surprisingly little about SVO in non-student samples. We rise to the occasion and use a representative sample of the Dutch population.

In the remainder of this paper, we first discuss previous research. We then describe the data collection process and our sample, followed by a presentation of our findings. We end this paper with prospects for future research and some concluding remarks.

2 What we know thus far

Previous assessments of SVOSM's test-retest reliability are encouraging. With approximately one hour between two waves (N = 124), Ackermann & Murphy (2019) report a correlation of 0.72 between SVOSM scores. Most other studies use a two measurements design, one week apart. One study with N = 872 reports a correlation of 0.79 between continuous SVOSM scores (Höglinger & Wehrli, 2017). The developers of the SVOSM report a correlation of 0.92 with a sample of 46 students (Murphy et al., 2011). Another study reports a correlation of 0.75 in a "non-monetary" condition (only show-up fee; N = 155 students) and a correlation of 0.35 in an incentivized "monetary" condition with N = 62 (Reyna et al., 2018). Hence, the SVOSM seems relatively stable across one-week periods but some measurement-to-measurement variation is present. There is only one study, to our knowledge, investigating the test-retest reliability of the SVOSM in a much longer time frame. Bakker & Dijkstra (2021) report a correlation of 0.60 between continuous SVOSM scores, relying on N = 86 students and two measurements three months apart. On the one hand, temporal instability may result from random measurement errors. On the other, it may result from SVO being systematically affected by, for example, personal experiences. All in all, prior research on the temporal stability of the SVOSM suffers from two defects: (*i*) studies either use very short time frames or have low sample sizes when using longer time frames, and (*ii*) studies rely exclusively on student samples. We remedy both shortcomings.

3 Method

3.1 Social value orientation slider measure

The SVOSM has six items. Per item, respondents are asked to make a decision indicating how they wish to allocate units of some hypothetically valuable good between themselves and a random other person. Each item contains several alternative resource allocations, with the ranges of own and others' payoff changing across items. Respondents were informed about the hypothetical nature of the questions and did not earn extra money in addition to their participation fee. We chose this non-incentivized design because most existing SVO studies do not use monetary incentives. In light of the findings of Reyna et al. (2018) mentioned above, however, investigating test-retest reliability across longer time frames in non-student samples in incentivized designs also seems valuable. We leave this question

for future research. Finally, to calculate SVOs we need to compute each respondent's SVO degree. We first calculated the mean payoff allocated to themselves and the other for all items and then measure a single SVO degree score based on the mean-self to mean-other ratio (see Murphy et al., 2011; Murphy & Ackermann, 2015, for more information on the measure and how to compute continuous SVO scores).

3.2 Data collection and our sample

We used a 6-month repeated measures design where respondents filled in the SVOSM each month in a non-experimental context. The first wave of data collection occurred in January 2021 with subsequent waves administered in February (wave 2), March (wave 3), April (wave 4), May (wave 5), and June (wave 6). Questionnaires started with an introduction, followed by an example SVO question to get acquainted with the type and format of SVO questions. Then respondents answered six allocation questions. Data were collected by the Longitudinal Internet studies for the Social Sciences (LISS) panel administered by CentERdata (Tilburg University, the Netherlands). The LISS panel is a representative sample of Dutch individuals. The panel is based on a true probability sample of households drawn from the population register and consists of 4500 households, comprising 7000 individuals. Our sample (N = 495) is a random subset of the panel. We expect the full data set to be publicly available in due course at https://www.lissdata.nl/.

3.3 Consistency check

A key property of the SVOSM is its consistency check, allowing researchers to exclude respondents who are inconsistent in their allocation preferences. This may indicate random answers or a lack of understanding. Murphy et al. (2011) suggested excluding respondents whose answers result in intransitive preferences over SVOs. As an alternative, Bakker & Dijkstra (2021) suggest excluding respondents whose answers were so inconsistent that their resulting vector is too short (i.e., whether distance, D, is smaller than some cutoff value). The more consistently a respondent chooses allocations corresponding to a particular SVO, the longer their D will be. Vectors shorter than 35 are considered inconsistent and are excluded from the sample.[1] For more information on computing vector lengths and the mathematical function, we refer the reader to the supplementary file attached to Bakker & Dijkstra (2021). We find that across all waves 9 percent of answer profiles are intransitive while 7 percent fail the vector length criterion. A total of 302 (approximately 14%) out

[1]The choice for 35 as the criterion is based on 39.99 (mean vector length) $- 2 * 2.47$ (standard deviation). Additional analyses using 37.5 (applied by Bakker & Dijkstra, 2021) or 40 show that relying on stricter vector length criteria leads to more stringent filtering of then considered inconsistent answer profiles: excluding 12.5 and 36.4% of responses respectively. The intra-class correlation coefficient (see section 4.2) goes up from 0.78 (D = 35, N = 230) to 0.81 (D = 37.5, N = 214) and 0.90 (D = 40, N = 103). A more conservative vector criterion leads, as expected, to fewer inconsistencies in answer profiles and higher test-retest reliability.

of 2176 responses were excluded because they failed to meet the transitivity criterion, the vector length criterion, or both.

4 Results

4.1 Distribution of SVO in our sample

In Table 1, we provide descriptive statistics for all six items in our questionnaire. Generally, we find that the mean scores do not vary that much. The standard deviations across payoffs allocated to themselves and the other, however, do show some variance. Especially items 1-other, 6-other, 4, and 5 show differences in allocation choices. Murphy et al. (2011) denote that, next to the self-other dimension, SVO items capture differences in preferences for maximizing own and others' outcomes and (in)equality. If respondents

TABLE 1: Inspecting the six SVO items separately, average payoffs to self and the other, and average SVO scores. Payoff ranges of items 1 to 6 are reported in a note below.

Item	Wave 1 *M* (*SD*)	Wave 2 *M* (*SD*)	Wave 3 *M* (*SD*)	Wave 4 *M* (*SD*)	Wave 5 *M* (*SD*)	Wave 6 *M* (*SD*)
1-self	85 (0)	85 (0)	85 (0)	85 (0)	85 (0)	85 (0)
1-other	79.1 (14.3)	81.3 (11.5)	82.2 (10.0)	82.1 (10.6)	82.0 (10.4)	82.1 (10.8)
2-self	99.3 (2.2)	99.5 (1.6)	99.5 (1.5)	99.5 (1.7)	99.5 (1.6)	99.7 (0.9)
2-other	48.3 (5.1)	48.8 (3.7)	48.9 (3.6)	48.9 (4.0)	48.9 (3.8)	49.2 (2.1)
3-self	82.8 (6.5)	83.4 (5.4)	83.0 (6.5)	82.8 (6.8)	83.1 (6.1)	83.4 (5.3)
3-other	85.9 (2.8)	85.7 (2.3)	85.9 (2.8)	85.9 (2.9)	85.8 (2.6)	85.7 (2.3)
4-self	68.7 (10.1)	69.2 (10.0)	68.6 (9.8)	68.6 (9.6)	68.6 (9.8)	69.5 (10.2)
4-other	54.6 (24.6)	53.4 (24.2)	54.8 (23.9)	54.9 (23.3)	54.9 (23.7)	52.7 (24.7)
5-self	83.8 (11.5)	83.3 (11.4)	83.0 (11.6)	81.7 (11.4)	82.3 (11.6)	83.4 (11.5)
5-other	66.2 (11.5)	66.7 (11.4)	67.0 (11.6)	68.3 (11.4)	67.7 (11.6)	66.6 (11.5)
6-self	89.7 (6.2)	89.1 (5.9)	89.1 (6.1)	88.9 (5.8)	89.0 (6.0)	89.3 (6.2)
6-other	74.0 (14.5)	75.4 (13.8)	75.4 (14.1)	57.9 (13.6)	75.6 (14.0)	75.1 (14.4)
Self	84.9 (4.4)	84.9 (4.2)	84.7 (4.4)	84.4 (4.3)	84.6 (4.4)	85.0 (4.4)
Other	68.0 (8.7)	68.5 (8.2)	69.0 (8.0)	69.3 (7.9)	69.1 (8.0)	68.6 (8.3)
SVO	27.3 (13.7)	28.0 (12.9)	28.9 (12.8)	29.4 (12.5)	29.1 (12.7)	28.0 (13.0)

Note. *M* = mean; *SD* = standard deviation; Ranges of the SVOSM items in the questionnaire comprise from left to right: 1-self = 85, 1-other = 85 to 15, 2-self = 85 to 100, 2-other = 15 to 50, 3-self = 50 to 85, 3-other = 100 to 85, 4-self = 50 to 85, 4-other = 100 to 15, 5-self = 100 to 50, 5-other = 50 to 100, 6-self = 100 to 85, and 6-other = 50 to 85.

favor maximizing their payoff, then they tend to select a self-payoff of 85 (other-payoff = 15) in item 4. Similarly, if respondents prefer equality in outcomes, then they would choose an allocation in, for example, item 5 that leads to an equal distribution. Yet, respondents favoring inequality in outcomes choose either a higher payoff for themselves or the other. The standard deviations in said items show variation in payoff allocations across waves, attesting to the need of assessing test-retest reliability of SVO via distributions and not solely based on mean scores or discrete categories. Finally, the observed average payoff allocated to themselves of all six items combined ranges from 67 to 93, with a mean of 84.7 (SD = 4.3). Conversely, the average payoff allocated to the other ranges from 38 to 87, with a mean of 68.8 (SD = 8.2). The mean payoff scores to self and the other vary little across waves.

We now turn to the distribution of SVO degrees (the continuous scores) in our sample (Table 1 and Figure 1). Individual SVO degree scores are based on answer profiles on all six items, summarized as the mean payoffs allocated to themselves and the other (as shown

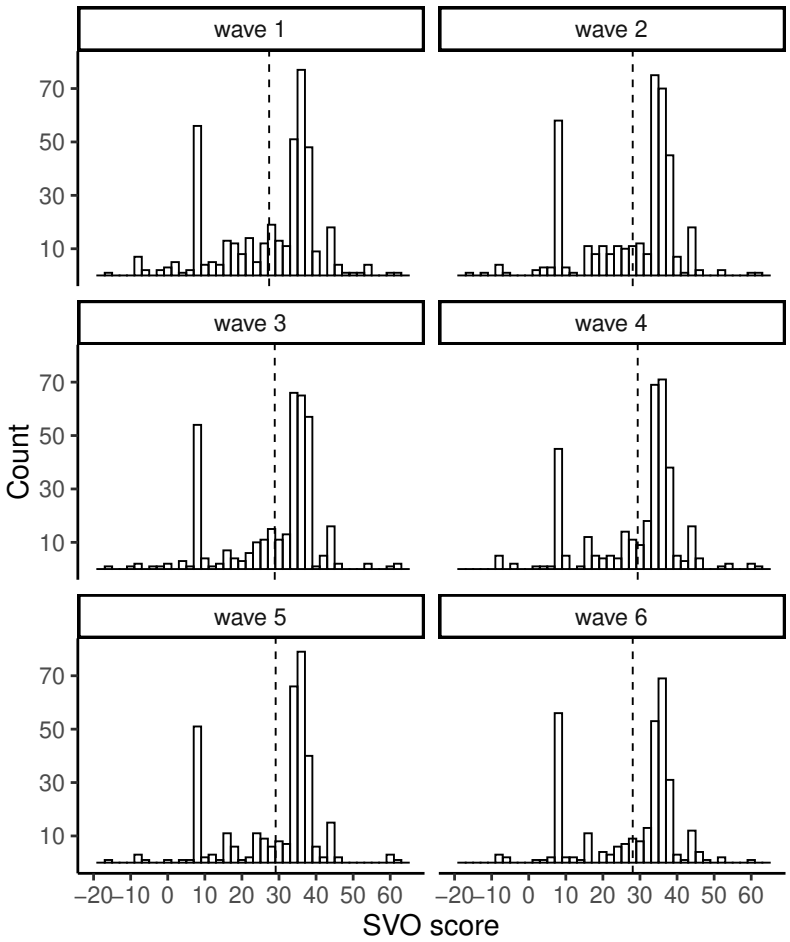

FIGURE 1: Visualizing SVO scores per wave. The mean is shown via a dashed line.

in Table 1) to provide a single index score per wave. Observed SVO degrees range from −16.1 to 61.4, with a mean across all respondents and waves of 28.4 (*SD* = 13.0). Lower scores on the scale indicate a more proself orientation while higher scores indicate that the respondents orient more prosocially. Table 1 shows that mean SVO scores vary marginally across waves but the high standard deviations point to substantial variance in SVO. Figure 1 provides us with a visual inspection of SVO distributions across waves. In particular, we see two major spikes, one approximately at score 8 and one near score 35. These represent respondents who consistently select either prosocial (score 35) or individualistic (score 8) allocations (Bakker & Dijkstra, 2021; Murphy et al., 2011). The descriptive analyses in Table 1 and Figure 1 point to the presence of variation, showing the need to explore intra-individual differences in SVO rather than wave-by-wave comparisons of mean scores.

Our sample suffered from attrition. Almost 33 percent of respondents dropped out from waves 1 to 6. In brief, attrition did not significantly affect the distribution of SVO in our sample. For example, comparing the wave 1 distributions of SVO scores between respondents who *had* and *had not* dropped out by waves 2 to 6, using a Kolmogorov–Smirnov test (which quantifies whether two distributions differ significantly from one another), shows no significant differences. For more information on the impact of attrition on the distribution of SVO, we refer to Appendix B.

4.2 Test-retest reliability of SVO scores

We use Pearson correlations, k-sample tests, and the intra-class correlation coefficient to indicate the test-retest reliability of SVO distributions. First, SVO scores correlate positively and significantly across waves (Table 2), meaning that respondents' SVO scores tend to be similar across all wave comparisons. Next, the similarity in SVO score distributions is confirmed by the Anderson-Darling (AD) k-sample test: the *p*-value of 0.57 indicates that we cannot reject the equality of SVO score distributions across waves. This result is in line with the distributions in Figure 1, the small differences in mean continuous SVO scores per wave, and the strong positive correlations reported in Table 2.[2] The distribution of SVO scores is fairly constant on the whole. Furthermore, we utilize the intra-class correlation coefficient (ICC) coefficient to inspect intra-individual consistency in SVO continuous scores. We find an ICC score for SVO scores of 0.78 (95% CI = [0.74, 0.82]) among respondents who participated in all six waves (N = 230). The high ICC score indicates that SVO continuous scores have a high test-retest reliability.

Moreover, we employ a multilevel linear regression to inspect whether prior SVO continuous scores are predictive of later SVO scores. Using a multilevel model, we control for the nested structure of our data in which SVO measures are nested within individuals.

[2]AD k-sample test *p*-value of the slider measure items comprise: 1-self = *not applicable*, 1-other = 0.25, 2-self = 0.08, 2-other = 0.08, 3-self = 0.38, 3-other = 0.38, 4-self = 0.95, 4-other = 0.95, 5-self = 0.07, 5-other = 0.07, 6-self = 0.94, and 6-other = 0.94. The *p*-values above 0.05 indicate that we cannot reject equality of distributions. Items are thus similarly distributed over time.

TABLE 2: Pearson correlations of SVO scores across waves.

	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
Wave 1	—					
Wave 2	0.75	—				
Wave 3	0.72	0.78	—			
Wave 4	0.71	0.75	0.82	—		
Wave 5	0.72	0.78	0.83	0.84	—	
Wave 6	0.66	0.70	0.81	0.83	0.84	—

Note. All p-values are < 0.0001.

We include a lagged variable of SVO continuous scores, representing one's SVO score at wave minus 1 ($t - 1$). The results are reported in Table 3. Notably, the SVO score in the previous wave is significantly and highly indicative of later SVO scores (estimate = 0.79, $SE = 0.01$, $p < 0.001$). Note that the wave coefficients represent the difference between the respective waves and the intercept coefficient (wave 2). Hence, the wave 3 coefficient is the combination of the intercept and wave 3 parameters, i.e., the estimate is 6.91 (wave 2 plus wave 3 estimates).

We visualize test-retest reliability based on the results of Table 3 in Figure 2. In the six plots, the diagonal black unity line represents perfect test-retest reliability in prior and consecutive SVO scoring. The x-axis shows a respondent's SVO score in the prior wave (referred to in Figure 2a as $t - 1$), showing waves 1 to 5. The y-axis shows the SVO score in wave t, ranging from wave $t = 2$ to 6. In Figures 2b-f we show a pairwise wave-to-wave comparison. Figure 2 also shows the result of a linear regression—blue line—with the prior SVO score as the independent variable and the current SVO score as the dependent variable. Each data point is a paired observation, showing the SVO score at $t - 1$ and t. We furthermore include a marginal distribution of the SVO score of dropouts in wave $t -$

TABLE 3: Results of the multilevel linear regression for estimating predictors of SVO scores.

Parameter	estimate	*SE*	p-value
Intercept (wave 2)	6.11	0.58	<0.001
Wave 3	0.80	0.58	0.169
Wave 4	0.56	0.58	0.334
Wave 5	0.12	0.59	0.842
Wave 6	−0.95	0.60	0.115
SVO score $t - 1$	0.79	0.01	<0.001

Note. $N = 426$ with 1700 decisions; *SE* = standard error.

1. Further analyses in Appendix B show that prior SVO scores are not key predictors of dropping out. In brief, Figure 2 shows that there is a strong tendency to score similarly in SVO across waves.

Results in Tables 1, 2, and 3 and Figures 1 and 2 show that the majority is largely similar in their SVO over time, but some variation persists. We quantify to which extent differences in SVO scores occur.[3] We calculate differences in SVO by subtracting the absolute value of a respondent's SVO score at $t - 1$ from the absolute value of the SVO score at t for respondents who participated in all six waves (N = 230 respondents with a total of 1380 scores). The mean difference between SVO at $t - 1$ and t is 3.36 (SD = 5.9).[4] The mean difference of 3.36 shows that respondents on the whole tend to marginally differ in SVO over time. In what follows, we provide aggregated percentages of absolute differences in SVO—and not separated per wave. Almost 55 percent of scores comparing SVO between $t - 1$ and t, a total of 752 scores, is smaller than 1 (628 cross-wave comparisons differ more than 1 unit in SVO). We see an increase to 71%, when we take 3 as an unit, instead of 1, as the dichotomous cutoff value in comparing differences in absolute SVO scores between $t - 1$ and t. Next, we use the standard deviation of 6.8 and two times the SD as cutoff values. Almost 79% and 93% report a difference lower in SVO across waves for 6.8 and 13.6 respectively. In sum, the majority of respondents report minor gradual differences in SVO scores over time, once again attesting to sufficiently high test-retest reliability.

Finally, although a major perk of the SVOSM is its potential to rely on continuous scores, it remains a largely standard practice in SVO research to compute either four or two SVO categories based on continuous SVO scores (Balliet et al., 2009; Bakker & Dijkstra, 2021; Murphy et al., 2011). Appendix A provides an overview of the distribution of SVO categories in our study as well as investigates the test-retest reliability of treating SVO as a categorical construct. The results are fairly the same: we find that respondents tend to orient similarly over time, while some measurement-to-measurement variation persists.

5 Discussion

The social value orientation slider measure (SVOSM) is favored over other SVO measures due to its easy implementation, low burden on respondents, clear consistency checks, high test-retest reliability, and usage of continuous SVO scores (Bakker & Dijkstra, 2021; Murphy et al., 2011; Murphy & Ackermann, 2014). Open questions were whether the SVOSM is reliable in non-student samples and over longer periods than one-week test-retest schemes. Our results show that this is indeed the case. Moreover, additional analyses in Appendix A allow us to recommend refraining from categorizing continuous SVO scores since imposing boundaries on a continuous SVO scale heavily affects the stability of SVO. Appendix B

[3]Multiple visualizations of individual trajectories of SVO are provided in our open access OSF folder.

[4]Mean difference in SVO per wave is: wave 1 → wave 2 = 4.99 (SD = 8.7), wave 2 → wave 3 = 4.30 (SD = 8.2), wave 3 → wave 4 = 3.75 (SD = 7.3), wave 4 → wave 5 = 3.41 (SD = 6.5), and wave 5 → wave 6 = 3.68 (SD = 6.3).

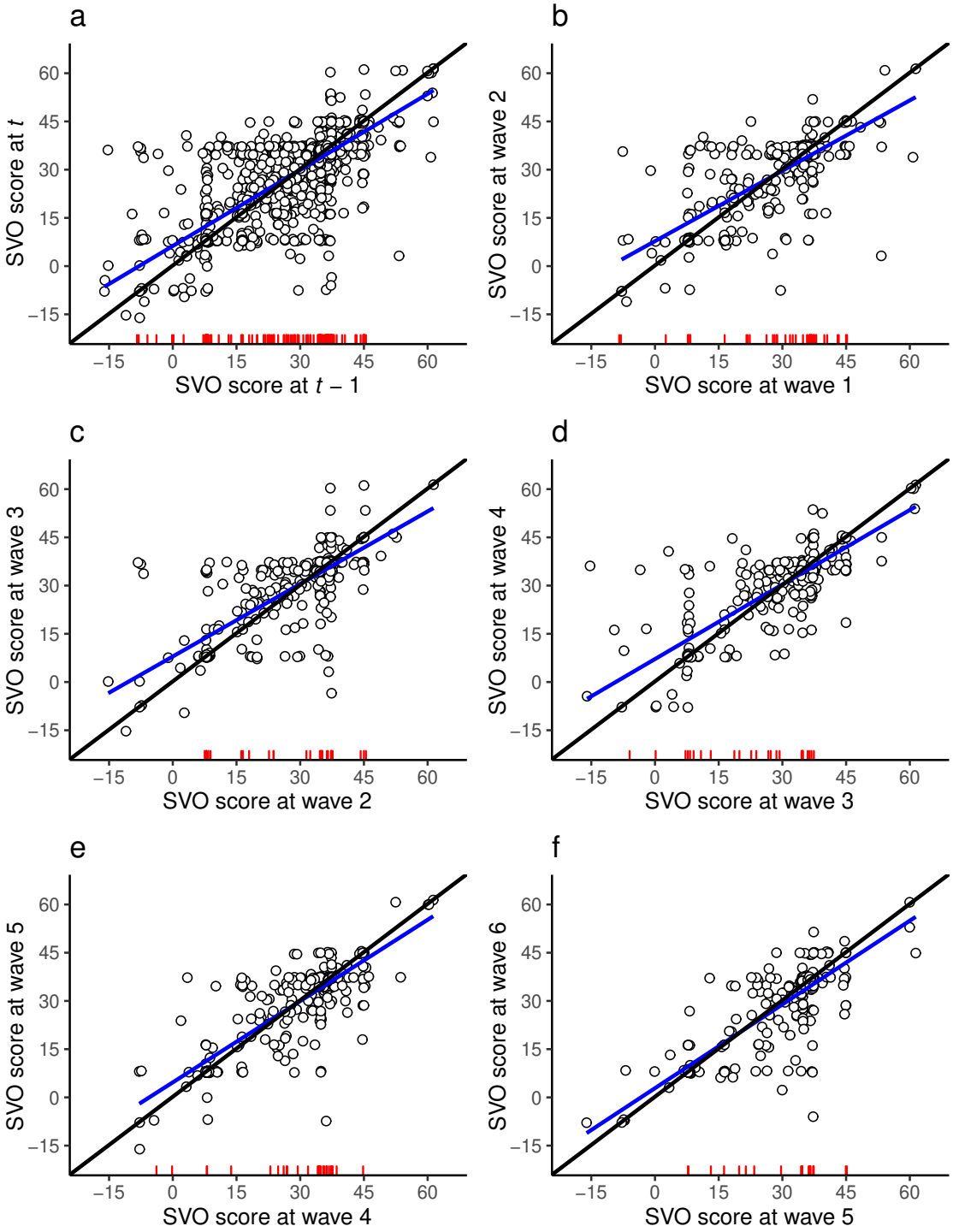

FIGURE 2: Test-retest reliability scatter plots. The diagonal black line represents perfect test-retest reliability. The blue line shows a linear regression with prior SVO ($t - 1$) as the independent variable and the current SVO score as the dependent variable (t). We show the marginal distribution of dropouts (no SVO score at t) in red. Panel a shows all waves combined, while panels b to f provide a wave-to-wave comparison of test-retest reliability.

shows that, while our study suffered sample attrition, dropouts do not differ significantly in SVO from respondents who did participate in later waves. Moreover, even with attrition, we had a sizeable sample of respondents.

Future work should investigate whether the stability of SVO also translates into stable predictions of cooperative behavior over time. Although major differences in SVO depending on monetary or non-monetary incentives are generally not expected (Balliet et al., 2009), findings from Reyna et al. (2018) suggest otherwise. Thus, future research should study the extent to which incentives affect the stability and predictive power of the SVOSM. Specifically, it would be valuable to know whether cooperative behavior is better predicted by monetary or by non-monetary incentivized measurements of SVO. Future research may also want to consider how individual characteristics and personal or social events — such as changes in income or occupation, experiences with voluntary work, ego depletion of guilt which shows to reduce prosocial behavior (Baumeister et al., 1994), or social integration — influence the stability of one's SVO.

The high test-retest reliability found in this study resembles the stability observed for other personality measures related to cooperativeness. Van Lange (1999) reported a 59 percent consistency score over 19 months with repeated measures using the SVO triple-dominance and ring measure. Bakker & Dijkstra (2021) found consistency percentages of 78, 71, and 67 for the slider, ring, and triple-dominance SVO measure, respectively, over a three-month period. Still, Van Lange and Bakker and Dijkstra utilized said SVO measures as categorical SVOs even though the ring and slider measure allows to assess SVO as a continuous construct. Akin to our results in Appendix A, there is some long-term variation found in the test-retest reliability among categorical SVOs. Moreover, similar accounts of high test-retest reliability are reported for the HEXACO (Dunlop et al., 2021), NEO (McCrae et al., 2011), and Big Five (Henry & Mõttus, 2020) personality inventories in which prosociality related to cooperation is assessed. We show that the slider measure can be added to the list.

The prime contribution of the current paper lies in answering the empirical question of whether SVOs, as measured by the slider measure, are *relatively stable* over time in non-student samples. Our results support classifying SVO as a personality trait.

References

Ackermann, K. A. & Murphy, R. O. (2019). Explaining cooperative behavior in public goods games: How preferences and beliefs affect contribution levels. *Games*, *10*(1), 15, https://doi.org/10.3390/g10010015.

Au, W. & Kwong, J. (2004). Measurements and effects of social-value orientation in social dilemmas: A review. In R. Suleiman, D. Budescu, I. Fischer, & D. Messick (Eds.), *Contemporary psychological research on social dilemmas* (pp. 71–98). New York, NY: Cambridge University Press.

Bakker, D. M. & Dijkstra, J. (2021). Comparing the slider measure of social value orientation with its main alternatives. *Social Psychology Quarterly*, *84*(3), 235–245, https://doi.org/10.1177/01902725211008938.

Balliet, D. P., Parks, C. D., & Joireman, J. A. (2009). Social value orientation and cooperation in social dilemmas: A meta-analysis. *Group Processes and Intergroup Relations*, *12*(4), 533–547, https://doi.org/10.1177/1368430209105040.

Baumeister, R. F., Stillwell, A. M., & Heatherton, T. F. (1994). Guilt: An Interpersonal Approach. *Psychological Bulletin*, *115*(2), 243–267, https://doi.org/10.1037/0033-2909.115.2.243.

Bogaert, S., Boone, C., & Declerck, C. (2008). Social value orientation and cooperation in social dilemmas: A review and conceptual model. *British Journal of Social Psychology*, *47*(3), 453–480, https://doi.org/10.1348/014466607X244970.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, *20*(1), 37–46, https://doi.org/10.1177/001316446002000104.

DeYoung, C. G. (2015). Cybernetic big five theory. *Journal of Research in Personality*, *56*, 33–58, https://doi.org/10.1016/j.jrp.2014.07.004.

Dunlop, P. D., Bharadwaj, A. A., & Parker, S. K. (2021). Two-year stability and change among the honesty-humility, agreeableness, and conscientiousness scales of the HEXACO100 in an Australian cohort, aged 24–29 years. *Personality and Individual Differences*, *172*, 110601, https://doi.org/10.1016/j.paid.2020.110601.

Fleeson, W. (2001). Toward a structure-and process-integrated view of personality: Traits as density distributions of states. *Journal of personality and social psychology*, *80*(6), 1011–1027, https://doi.org/10.1037/0022-3514.80.6.1011.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, *76*(5), 378–382, https://doi.org/10.1037/h0031619.

Henry, S. & Mõttus, R. (2020). Traits and Adaptations: A Theoretical Examination and New Empirical Evidence. *European Journal of Personality*, *34*(3), 265–284, https://doi.org/10.1002/per.2248.

Höglinger, M. & Wehrli, S. (2017). Measuring social preferences on amazon mechanical turk. In B. Jann & W. Przepiorka (Eds.), *Social dilemmas, institutions, and the evolution of cooperation* (pp. 527–546). Berlin, Boston: De Gruyter Oldenbourg.

LaHuis, D. M., Hartman, M. J., Hakoyama, S., & Clark, P. C. (2014). Explained variance measures for multilevel models. *Organizational Research Methods*, *17*(4), 433–451, https://doi.org/10.1177/1094428114541701.

Manesi, Z., Van Lange, P. A. M., Van Doesum, N. J., & Pollet, T. V. (2019). What are the most powerful predictors of charitable giving to victims of typhoon Haiyan: Prosocial traits, socio-demographic variables, or eye cues? *Personality and Individual Differences*, *146*, 217–225, https://doi.org/10.1016/j.paid.2018.03.024.

McCrae, R. R., Kurtz, J. E., Yamagata, S., & Terracciano, A. (2011). Internal Consistency, Retest Reliability, and Their Implications for Personality Scale Validity. *Personality and Social Psychology Review*, *15*(1), 28–50, https://doi.org/10.1177/1088868310366253.

Messick, D. M. & McClintock, C. G. (1968). Motivational bases of choice in experimental games. *Journal of Experimental Social Psychology*, *4*(1), 1–25, https://doi.org/10.1016/0022-1031(68)90046-2.

Murphy, R. O. & Ackermann, K. A. (2014). Social value orientation: Theoretical and measurement issues in the study of social preferences. *Personality and Social Psychology Review*, *18*(1), 13–41, https://doi.org/10.1177/1088868313501745.

Murphy, R. O. & Ackermann, K. A. (2015). Social preferences, positive expectations, and trust based cooperation. *Journal of Mathematical Psychology*, *67*, 45–50, https://doi.org/10.1016/j.jmp.2015.06.001.

Murphy, R. O., Ackermann, K. A., & Handgraaf, M. J. J. (2011). Measuring social value orientation. *Judgment and Decision Making*, *6*(8), 771–781.

Reyna, C., Belaus, A., Mola, D., Ortiz, M. V., & Acosta, C. (2018). Social values orientation slider measure: Evidences of validity and reliability among Argentine undergraduate students. *Testing, Psychometrics, Methodology in Applied Psychology*, *25*(3), 395–408, https://doi.org/10.4473/TPM25.3.5.

Snijders, T. A. B. & Bosker, R. J. (1994). Modeled variance in two-level models. *Sociological Methods & Research*, *22*(3), 342–363, https://doi.org/10.1177/0049124194022003004.

Thielmann, I., Spadaro, G., & Balliet, D. P. (2020). Personality and prosocial behavior: A theoretical framework and meta-analysis. *Psychological Bulletin*, *146*(1), 30–90, https://doi.org/10.1037/bul0000217.

Van Lange, P. A. M. (1999). The pursuit of joint outcomes and equality in outcomes: An integrative model of social value orientation. *Journal of Personality and Social Psychology*, *77*(2), 337–349, https://doi.org/10.1037/0022-3514.77.2.337.

Van Lange, P. A. M., Balliet, D. P., Parks, C. D., & Van Vugt, M. (2014). *Social dilemmas: Understanding human cooperation*. Oxford, UK: Oxford University Press.

Appendix A

Descriptive analysis of SVO as a categorical construct

We classify the observed SVO scores into categories: *prosocials* assign more weight to others' outcomes than *individualistic* types, while *competitive* (*altruistic*) types want to maximize the positive difference in outcomes between themselves (others) and others (themselves). Altruists have a score greater than 57.15. Prosocial scores lie between 22.45 and 57.15. Individualists have a score between −12.04 and 22.45. Respondents with a score less than −12.04 are classified as being competitively oriented. Most studies lump altruistic and prosocial types into a "prosocial" category and competitive and individualistic types into a "proself" category since altruistic and competitive types are rare.

Table 4 shows the count and percentage per SVO category and per wave in our sample. Most respondents are prosocially oriented while a good number have an individualistic orientation. Our sample contains hardly any competitive or altruistic respondents. Note that the *N* per column in Table 4 varies due to the post-hoc removal of intransitive and small vector length responses separately per wave (row 'excluded'). Ignoring the missing values due to sample attrition, we find that the percentage of prosocials (altruistic and prosocial types) is rather constant, floating within the bandwidth of 67 to 76 percent. At lower percentages, the same holds for proself types who show a consistent presence of around 24 to 33 percent (competitive and individualistic types). Consistent with these findings, most

TABLE 4: Count and percentage of respondents per SVO category per wave.

SVO type	Wave 1	Wave 2	Wave 3	Wave 4	Wave 5	Wave 6
competitive	1 (0.2%)	1 (0.2%)	1 (0.2%)	0 (0%)	1 (0.2%)	0 (0%)
individualistic	135 (32.5%)	111 (25.6%)	91 (20.5%)	85 (18.9%)	82 (17.6%)	88 (19.2%)
prosocial	277 (66.8%)	275 (63.5%)	275 (62.1%)	267 (59.1%)	253 (54.2%)	219 (47.8%)
altruistic	2 (0.5%)	2 (0.5%)	3 (0.7%)	3 (0.7%)	4 (0.9%)	1 (0.2%)
missing (NA)	0 (0%)	44 (10.2%)	73 (16.5%)	97 (21.5%)	127 (27.2%)	150 (32.8%)
excluded	80	62	52	43	28	37

Note. Excluded refers to intransitive and too short vector length cases; *N* per wave (without missing values and excluded cases): wave 1 = 415, wave 2 = 389, wave 3 = 370, wave 4 = 355, wave 5 = 340, and wave 6 = 308.

work usually reports that roughly two-thirds of their sample classifies as prosocial while approximately one-third is proself (Bakker & Dijkstra, 2021; Höglinger & Wehrli, 2017).

We assess the extent to which such SVO categorization leads to the loss of explained variance in SVO continuous scores. We estimate a multilevel linear regression model to account for the nested data structure. We take SVO continuous scores as the dependent variable and either the four or two SVO categories as the independent variable. Findings indicate that the four (R^2 = 0.81) and two (R^2 = 0.77) category implementations have a high and roughly similar degree of explanatory power.[5] This statistical finding supports the prosocial-proself dichotomy usually employed by researchers using the SVOSM. Still, some variance in SVO remains unexplained due to categorization.

SVO categorical test-retest reliability

Previous assessments of the test-retest reliability regarding SVOSM as categories point to a stable construct over two measurements one week apart. Höglinger & Wehrli (2017) show that 86 percent of SVO categorical classifications remained similar (N = 872). The developers of the SVOSM report an 89 percent consistency score (Murphy et al., 2011), while Bakker & Dijkstra (2021) report a 78 percent categorical type consistency score with two measurements three months apart. The categorical test-retest reliability decreased to 67 percent over a period of 1.5 years (Bakker & Dijkstra, 2021), but with only N = 27. In the current study, the mean overall instability across all waves is 0.12 (percentage stability is 88%), indicating that on average about 12% of respondents change categories from one wave to the next. The measurement-to-measurement variation in categorical instability is as follows: wave 1 → wave 2 = 0.17, wave 2 → wave 3 = 0.14, wave 3 → wave 4 = 0.10, wave 4 → wave 5 = 0.10, and wave 5 → wave 6 = 0.09. The trend appears to show an increasingly stable classification.

We use the following two statistics to formally inspect test-retest reliability in SVO categories: Cohen's (1960) and Fleiss' (1971) Kappa (κ). First, Cohen's κ allows us to check whether respondents stick to their categories in consecutive waves. We find that respondents are consistent in their SVO: wave 1 → wave 2 = 0.65 (N = 343), wave 2 → wave 3 = 0.70 (N = 342), wave 3 → wave 4 = 0.77 (N = 328), wave 4 → wave 5 = 0.76 (N = 317), and wave 5 → wave 6 = 0.79 (N = 299). Second, Fleiss' κ is an adaptation of Cohen's κ and allows to assess consistency across all waves at the same time. We find a high Fleiss' κ of 0.70 of respondents who participated in all six waves (N = 230).

Figure 3 visualizes variation in SVO. The dark grey block indicates prosocials, while the light grey block represents proselfs (white is NA). Figure 3 shows how the prosocial and proself categories exchange members over time, while a stable flow of respondents drops out at every transition. The pool of dropouts consists mainly of prosocials, which

[5]Assessing explained variance in multilevel models can be done via multiple R^2 measures (LaHuis et al., 2014). We rely on the Snijders & Bosker (1994) R^2 measure because it captures variance in two-level models, which we have.

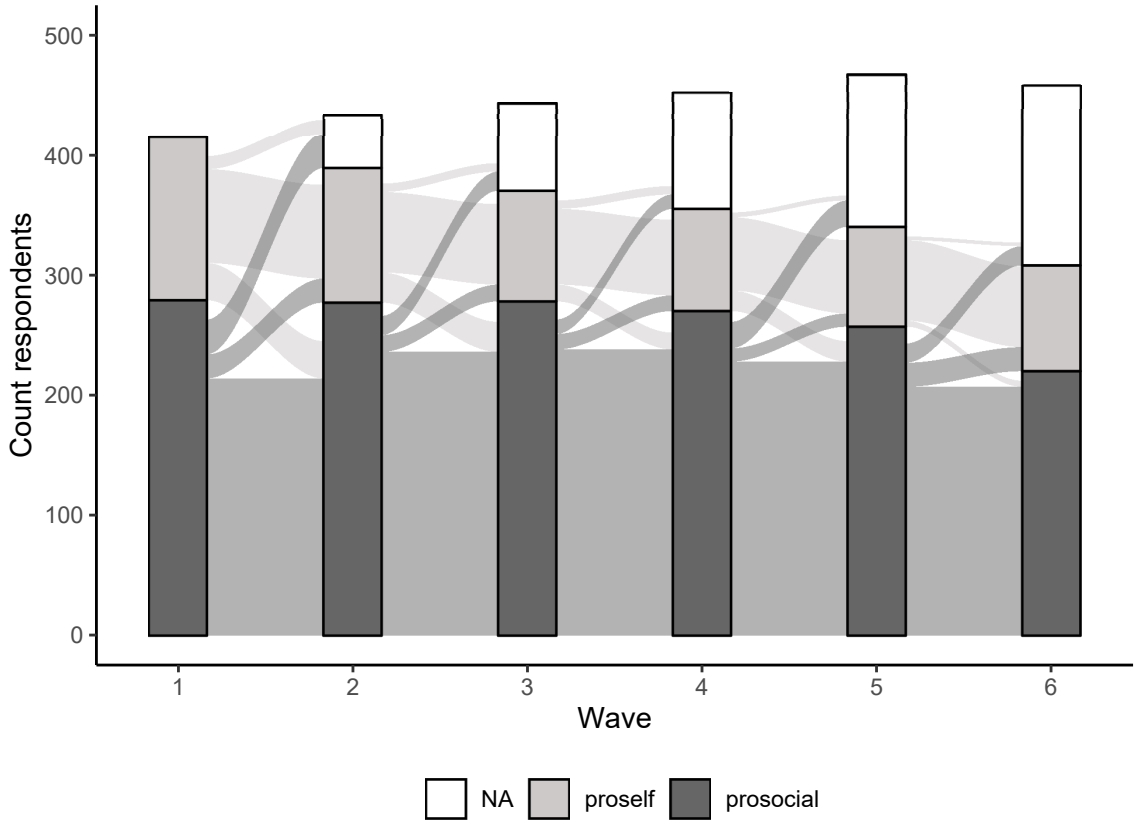

FIGURE 3: An alluvial diagram visualizing changes in SVO. Respondents with similar prior and current SVO are bundled together. Respondents with prospected intransitive answer profiles do not have a visualized trajectory in-between waves. The *N* per column varies due to the post-hoc removal of intransitive and too small vector length responses per wave.

is unsurprising given that they make up about two-thirds of our sample. The *N* per wave varies due to the post-hoc removal of intransitive and small vector length responses.

Prior research indicates that respondents scoring near the classification boundaries are more likely to switch SVO category classification (Bakker & Dijkstra, 2021), attesting to the importance of using continuous scores. To investigate whether this holds in our sample, we estimate a multilevel logistic regression model. The dependent variable is whether individuals changed in terms of SVO category from one wave to the next (1 = change and 0 = no change). To calculate proximity to the prosocial-proself boundary, we first subtracted the number 22.45 (the category boundary in degrees) from the continuous SVO scores, followed by converting scores to absolute values. We then reversed the variable by subtracting the calculated distance from the maximum possible distance, so that higher scores indicate proximity to the boundary. Results show that respondents scoring near the boundary are more likely to change SVO categories than respondents farther from the boundary (estimate = 0.05, $p = 0.002$). This finding is in sync with prior research (Bakker & Dijkstra, 2021). Minor gradual changes in SVO categories may thus lead to major

consequences in SVO stability in the long haul. Also, the negative waves effects (e.g., wave 1 → 2 estimate = −3.12, $p < 0.001$) indicate that changing SVO categories is not very likely from the outset and becomes even less likely in later waves (e.g., wave 5 → 6 estimate = −4.07, $p < 0.001$), which confirms the low instability percentages discussed earlier. Furthermore, prosocially oriented respondents are less likely to change their SVO category than proselfs (estimate = −0.93, $p < 0.001$).

Appendix B

Attrition in our sample

Data collection started with $N = 495$ in wave 1 and ended up with $N = 345$ in wave 6 (see Table 4 in Appendix A). A total of 44 respondents dropped out in wave 2, 73 in wave 3, 97 in wave 4, 127 in wave 5, and 150 in wave 6. The total attrition rate is 33% percent when comparing the sample size from waves 1 to 6. Figures 4a and b visualize the distribution of prosocial and proself categories with and without the NA, *not available*, cohort. The percentages of types remain rather similar over time. This is a first indication that attrition did not substantively affect the distribution of prosocial and proself types in our sample.

We formally test the role of attrition on SVO using Fisher's exact test (categorical SVO) and Kolmogorov-Smirnov test (continuous SVO). Applying Fisher's exact test to the distributions of SVO categories of wave 1 respondents who *had* and *had not* dropped out by wave 2 to 6, we find no statistical difference ($p = 1$ for all wave 1 to future wave, 2 to 6, comparisons). Thus, the impact of attrition on the SVO category distribution seems minimal. The same holds for treating SVO as a continuous construct. A Kolmogorov–Smirnov test shows that the SVO continuous score distribution of wave 1 respondents who, again,

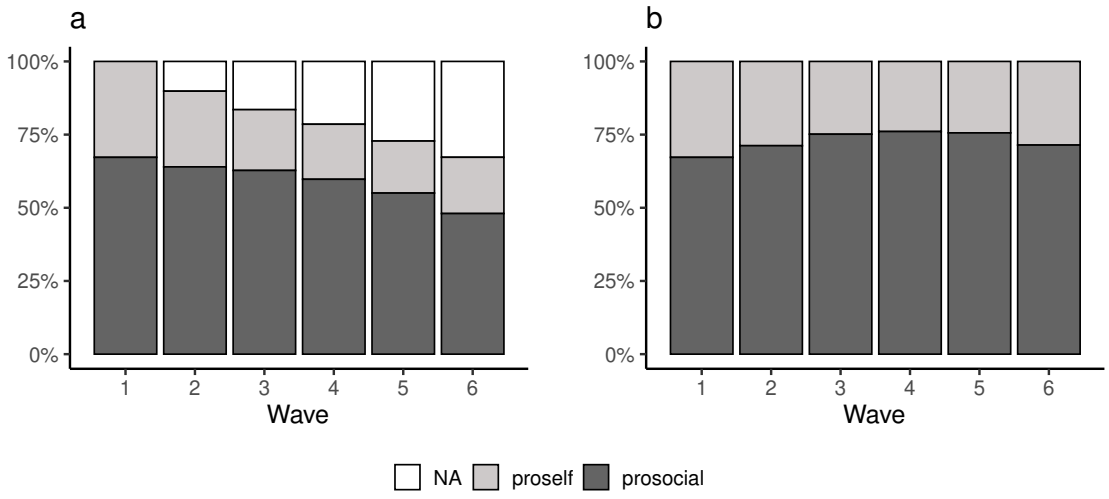

FIGURE 4: Visualizing percentages prosocial and proself types with (a) and without (b) NA's.

had and *had not* dropped out by wave 2 to 6 are equally distributed. To be clear, the Kolmogorov–Smirnov test p value per wave as follows: wave 1 → wave 2 = 0.24, wave 1 → wave 3 = 0.45, wave 1 → wave 4 = 0.28, wave 1 → wave 5 = 0.36, and wave 1 → wave 6 = 0.73. Attrition thus did not lead to significant differences in SVO distributions.

Next, we investigate whether changing in SVO categories is a prerequisite for dropping out in later waves. We conducted supplementary logistic regression analyses with dropping out measured at waves 3, 4, 5, and 6 as dependent variables (1 = dropping out, 0 = maintaining participation). To be clear, we tested in four separate logistic models whether, for example, changing in SVO from wave 1 to 2 increases the likelihood to drop out in wave 3. Subsequent models includes changing SVO from wave 2 to 3, 3 to 4, and 4 to 5 as independent variables and dropping out in respectively waves 4, 5, and 6 as dependent variables. We included changing in SVO categories from wave 1 to 2 (estimate = 0.14, SE = 0.57, p = 0.81), 2 to 3 (estimate = −0.97, SE = 1.04, p = 0.35), 3 to 4 (estimate = −0.27, SE = 0.76, p = 0.73), and 4 to 5 (estimate = −0.58, SE = 1.05, p = 0.58) as independent variables. In general, our analyses reveal that instability in SVO in the past is not a significant predictor of dropping out in future waves.

We furthermore explore whether a proself vs. prosocial orientation as well as orientating high or low on the continuous SVO scale is a predictor of dropping out in our sample. The dependent variable is again dropping out (1) or not (0). We include SVO at $t - 1$ as a dichotomous or continuous independent variable. The continuous SVO scores in $t - 1$ are generally not predictive of dropping out in wave 2 (estimate = 0.01, SE = 0.01, p = 0.46), wave 3 (estimate = −0.00, SE = 0.01, p = 0.82), wave 5 (estimate = 0.00, SE = 0.02, p = 0.96), or wave 6 (estimate = 0.01, SE = 0.02, p = 0.69). The sole exception is dropping out in wave 4 (estimate = −0.03, SE = 0.01, p = 0.04). Respondents with higher SVO scores are less likely to drop out than respondents lower on SVO. Twenty-four respondents (6%) dropped out in wave 4 and 355 (94%) maintained to participate in our study. The mean SVO score of those 24 respondents is 22.7 (SD = 13.3) vs. 28.6 (SD = 13.2) of the *stayers*. An additional Kolmogorov–Smirnov test shows that the SVO continuous score distribution from wave 3 ($t - 1$) among dropouts and stayers in wave 4 does not significantly differ according to $p < 0.05$ standards: p-value = 0.06. The analysis of SVO as a category does not confirm the higher chances of proself types to drop out more readily than their prosocial counterparts: wave 4 estimate = −0.71, SE = 0.43, p = 0.10). The non-effect of prosocial and proself categorization is confirmed in other waves: wave 2 (estimate = 0.30, SE = 0.36, p = 0.39), wave 3 (estimate = −0.13, SE = 0.42, p = 0.76), wave 5 (estimate = 0.54, SE = 0.51, p = 0.28), or wave 6 (estimate = −0.13, SE = 0.49, p = 0.80).

Finally, we study whether having an *extreme* SVO — for example, preferring to maximize payoffs to themselves or the other — is a predictor of dropping out. Extremeness in SVO is calculated as follows: as a benchmark, we take the diagonal line in the distribution between payoffs to self and the other (as visualized in Murphy et al., 2011, p. 773, Figure 2). The

diagonal line represents a SVO score of 45.[6] Then, we calculate the absolute distance to 45 in SVO in $t-1$ and use that indicator as an explanatory variable for the logistic regression. The dependent variable is dropping out (1) or not (0). Distance to the 45 benchmark SVO score in $t-1$ is generally not predictive of dropping out in wave 2 (estimate = -0.01, $SE = 0.01$, $p = 0.34$), wave 3 (estimate = 0.00, $SE = 0.02$, $p = 0.86$), wave 5 (estimate = -0.00, $SE = 0.02$, $p = 0.83$), or wave 6 (estimate = -0.01, $SE = 0.02$, $p = 0.59$). The sole exception is again the impact of distance to the benchmark in dropping out in wave 4 (estimate = 0.03, $SE = 0.01$, $p = 0.04$). Respondents with more extreme SVO preferences — preferring to maximize differences to the other, either beneficial for themselves or the other — are generally more likely to drop out in wave 4. In the previous paragraph, we already stressed that especially respondents with low SVO scores drop out in wave 4, suggesting that these respondents generally favor higher payoffs allocated to themselves than the other. In brief, respondents with a particular SVO score do not disproportionately drop out in our study—cf. respondents with lower SVO degrees in wave 4.

[6]Mathematically, the benchmark represents $\pi/4$, a perfectly straight diagonal line in a plot. For more information on describing SVO as a degree angle score (SVO°), we refer to Murphy et al. (2011).